Article scientifique   Article   2023

Published version   Open Access

Exploring the Latest Highlights in Medical Natural Language Processing across Multiple Languages : A Survey

---

Shaitarova, Anastassia; Zaghir, Jamil; Lavelli, Alberto; Krauthammer, Michael; Rinaldi, Fabio

This publication URL: https://archive-ouverte.unige.ch/unige:176613
Publication DOI: 10.1055/s-0043-1768726

Exploring the Latest Highlights in Medical Natural Language Processing across Multiple Languages: A Survey

Anastassia Shaitarova[1*], Jamil Zaghir[2,3*], Alberto Lavelli[4], Michael Krauthammer[5,6], Fabio Rinaldi[4,5,7,8]

1   Department of Computational Linguistics, University of Zurich, Zurich, Switzerland
2   Division of Medical Information Sciences, University Hospitals of Geneva, Geneva, Switzerland
3   Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland
4   Natural Language Processing Research Unit, Center for Digital Health and Wellbeing, Fondazione Bruno Kessler, Trento, Italy
5   Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland
6   Biomedical Informatics, University Hospital Zurich, Zurich, Switzerland
7   Dalle Molle Institute for Artificial Intelligence Research, Lugano, Switzerland
8   Swiss Institute of Bioinformatics

* These authors contributed equally to this work

Summary

Objectives: This survey aims to provide an overview of the current state of biomedical and clinical Natural Language Processing (NLP) research and practice in Languages other than English (LoE). We pay special attention to data resources, language models, and popular NLP downstream tasks.

Methods: We explore the literature on clinical and biomedical NLP from the years 2020-2022, focusing on the challenges of multilinguality and LoE. We query online databases and manually select relevant publications. We also use recent NLP review papers to identify the possible information lacunae.

Results: Our work confirms the recent trend towards the use of transformer-based language models for a variety of NLP tasks in medical domains. In addition, there has been an increase in the availability of annotated datasets for clinical NLP in LoE, particularly in European languages such as Spanish, German and French. Common NLP tasks addressed in medical NLP research in LoE include information extraction, named entity recognition, normalization, linking, and negation detection. However, there is still a need for the development of annotated datasets and models specifically tailored to the unique characteristics and challenges of medical text in some of these languages, especially low-resources ones. Lastly, this survey highlights the progress of medical NLP in LoE, and helps at identifying opportunities for future research and development in this field.

Keywords

Multilingualism, natural language processing, datasets as topic, language models, shared tasks

1 Introduction

Natural Language Processing (NLP) has established itself in the medical field as a valuable method of data processing. It is widely used in biomedical research as well as in several clinical applications [1–3]. While the initial success of NLP stemmed from work in the English language, the need for reliable NLP tools in other languages is ever-growing, particularly regarding healthcare [4–6]. The coronavirus pandemic brought to light how much cross-linguistic barriers impede scientific development [7]. A large number of biomedical papers and clinical texts are produced in Languages other than English (LoE), yet most language technology and text processing systems are developed for English only.

As NLP is a rapidly evolving field, characterized by frequent releases of new resources and solutions, there is an ongoing need for literature reviews and summaries. The first overview of clinical NLP in LoE was published by Névéol *et al.* in 2018 [4]. The authors surveyed 114 papers on 13 languages spanning over a decade of clinical NLP. Aside from confirming a steady interest in NLP in LoE, the authors stated the need for shared tasks and datasets in LoE. The latest IMIA NLP surveys [8, 9] covered several language-specific adaptations of transformer models as well as the transfer learning method. We pick up where the previous reviews left off and highlight clinical and biomedical NLP advances for LoE from the years 2020-2022. We also expand slightly beyond the purely clinical field to include corpora, shared tasks, and tools that deal with medical text more broadly, yet excluding more biologically oriented NLP activities.

We queried online databases and surveyed 129 papers dealing with 27 languages. Our methods are described in Section 2, while Section 3 details our findings.

2 Methods

We collected most of the papers reviewed for this study in November 2022 using Google Scholar. Although not perfect, this search engine aggregates publications from various databases. Additionally, some publicly available archives and anthologies (such as the Anthology of the Association for Computational Linguistics) rely on Google's indices for their paper retrieval. To automate our search, we utilized an available API[1]. Due to time and resource constraints, we did not create individual queries for LoE. Instead, we used the general terms „multilingual" and „cross-lingual," as these terms are frequently mentioned in papers addressing tasks in LoE. Furthermore, in the existing NLP literature, the term "multilingual" can refer to a wide range of topics, including the implementation of language-agnostic pipelines [10], learning language representations in multilingual space, and experimenting with downstream tasks in LoE. Multilingual language models are often used in the latter case [11, 12].

Our main query was `(multilingual OR cross-lingual) AND (medical OR clinical) AND (NLP OR „natural language processing")`, with a time window restriction of 2020-2022. To retrieve results for our selected topics, we used targeted queries with specific keywords, including `(corpus OR database)`, `(NER OR „named entity recognition")`, and `negation`. These targeted queries resulted in 7,990, 3,280, and 5,950 hits, respectively. We ordered our findings based on relevance[2] and manually inspected the top 100 titles and abstracts in each category, selecting papers that address the predefined downstream tasks in LoE, and publicly accessible non-English and, when available, multilingual models and datasets. Despite relevance-based ordering, we encountered many irrelevant hits, confirming the limitations of automatic querying of online databases. We used the same query to search PubMed and medRxiv, selecting 10 papers from PubMed and one already published

paper from medRxiv that met our criteria. Furthermore, we analyzed the retrieved publications to assemble a list of shared tasks from 2020 to 2022, examining their proceedings for multilingualism, cross-lingual methods, and LoE. Eleven papers in our survey come from arXiv.org, with seven of them having peer-reviewed publications available at a later date.

As an additional point of reference, we collected all available NLP review and survey papers in the biomedical and clinical domain that were published in the last two years and mentioned multilingualism [1, 3, 13–15]. Although we tried to be as complete as possible, this paper is not meant to be a comprehensive review. There could be limitations due to the usage of search engines, language-agnostic queries, and the fact that we focused on papers written in English.

3 Results

Section 3.1 introduces new corpora in LoE as well as some multilingual resources. Section 3.2 lists new large Pretrained Language Models (PLMs) in LoE and discusses various pretraining strategies. Sections 3.3 and 3.4 present new publications and advances in two foundational medical NLP tasks, namely Named Entity Recognition (NER) and negation detection. Special attention is given to the recently shared tasks in LoE, which we provide in Section 3.5. While conducting our survey, we identified an unparalleled rise in Spanish medical NLP. We dedicate Section 3.6 to this phenomenon.

3.1 New Multilingual Resources and Monolingual Datasets in LoE

There is a well-known need for biomedical databases and annotated corpora in different languages. The issue remains at the center of attention for many years while slowly improving. In their 2018 survey, Névéol *et al.* listed annotated corpora in French, German, Greek, Japanese, Portuguese, Spanish, and Swedish [4]. In our survey, we address annotated corpora in LoE that have been made public in the years 2020-2022. We note that the number of languages has more than doubled and now we

have data in Arabic, Basque, Bengali, Catalan, traditional and simplified Chinese, Dutch, French, Spanish, German, Italian, Japanese, Korean, Portuguese, Romanian, Russian, Serbian, Spanish and Vietnamese. The publicly available datasets can be found in Table 1.

A notable trend concerns the availability of multilingual corpora. The European Clinical Case Corpus (E3C) is a collection of clinical narratives and descriptive clinical documents in five languages [16]. The corpus contains annotations for temporal information as well as clinical entities. While English, French and Spanish texts were sourced from publicly available corpora, clinical cases in Italian and Basque had to be manually extracted from various sources. Significant advances in the quality of Neural Machine Translation (NMT) opened new opportunities for the creation of silver standard corpora. To foster the development of multilingual tools, the LivingNER [17] and DisTEMIST [18] corpora provide gold standard annotated data in Spanish together with the machine-translated and post-edited counterparts in English, Portuguese, French, Italian, Romanian, Catalan, and Galician. The two corpora feature annotations of species mentions and disease mentions, respectively, which have been transferred to all the languages. NMT facilitated the development of RuMedNLI [19], the first comprehensive open Russian medical language understanding benchmark, which is a Russian version of MedNLI [20]. The corpus was initially translated using Google Translate and DeepL MT systems, followed by post-editing from a team of native speakers and medical specialists who corrected numerous domain-specific translation errors, as well as cultural and localization issues, including drug name adaptations and measurement conversions.

Several corpora have been developed for German, most of them being an approximation for real data. There now exists an MT-produced GERNERMED [21], as well as a fully generated GPTNERMED [22] corpus. Both of these silver standard corpora were created using open resources and utilized to train models which were then evaluated on gold standard data. The Graz Synthetic Clinical Corpus (GraSCCo) is composed of original clinical documents that have undergone several rounds of anonymization and significant

---

Shaitarova et al

Table 1 Multilingual resources and non-English datasets.

Corpus and publication	Language(s)[6]	Target tasks	Data description	Availability
Boudjellal et al., 2020 [34]	ara	NER, UMLS Entity Linking	49,856 biomedical sentences made from a dictionary-based automatic annotation	Available upon request
BanglaBioMed [32] Sazzed, 2022	ben	NER	818 sentences, 11,196 tokens, health articles	https://github.com/sazzadcsedu/BanglaBioMed.git
BRONCO [26] Kittner et al., 2021	deu	NER, Negation, Speculation	200 oncological discharge summaries, 11,434 sentences, 89,942 tokens	Available upon request https://www2.informatik. hu-berlin.de/~leser/bronco/index.html
GGPONC [24] Borchert et al., 2020	deu	NER	8,414 text segments, 1.3M tokens	Available upon request https://www.leitlinienprogramm-onkologie.de/projekte/ggponc-english/
GGPONC 2.0 [25] Borchert et al., 2022	deu	NER	10K text segments, 78K sentences, 1.8M tokens	Available upon request https://www.leitlinienprogramm-onkologie.de/projekte/ggponc-english/
GPTNERMED [22] Frei and Kramer, 2022	deu	NER	GPT-generated text and annotations: 245,107 tokens annotated with doses, medications and diagnoses	https://github.com/frankkramer-lab/GPTNERMED
GraSCCo [23] Modersohn et al., 2022	deu	Classification	63 alienated clinical case reports, about 43K tokens covering 13 medical topics	https://zenodo.org/record/6539131#.Y7WAUnbMKUk
CAS [27] Grabar et al., 2020	fra	NER, Negation, Uncertainty	4,900 clinical cases, ~1.7M words	Available upon request http://natalia.grabar.free.fr/resources.php#cas
CLISTER [28] Hiebel et al., 2022	fra	Semantic Text Similarity	1K medical sentence pairs with a similarity score (0-5)	https://gitlab.inria.fr/codeine/clister
Real-MedNLP [29] Yada et al., 2022	jpn	NER	Case reports, radiology reports	https://sociocom.naist.jp/real-mednlp/wp-content/uploads/sites/3/2021/06/Yada2020.lrec-1.561-1.pdf
Kim and Lee, 2021 [30]	kor	NER	536 Question-Answer pairs	Available upon request https://github.com/labihem/BERT-NER-KO
Kim et al., 2022 [31]	kor	NER, Query intent detection	4,562 Question-Answer pairs annotated for medical NER	Available upon request
MedQA [141] Jin et al., 2021	multiling: eng, zho, simplified zho	QA	Eng - 12,723 questions, zho - 34,250; simplified zho - 14,123	https://github.com/jind11/MedQA
E3C [16] Magnini et al., 2021	multiling: eus, eng, fra, ita, spa	NER, temporal information, Factuality	25K+ tokens (L1), 50-100K tokens (L2), 1M tokens (L3) Clinical cases	https://e3c.fbk.eu/
LivingNER [17] Miranda-Escalada et al., 2022	multiling: spa translated into cat, glg, fra, ita, por, ron	NER	2K clinical case reports	https://zenodo.org/record/6768606
DisTEMIST [116] Miranda-Escalada et al., 2022	multiling: spa translated into cat, glg, fra, ita, por, ron	NER	2K clinical case reports	https://zenodo.org/record/6671292#.Y6AtLHbMKUk
ICR [142] Andrade et al., 2021	multiling: spa, por	NER, NEL	1,555 parallel articles in oncological domain	https://github.com/lasigeBioTM/ICERL_system-ICR_Corpus
ECCParaCorp [143] Ma et al., 2020	multiling: zho, eng	-	6,685 aligned eng-zho cancer texts	http://www.phoc.org.cn/ECCParaCorp/
Olthof et al., 2022 [51]	nld	Text classification	2,256 radiology requests and reports	https://data.overheid.nl/dataset/14845-covid-19-ziekenhuisopnames--volgens-de-nice-registratie--per-gemeente-per-ziekenhuisopnamedatum-e#panel-resources
SemClinBr [79] Oliveira et al., 2022	por	NER, Negation	1,000 clinical notes, labeled with 65,117 entities and 11,263 relations	Available upon request https://github.com/HAILab-PUCPR/SemClinBr
SiMoNERo [144] Mititelu and Mitrofan, 2020	ron	NER	4,861 sentences, 108,415 tokens	https://github.com/UniversalDependencies/UD_Romanian-SiMoNERo/find/master
RuMedBench [145] Blinov et al., 2022	rus	Diagnosis prediction, Symptoms recommendation, QA	Contains several datasets, including 7,625 de-identified case reports	https://github.com/pavel-blinov/RuMedBench

6   We use the standardized nomenclature ISO 639-3 for the language codes (https://iso639-3.sil.org/code_tables/639/data).

Table 1 (continued)   Multilingual resources and non-English datasets.

Corpus and publication	Language(s)	Target tasks	Data description	Availability
RuMedNLI [19] Blinov et al., 2022	rus	NLI (Natural Language Inference)	Silver-standard from MedNLI, an English dataset for NLI based on MIMIC-III records	https://www.physionet.org/content/rumednli-russian-inference/1.0.0/
NUBes [78] Lima Lopez et al., 2020	spa	Negation, Uncertainty	29,682 sentences from EHR	https://github.com/Vicomtech/NUBes-negation-uncertainty-biomedical-corpus
CANTEMIST [70] Miranda-Escalada et al., 2020	spa	NER	1.3K synthetic oncological clinical case reports	https://zenodo.org/record/3978041#. Y6At9HbMKUk
CodiEsp [92] Miranda-Escalada et al., 2021	spa	NER	1K clinical cases annotated with ICD-10 codes	https://zenodo.org/record/3758054#. Y7VDOXbMKUk
MEDDOPROF [88] Lima-Lopez et al., 2021	spa	NER, Text classification	1.8K annotated clinical cases	https://zenodo.org/record/5070541#. Y7U9BXbMKUk
SpRadIE [99] Cotik et al., 2021	spa	NER, Text classification	513 annotated ultrasonography pediatric cases	https://sites.google.com/view/spradie-2020/data
CT-EBM-SP [146] Campillos-Llanos et al., 2021	spa	NER	1,200 clinical trial texts annotated with UMLS semantic groups	http://www.lllf.uam.es/ESP/nlpmedterm_en.html
Cancer Dataset [77] Pabón et al., 2022	spa	Negation, uncertainty	2,700 sentences manually annotated	Available upon request cristina.avendano@uam.es
UIT-ViNewsQA [33] Van Nguyen et al., 2022	vie	QA	22,057 human-generated question answer pairs from healthcare news articles	Available upon request https://nlp.uit.edu.vn/datasets

linguistic alterations [23]. The corpus was subsequently compared syntactically and semantically with real, non-shareable German clinical data and found to be a suitable dataset for training clinical language models. GGPONC [24] and GGPONC 2.0 [25] contain German clinical guidelines for oncology and can be used as a proxy for real clinical data. Finally, BRONCO [26] is a dataset of de-identified real clinical documents. However, this corpus is rather small and its sentences are randomly shuffled across documents to prevent deanonymization.

Two medical French datasets, CAS [27] and CLISTER [28], comprise clinical cases extracted from the scientific literature. Yada *et al.* [29] created a Japanese dataset for NER, consisting of case and radiology reports. Additionally, there are now two Korean datasets for NER that feature multiple question-answer pairs [30, 31]. Other resources include SIMONERO, a Romanian medical treebank with gold standard morphological annotations; BanglaBioMed [32], a biomedical named-entity annotated Bengali corpus; and UIT-ViNewsQA [33], a crowd-sourced Vietnamese dataset. Boudjellal *et al.* [34] created an Arabic dataset for NER

and entity linking, employing a silver standard annotation scheme. This dataset was later utilized for fine-tuning a NER model, yielding good results.

Unfortunately, many corpora in LoE remain unavailable to the public for various reasons (ethics, data sensitivity, company policy, etc.). Nevertheless, they are often featured in publications that carry detailed and valuable information on the specificities of a particular LoE (Ukrainian [35]); the resource selection (Arabic [36]); the annotation process (Tibetan [37]), or the evaluation of different machine learning methods (French [38]).

Aside from annotated corpora, there is an interest in generalizable methods of corpus creation that are also extendable to other languages or adaptable to other domains. Vivaldi and Rodriguez [39] pooled together 13 available mono-, multi-, and cross-lingual resources in seven languages, including high (English), medium (French, German, Spanish), and low resource languages (Arabic, Basque, and Catalan), to create a multilingual terminology and a language-agnostic methodology for medical semantic tagging.

Overall, we report 30 newly available medical corpora in LoE (Table 1), with 20 of them evaluated on NER and four on negation and uncertainty, for a total of 15 distinct languages (19 if silver standard corpora are also taken into account). Moreover, nine of these datasets are used in the shared tasks in Spanish, French, and Japanese. Considering that multiple datasets are now available for some LoE (for example, Spanish and German), there is an opportunity to pool them into benchmarks for evaluating language models on biomedical and clinical tasks. This would make a great contribution to biomedical NLP research since domain-appropriate comprehensive benchmarks and leaderboards are still lacking for many languages.

3.2 Pretrained Language Models for LoE

Large Pretrained Language Models (PLMs) have become the state-of-the-art method for solving various NLP tasks. There is now a plethora of biomedical and clinical PLMs available in English [13, 40]. Moreover, there is a steadily growing number of PLMs

in LoE. We report biomedical and clinical language-specific PLMs that have become available since 2020 in Table 2. Most models are initialized with the weights of their language-specific general-domain counterparts [41–43] which has been the go-to method for the creation of domain-specific models. However, the general-domain vocabulary created within the model during pretraining is not representative of the biomedical or clinical domain and can endanger the downstream performance. Similarly, training models from scratch on mixed-domain data is disputed by the biomedical NLP community, especially in a low-resource setting [44]. Overall, in-domain pretraining is arguably the best option, given the availability of training data and computational resources [45–47].

Nevertheless, mixing closely related domains, such as biomedical and clinical data can boost performance even within a mid-resource scenario [48]. Moreover, mixed-domain pretraining can be unavoidable when resources in the target language are scarce. Despite the controversy about mixed-domain pretraining, there persists an assumption that out-of-domain text contributes to better language modeling. Pretraining on a domain-specific corpus alone can be insufficient to learn proper language representations [43].

Another question that arises when it comes to PLMs in the biomedical domain is whether to use monolingual or multilingual models for tasks in LoE. General-domain monolingual BERT models tend to outperform multilingual models (for example, BERTje for Dutch [49] or EstBERT for Estonian [50]). Similar tendencies can be observed in the medical domain. The language-specific Dutch model, RobBERT, outperformed multilingual BERT (mBERT) on the multilabel classification of chest imaging requests and report items [51]. The Swedish KB-BERT model outperformed mBERT when fine-tuned for the de-identification task, albeit marginally [52].

The characteristics of training and fine-tuning data can strongly affect the outcome of fine-tuning. The large French BERT model CamemBERT [53], fine-tuned on French biomedical articles yielded worse results than the fine-tuned

Table 2 Medical language models in LoE from the years 2020-2022.

Model + Paper	Training + Data	Language	Evaluated on	Availability
ABioNER [41] Boudjellal et al., 2021	AraBERT weights; Arabic biomedical literature	ara	NER	Upon request from the authors.
RAD-BERT [43] Bressem et al., 2020	German BERT weights; 3.8M radiology reports	deu	classification	https://github.com/rAIdiance/bert-for-radiology
GERNERMED++ [64] Frei et al., 2022	Spacy NER model; MT-translated n2c2 NLP 2018 Track2 set with anno projection	deu	NER	https://github.com/frankkramer-lab/GERNERMED-pp
Cross-lingual SAPBERT [71] Liu et al., 2021	UMLS 2020AB	eng, spa, jpn, rus, deu, kor, zho, fin, tur, tha	Cross-lingual BEL	https://github.com/cambridgeltl/sapbert
UTH-BERT [60] Kawazoe et al., 2021	120M clinical texts from the University of Tokyo Hospital	jpn	MLM, NSP, multilabel classification of tweets (NTCIR 13)	https://ai-health.m.u-tokyo.ac.jp/home/research/uth-bert
KM-BERT [42] Kim et al., 2022	KR-BERT weights; medical textbooks, health information news, and medical research articles	kor	MLM, NSP MedSTS, NER	https://github.com/KU-RIAS/KM-BERT-Korean-Medical-BERT
BioBERTpt [54] Schneider et al., 2020	mBERT weights; 1) 2M+ clinical notes 2) Scielo and PubMed	por	NER, SemClinBr	https://github.com/HAILab-PUCPR/BioBERTpt
Romanian BioNER [62] Mitrofan and Pais, 2022	XLM-R + inhibitory layer SiMoNERo corpus with COVID terms	ron	NER	https://relate.racai.ro/index.php? path=lrlt/models
ClinicalCodingTransformerES [138] López-García et al., 2021	Continued training on 30.9K oncology texts, fine-tuned on the CodiEsp corpus	spa	clinical coding	https://github.com/guilopgar/ClinicalCodingTransformerES
bsc-bio [46,48] Carrino et al., 2021-2022	RoBERTa from scratch; 514K clinical docs, 2.5M biomedical texts	spa	NER	https://huggingface.co/PlanTL-GOB-ES
BioBERTurk [44] Turkmen et al., 2022	Biomedical papers and theses on radiology	tur	classification of radiology reports	https://github.com/hazalturkmen/BioBERTurk
Chinese CNER BERT [47] Li et al., 2020	English BERT weights; large Chinese clinical corpus	zho	NER	https://github.com/lxy444/bertcner

mBERT base model. The amount of biomedical fine-tuning data was supposedly insufficient to influence the knowledge gained by CamemBERT during pretraining [12]. The large Portuguese BERT model also performed worse than mBERT in a related study [54]. The authors suspected that "catastrophic forgetting" was the culprit, i.e. fine-tuning the monolingual Portuguese BERT on clinical data might have "erased" the knowledge learned during pretraining.

This might be due to vast differences in the linguistic characteristics of the pretraining and fine-tuning data. The effect of forgetting might be less noticeable in the multilingual model because of the larger variability of its pretraining data.

In summary, there is an undisputed need for better biomedical and clinical representations in many languages. Training and releasing domain-specific LMs for LoE is a great way to fulfill this need. However,

many factors need to be taken into account, including the size, compatibility, and homogeneity of training and fine-tuning data, linguistic transferability of involved languages, optimization of hyperparameters, as well as hardware availability. Combining large generic data with minimal in-domain data when training a model from scratch might not be effective. Yet continual pretraining on a small-sized domain-specific corpus can be the right solution, depending on the
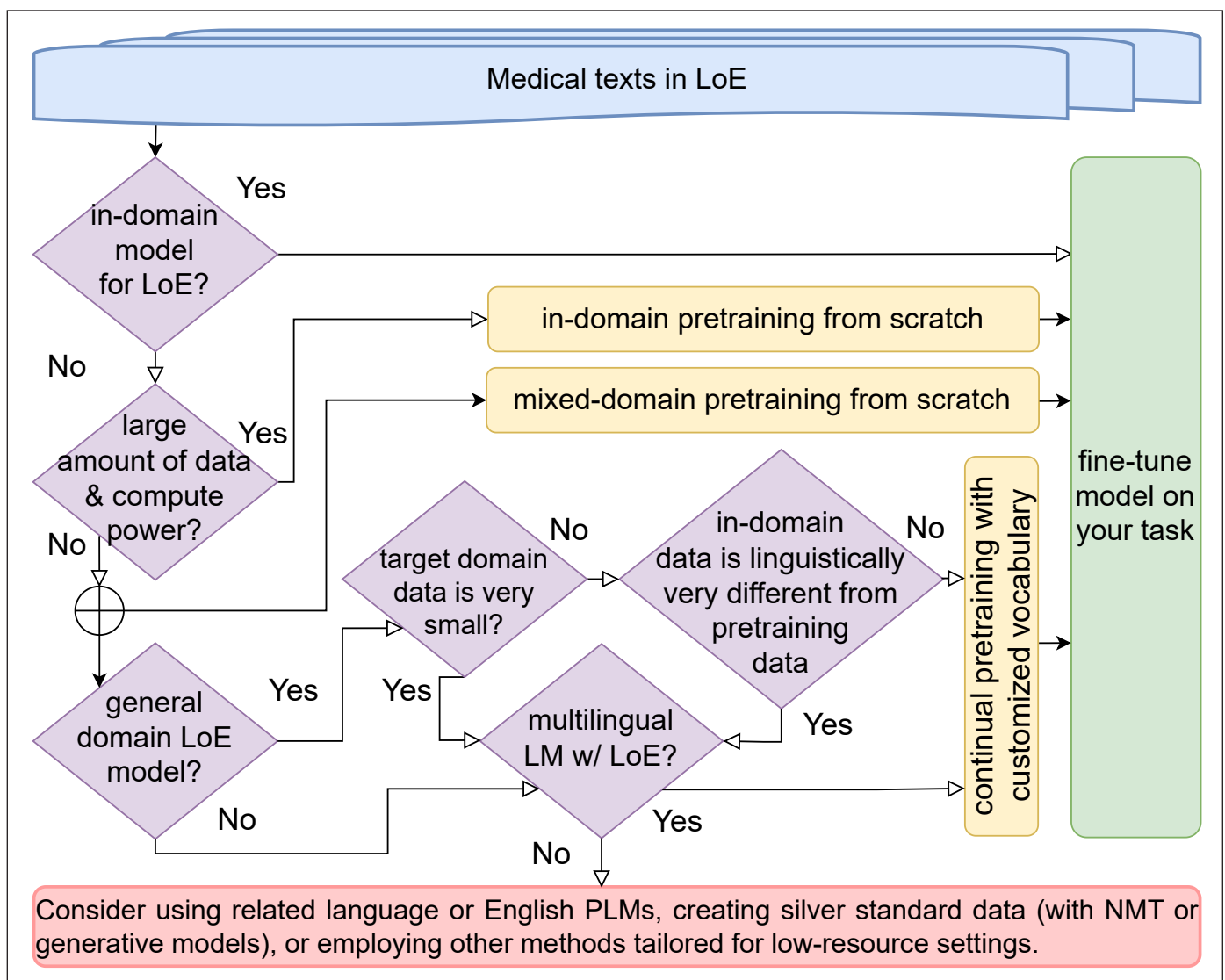

Fig. 1 Tree diagram describing the choice of the optimal type of LM according to the nature of the data.

task. Finally, the size of the vocabulary and the number of parameters impact the carbon footprint of a model's training, which is becoming a commonly reported issue for consideration [12]. Figure 1 provides an overview of the decision-making process for training a model in a new language, informed by insights from the reviewed papers.

3.3 Named Entity Recognition, Normalization, and Linking for LoE

Named Entity Recognition (NER) is a foundational task for efficient information extraction. Unsurprisingly, it is known as "*the most studied task in the biomedical and clinical NLP literature*" [48]. Medical named entities have a more complicated structure than entities in other domains. They represent a wide range of specific information including diseases, symptoms, treatments, drugs, and anatomical concepts. Such a variety of labels complicates NER and calls for domain-specific resources and solutions which are often lacking.

NER has specific challenges in each given language, which requires a variety of language-specific solutions. For instance, capitalization is a boolean morphological feature that may help identify named entities in English or French. This does not apply to German, however, where all nouns are capitalized, or to Semitic languages such as Arabic and Hebrew that do not use capital letters at all. Semitic languages represent vowels as optional diacritics, which increases potential ambiguity. Transliteration from Latin characters into Semitic languages and cross-linguistic correspondences between consonants result in orthographic variations of proper nouns. Bitton *et al.* [55] propose an unsupervised method to build a synthetic dataset linking possible Hebrew transliterations with Unified Medical Language System (UMLS [39]) concepts. This dataset is then used to train a model that maps the transliterated Hebrew entities to the corresponding Concept Unique Identifier (CUI) code. Further, Arabic clinical documents typically contain terms written in the Latin alphabet, such as locations, first names, or proper nouns. These terms can be easily removed (if the task allows it) to process the

documents with a purely Arabic model [36].

Non-segmented languages, such as Japanese, Chinese, or Tibetan, pose a real challenge to proper tokenization and NER. Leveraging multi-level representations can alleviate the loss of semantics caused by erroneous segmentation in non-segmented languages [37, 56]. Wang *et al.* [57] outperformed the current Chinese NER state-of-the-art by navigating character- and word-level trees via different-length paths in order to combine character representations with word and position embeddings.

Despite the constant need for training data, the usual data augmentation methods such as paraphrasing, noising, and translation do not always work for NER. These methods can negatively affect the original semantics, sample diversity, and domain specificity. A novel medical data augmentation (MDA) method based on medical knowledge graphs was evaluated on Chinese NER and relational classification with good results [58].

Transfer learning and pretrained language models, such as mBERT and BERT, have shown considerable success in NER tasks for LoE, often outperforming traditional methods like BiLSTM-CRF and CRF. For example, the mBERT model significantly outperformed the character-level BiLSTM-CRF method for Korean clinical NER [59]. In an extremely low-resource setting, the Japanese medical UTH-BERT [60] provided benefits only for radiology reports, owing to their linguistic similarity to the training data; for other cases, the general BERT model produced more favorable results [29]. These models have also demonstrated their effectiveness in languages such as Portuguese [61] and Romanian [62]. Kepler *et al.* conducted a comprehensive evaluation of state-of-the-art NER methods on Serbian clinical narratives and found that combining existing models in a majority voting ensemble produced the best F1 score of 89.2%, showcasing the potential of hybrid approaches [63].

German clinical NER research has flourished in recent years, owing to the release of several annotated corpora, the development of PLMs, and the availability of high-quality English-German MT systems. The novel pipelines [21, 64] involve

training a model on machine-translated open-access, high-quality, high-resource (English) datasets, thus bypassing potential issues related to privacy, security, and bureaucracy, and projecting annotations with the assistance of neural word alignment tools. The model is then evaluated on either an existing gold standard corpus or a custom-made out-of-distribution dataset. Alternatively, training data can be entirely synthesized by a generative model using a few-shot method.

The NER task is often combined with entity linking, which consists of **Biomedical Entity Linking (BEL)**, or grounding, where a medical term is disambiguated by linking it to a unique concept identifier in a knowledge base or terminology such as the UMLS. Entity linking generally depends on the availability of terminological resources in the target language, which creates problems for LoE. For example, French is one of the most represented LoE in the UMLS metathesaurus, yet only 4% of all the available concepts can be associated with at least one label term in French [65]. One of the most straightforward ways to overcome this challenge has been the use of human or machine translation. Entities in LoE are simply translated into English and then linked to a concept identifier [66], however translating medical jargon poses its own difficulties.

Overall, the resources for BEL in LoE are extremely scarce. According to a recent review on BEL [67], there are only 3 BEL corpora available in LoE: one in French [68] and two in Spanish [69, 70]. Such scarcity of data has a profound negative effect of BEL in LoE. Notwithstanding, deep learning techniques and multilingual representations offer a promising alternative to annotation-dependent NER and BEL approaches in LoE. Leveraging data-driven knowledge of pretrained language models and available multilingual terminologies allows for good results without labeled data or translation [65]. The cross-lingual extension of SAPBERT (the model that currently achieves state-of-the-art results in BEL) pools all the synonyms in the UMLS that share the same CUI to obtain similar representations, regardless of the language [71].

3.4 Recent Advances in Negation Resolution for LoE

Negation is a complex linguistic phenomenon that changes the meaning of an utterance. It plays a crucial role in biomedical text mining since its identification can greatly affect the quality and veracity of retrieved information. Negation exhibits great diversity in its syntactic and morphological representation across languages which poses additional challenges to the use of such methods as transfer learning [72]. The lack of annotated data hampers research on negation even more than its linguistic particularities. Before 2020, there were nine LoE resources out of a total of 19 publicly available negation-annotated corpora [73, 74]. However, only two languages were annotated for the medical domain: the IULA Spanish Clinical Record Corpus [75] and the two corpora of clinical texts in French [27, 76]. Four additional corpora have been released since then: a cancer dataset [77] and a clinical corpus of negation and uncertainty [78] in Spanish, a multipurpose clinical corpus in Brazilian Portuguese [79], and a German corpus of discharge summaries [26] (Table 1).

Despite the advancement of transfer learning for negation detection [76, 79, 80], rule-based [27] and supervised machine learning approaches [76, 77, 81–83] for LoE continue to be researched and employed. One paper presented a corpus-free approach, which is an attractive prospect in a scenario where there is no annotated data [84]. However, this work relied on gazetteers and the similarity of anamneses. A particularly interesting and challenging aspect of negation detection is the resolution of its scope, i.e. the identification of all negated tokens in a sentence. We found six publications dedicated to biomedical negation scope resolution in LoE. Two papers explore zero-shot transfer-learning methods across languages and domains for French and Spanish [80, 85] with the best F1 score of 90.8% for Spanish. A supervised BiLSTM-based method achieves 84.8% in Brazilian Portuguese [83] and 90.0% in Spanish [78]. Overall, despite some reports on the inapplicability of negation detection across domains [74], the deep learning methods

prove to be feasible for transferring negation scope knowledge across corpora and languages [77, 80, 85, 86].

3.5 Shared Tasks

A major problem in the field of medical NLP is that contributions and results are often made using private, non-shareable, and ethically sensitive clinical data. This problem leads to a lack of common public benchmarks to use in order to evaluate and compare the quality of new works. This is even more true for data in LoE. To overcome data availability issues, shared tasks provide non-sensitive data extracted through public sources such as social media [87], or medical literature [27, 88]. In other cases, datasets might contain synthetic examples [70]. Furthermore, these datasets remain available to the public beyond the challenge time frame, making them solid medical NLP benchmarks [89]. Besides, such annotated datasets spare time and expenses for other research groups because the creation of an annotated gold standard requires months of work by trained domain experts [90, 91].Our query revealed a number of medical NER challenges for the Spanish language (Table 3), including CLEF eHealth (2020-21) [92–101], IberLEF (2020-22) [17, 70, 102–115], and CLEF BioASQ (2022) [116–119]. Furthermore, IberLEF 2022 [17] and BioASQ 2022 [116] released their datasets translated into seven other languages, encouraging future contributions to multilingual medical NLP. For the French language, the CAS corpus [27] was used in DEFT [53, 120–124], an annual French-speaking text-mining challenge. The 2020 edition of DEFT involved the automatic annotation of 13 different medical entity types, while the 2021 edition proposed to identify the patient's clinical profile through multilabel classification of diseases using the Medical Subject Headings (MeSH) thesaurus. In the same way as DEFT 2020, NTCIR 16 challenge [29, 125] consisted in finding a wide variety of medical entity types for the Japanese language.

Shared tasks provide excellent opportunities to test recently published PLMs and build new ones. For example, the Spanish

Medical RoBERTa model, released in 2021 [48], was tested and validated in the IberLEF 2022 [17], BioASQ 2022 [116], and SocialDisNER 2022 [87, 126–130] challenges. UTH-BERT, a model pretrained on Japanese clinical text published in 2021 [60], was among the top models for the Japanese NER task of NTCIR 16. Alternatively, some LMs were built specifically for a shared task:

- IberLEF 2020 [115]: XLM-RoBERTa [131] + Galén dataset (private real-world de-identified clinical cases in Spanish), mBERT + Galén, BETO [132] + Galén [133];
- DEFT 2020 [120]: CamemBERT [134] + French abstracts extracted from PubMed [121];
- DEFT 2021 [123]: FlauBERT [135] + corpus of medical emergencies cases obtained through house calls (SOS Médecins) [124].

Multilingual models such as mBERT and XLM-RoBERTa are frequently used in non-English language challenges. Due to domain constraints, English models might perform better than the available language-specific models, like in the case of the 16th NTCIR NER task. We have observed in shared tasks the tendency of creating a medical monolingual model when one is not available. This is often done through continual pretraining of existing models.

3.6 The Special Case of Spanish

Our survey uncovered a large increase in the biomedical NLP resources for the Spanish language. Spanish is one of the most spoken languages in the world, yet it lacked linguistic data and trained models up until very recently. BETO [132], the first publicly available BERT-based general-domain monolingual Spanish model, was released in 2020. Several additional general-domain Spanish LMs emerged since then, e.g. IXAmBERT [136] (also deals with Basque and English), and MarIA, an entire family of Spanish LMs based on RoBERTa and GPT2 [137].

López-García *et al.* obtained clinical language models in Spanish through the continual pretraining of the mBERT, BETO,

Table 3 Shared tasks in LoE related to clinical NLP from challenges in 2020-2022. Task: NER = Named Entity Recognition, EL = Entity Linking, MLC = Multi-Label Classification. Models: X = multilingual LM, I = monolingual LM pretrained on LoE, E = English LM.

Shared Task	Language	Tasks	Employed Models*
BioNLP 2019 / PharmacoNER	spa	NER, EL, MLC	XIE
CLEF eHealth 2020	spa	NER, EL,	XI
CLEF eHealth 2021 / SpRadIE	spa	NER	XI
SocialDisNER 2022	spa	NER	XI
IberLEF 2020 / CANTEMIST	spa	NER, EL	XI
IberLEF 2020 / eHealthKD	spa	NER	XI
IberLEF 2021	spa	NER	XI
IberLEF 2022 / LivingNER	spa	NER	XI
CLEF 2022 BioASQ / DisTEMIST	spa	NER, EL	XI
DEFT 2020	fra	NER	XIE
DEFT 2021	fra	NER, EL, MLC	XI
NTCIR 16 (2022) / Real-MedNLP	jpn	NER	XIE

* The column Employed Models refers to all models participating in the task

and XLM-RoBERTa models [138]. Carrino *et al.* released several Spanish models based on RoBERTa, evaluating mixed-domain pre-training [48] and training from scratch [46].

The published research on Spanish now dominates biomedical NLP papers concerning LoE. Indeed, based on the queries detailed in the Methods section, 59 out of 97 articles related to NER in LoE involve the Spanish language. In the case of negation papers, 9 out of 25 were about Spanish. Furthermore, the queried articles were related to 12 different shared tasks, with 9 of them using Spanish data (see Table 3).

The growth in quantity and quality of Spanish language resources could be attributed to investments made by the Spanish government through initiatives such as the Digital Agenda for Spain and the Spanish Strategy for Science, Technology, and Innovation. Among their various objectives, these initiatives specifically target information communication technology for support, striving to establish Spain as a global leader in language technology and innovation[3]. Specifically, the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL)[4], particularly from 2018 onwards, included a dedicated flagship project focused on Health and Biomedicine. The promoted shared tasks were directly related to high-impact biomedical use cases and healthcare data exploitation scenarios. More than 15 healthcare-related shared tasks with data in Spanish were organized over the past five years. The shared tasks promoted by the Plan TL stimulated global participation and research in biomedical NLP beyond research groups located in Spanish-speaking countries, as evidenced by the diverse teams participating from around the world.

4 Discussion and Conclusion

In recent years, novel NLP technologies have revolutionized various research areas dealing with text analysis, including biomedical and clinical text mining. This survey clearly indicates the trend, demonstrating that significant changes have taken place between 2020 and 2022. The period started with the dominance of transformer-based models, beginning with BERT, and currently, we witness the introduction and implementation of very large models that can achieve feats previously thought impossible to reach, such as GPT-3 [139]. These powerful models provide new opportunities, while accentuating the resource gap between different research labs. In fact, the release of the ChatGPT[5] conversational agent in November 2022 sent shockwaves through many industries, including the medical field [140]. Given its recent emergence, we leave the analysis of this extremely dynamic and highly impactful development to a future survey.

Medical NLP has undertaken new paradigms for processing data outside of the English language. The training of language- and domain-specific models is on the rise. Most often biomedical models for LoE are further trained or fine-tuned using the weights of their language-specific general domain counterparts. However, several factors must be considered when training a given model, such as the size and domain composition of the training data as well as its linguistic compatibility with the target texts (see Figure 1). Finally, while monolingual models often outperform their multilingual counterparts, the latter should always be considered when dealing with low-resource languages.

Another interesting trend of the past few years is the increased use of generated texts in the creation of corpora. The advanced quality of NMT allows for the creation of large, multilingual parallel silver standard corpora, such as LivingNER. For high-resource languages like German, it is now possible to bypass many constraints associated with data sensitivity and sparsity of annotated resources. High-quality open datasets can be automatically translated, and annotations projected using neural aligners. Furthermore, the GPT family of models offers an opportunity to create training data with just a few well-engineered prompts. Nevertheless, validation with gold-standard data remains essential.

Negation detection and NER are foundational tasks in biomedical NLP that require language-specific awareness and customized methods. For instance, lexical, morphological, and semantic characteristics of many non-Indo-European languages may be too specific to follow an NLP pipeline

---

[3] https://plantl.mineco.gob.es/tecnologias-lenguaje/PTL/Paginas/plan-impulso-tecnologias-lenguaje.aspx

[4] Plan for Advancement of Language Technologies

[5] https://chat.openai.com/chat

designed for English. Nevertheless, both tasks are often solved using PLMs and transfer-learning. Entity linking and negation scope resolution still suffer from the lack of annotated data in many LoE. It would be reasonable to investigate the power of large multilingual generative models to start closing that gap.

Overall, there is a noticeable acceleration in biomedical NLP research for LoE. We determined that significantly more papers were dedicated to this area of research in the last two years than over the previous decade [4], presenting new labeled corpora, trained models, and shared tasks. All these new resources provide an opportunity for the creation of new comprehensive benchmarks in LoE. The specific advances seen for the Spanish language, likely the result of targeted research funding and support, is a particularly good example of the current dynamic research environment that can lead to considerable advances in a limited period of time.

NLP techniques have the potential to render medical documents much more valuable for primary and secondary research. They can be used to make specific clinical documentation more easily accessible for medical practitioners (primary usage), and once de-identified and aggregated they can be used to derive previously unseen correlations among medical events, such as between specific treatments and outcomes (secondary usage). The fact that NLP tools for the medical domain are increasingly being developed in LoE goes a long way to enable their usability across different countries and language communities. The widespread usage of such technologies in turn has the potential to enable large-scale evidence-based medical research, which will have positive outcomes on public health.

Acknowledgements

References

1. Zhou B, Yang G, Shi Z, Ma S. Natural Language Processing for Smart Healthcare. IEEE Rev Biomed Eng 2022;1–17. doi:10.1109/RBME.2022.3210270.
2. Aramaki E, Wakamiya S, Yada S, Nakamura Y. Natural Language Processing: from Bedside to Everywhere. Yearb Med Inform 2022 Jun 2; doi:10.1055/s-0042-1742510.
3. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, et al. Neural Natural Language Processing for unstructured data in electronic health records: A review. Comput Sci Rev 2022 Nov 1;46:100511. doi:10.1016/j.cosrev.2022.100511.
4. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical Natural Language Processing in languages other than English: opportunities and challenges. J Biomed Semant 2018 Mar 30;9(1):12. doi:10.1186/s13326-018-0179-8.
5. Walpole SC. Including papers in languages other than English in systematic reviews: important, feasible, yet often omitted. J Clin Epidemiol 2019 Jul 1;111:127–34. doi:10.1016/j.jclinepi.2019.03.004.
6. Dalianis H. Characteristics of Patient Records and Clinical Corpora. Dalianis H, editor. Clinical Text Mining: Secondary Use of Electronic Patient Records. Cham: Springer International Publishing; 2018. p. 21–34. doi:10.1007/978-3-319-78503-5_4.
7. Soares F, Yamashita GH. On the crucial role of multilingual biomedical databases in epidemic events (SARS-CoV-2 analysis). Int J Infect Dis 2020 Jul;96:352–4. doi:10.1016/j.ijid.2020.05.023.
8. Grabar N, Grouin C. Year 2020 (with COVID): Observation of Scientific Literature on Clinical Natural Language Processing. Yearb Med Inform 2021 Aug;30(1):257–63. doi:10.1055/s-0041-1726528.
9. Laparra E, Mascio A, Velupillai S, Miller T. A Review of Recent Work in Transfer Learning and Domain Adaptation for Natural Language Processing of Electronic Health Records. Yearb Med Inform 2021 Aug;30(1):239–44. doi:10.1055/s-0041-1726522.
10. Yang F, Wang X, Ma H, Li J. Transformers-sklearn: a toolkit for medical language understanding with transformer-based models. BMC Med Inform Decis Mak 2021 Jul 30;21(Suppl 2):90. doi:10.1186/s12911-021-01459-0.
11. Jati BS, Widyawan S, Muhammad Nur Rizal ST. Multilingual Named Entity Recognition Model for Indonesian Health Insurance Question Answering System. Proceedings of the 3rd International Conference on Information and Communications Technology (ICOIACT). 2020. p. 180–4. doi:10.1109/ICOIACT50329.2020.9332027.
12. Gérardin C, Wajsbürt P, Vaillant P, Bellamine A, Carrat F, Tannier X. Multilabel classification of medical concepts for patient clinical profile identification. Artif Intell Med 2022 Jun;128:102311. doi:10.1016/j.artmed.2022.102311.
13. Wang B, Xie Q, Pei J, Tiwari P, Li Z, Fu J. Pre-trained Language Models in Biomedical Domain: A Systematic Survey. arXiv; 2021. Available at: http://arxiv.org/abs/2110.05006.
14. AlShuweihi M, Salloum SA, Shaalan K. Biomedical Corpora and Natural Language Processing on Clinical Text in Languages Other Than English: A Systematic Review. In: Al-Emran M, Shaalan K, Hassanien AE, editors. Recent Advances in Intelligent Systems and Smart Applications. Cham: Springer International Publishing; 2021. p. 491–509. (Studies in Systems, Decision and Control). doi:10.1007/978-3-030-47411-9_27.
15. Ge Y, Guo Y, Yang YC, Al-Garadi MA, Sarker A. Few-shot learning for medical text: A systematic review. arXiv; 2022. Available at: https://arxiv.org/abs/2204.14081.
16. Magnini B, Altuna B, Lavelli A, Speranza M, Zanoli R. The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases. Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 : Bologna, Italy, March 1-3, 2021. Torino: Accademia University Press; 2021. p. 258–64. (Collana dell'Associazione Italiana di Linguistica Computazionale). doi:10.4000/books.aaccademia.8663.
17. Miranda-Escalada A, Farre-Maduell E, Lima-Lopez S, Estrada D, Gasco L, Krallinger M. Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of the LivingNER shared task and resources. Procesamiento del Lenguaje Natural 2022;69:241–53. doi:10.26342/2022-69-21.
18. Miranda-Escalada, Antonio, Eulàlia Farré, Gasco L, Lima S, Krallinger M. DisTEMIST corpus: detection and normalization of disease mentions in Spanish clinical cases. Zenodo; 2022. doi:10.5281/ZENODO.6408476.
19. Blinov P, Nesterov A, Zubkova G, Reshetnikova A, Kokh V, Shivade C. RuMedNLI: A Russian Natural Language Inference Dataset For The Clinical Domain. PhysioNet; 2022. doi:10.13026/gxzd-cf80.
20. Shivade C. MedNLI - A Natural Language Inference Dataset For The Clinical Domain. PhysioNet; 2017. doi:10.13026/C2RS98
21. Frei J, Kramer F. GERNERMED -- An Open German Medical NER Model. arXiv; 2021. Available at: http://arxiv.org/abs/2109.12104.
22. Frei J, Kramer F. Annotated Dataset Creation through General Purpose Language Models for non-English Medical NLP. arXiv; 2022. Available at: http://arxiv.org/abs/2208.14493.
23. Modersohn L, Schulz S, Lohr C, Hahn U. GraSCCo-The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus. Stud Health Technol Inform 2022;296:66–72. doi:10.3233/SHTI220805.
24. Borchert F, Lohr C, Modersohn L. GGPONC: A Corpus of German Medical Text with Rich Metadata Based on Clinical Practice Guidelines. arXiv; 2020. Available at: https://arxiv.org/abs/2007.06400.
25. Borchert F, Lohr C, Modersohn L, Witt J, Langer T, Follmann M, et al. GGPONC 2.0-the German clinical guideline corpus for oncology: Curation workflow, annotation policy, baseline NER tag-

gers. Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022. p. 3650–60. Available at: https://aclanthology.org/2022.lrec-1.389/.

26. 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 Apr 1;4(2):ooab025. doi:10.1093/jamiaopen/ooab025.

27. Grabar N, Dalloux C, Claveau V. CAS: corpus of clinical cases in French. J Biomed Semant 2020 Aug 6;11(1):7. doi:10.1186/s13326-020-00225-x.

28. Hiebel N, Ferret O, Fort K, Névéol A. CLIS-TER : A Corpus for Semantic Textual Similarity in French Clinical Narratives. Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association; 2022. p. 4306–15. Available at: https://aclanthology.org/2022.lrec-1.459.

29. Yada S, Nakamura Y, Wakamiya S, Aramaki E. Real-MedNLP: Overview of REAL document-based MEDical Natural Language processing Task. Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies NII. 2022.

30. Kim YM, Lee TH. Korean clinical entity recognition from diagnosis text using BERT. BMC Med Inform Decis Mak 2020 Sep 30;20(7):242. doi:10.1186/s12911-020-01241-8.

31. Kim YM, Lee TH, Na SO. Constructing novel datasets for intent detection and ner in a korean healthcare advice system: guidelines and empirical results. Appl Intell 2022;53(1):1–21. doi:10.1007/s10489-022-03400-y.

32. Sazzed S. BanglaBioMed: A Biomedical Named-Entity Annotated Corpus for Bangla (Bengali). Proceedings of the 21st Workshop on Biomedical Language Processing. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 323–9. doi:10.18653/v1/2022.bionlp-1.31.

33. Van Nguyen K, Van Huynh T, Nguyen DV, Nguyen AGT, Nguyen NLT. New Vietnamese Corpus for Machine Reading Comprehension of Health News Articles. ACM Trans Asian Low-Resour Lang Inf Process 2022 Sep 23;21(5):105:1-105:28. doi:10.1145/3527631.

34. Boudjellal N, Zhang H, Khan A, Ahmad A, Naseem R, Dai L. A Silver Standard Biomedical Corpus for Arabic Language. Complexity 2020 Oct 9;2020:e8896659. doi:10.1155/2020/8896659.

35. Cherednichenko O, Kanishcheva O, Yakovleva O, Arkatov D. Collection and Processing of a Medical Corpus in Ukrainian. Proceedings of the 4th International Conference on Computational Linguistics and Intelligent Systems. Lviv: CEUR; 2020. Available at: https://ceur-ws.org/Vol-2604/paper21.pdf.

36. Hammoud J, Vatian A, Dobrenko N, Vedernikov N, Shalyto A, Gusarova N. New Arabic Medical Dataset for Diseases Classification. arXiv; 2021. Available at: http://arxiv.org/abs/2106.15236.

37. Zhuoma C, Cairang J, Sangjie D, Yangmao Z, Zhuoma Z. Tibetan Medical Named Entity Recognition Study for Tibetan Clinical Electronic Medical Records. SSRN; 2022 Feb 22; doi: 10.2139/ssrn.4040676.

38. Zaghir J, Goldman JP, Bjelogrlic M, Keszthelyi D, Gaudet-Blavignac C, Turbé H, et al. Performance of Machine Learning Methods to Classify French Medical Publications. Stud Health Technol Inform 2022;294:874–5. doi:10.3233/SHTI220613.

39. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267-70. doi: 10.1093/nar/gkh061.

40. Lewis P, Ott M, Du J, Stoyanov V. Pretrained Language Models for Biomedical and Clinical Tasks: Understanding and Extending the State-of-the-Art. Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: Association for Computational Linguistics; 2020. p. 146–57. doi:10.18653/v1/2020.clinicalnlp-1.17.

41. Boudjellal N, Zhang H, Khan A, Ahmad A, Naseem R, Shang J, et al. ABioNER: a BERT-based model for Arabic biomedical named-entity recognition. Complexity 2021;2021.

42. Kim Y, Kim JH, Lee JM, Jang MJ, Yum YJ, Kim S, et al. A pre-trained BERT for Korean medical natural language processing. Sci Rep 2022;12(1):1–10.

43. Bressem KK, Adams LC, Gaudin RA, Tröltzsch D, Hamm B, Makowski MR, et al. Highly accurate classification of chest radiographic reports using a deep learning natural language model pre-trained on 3.8 million text reports. Bioinformatics 2020 Nov 1;36(21):5255–61. doi:10.1093/bioinformatics/btaa668.

44. Turkmen H, Dikenelli O, Eraslan C, Callı MC. Bioberturk: Exploring Turkish Biomedical Language Model Development Strategies in Low Resource Setting. Research Square; 2022. doi:10.21203/rs.3.rs-2165226/v1.

45. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Trans Comput Healthc 2022 Jan 31;3(1):1–23. doi:10.1145/3458754.

46. Carrino CP, Llop J, Pàmies M, Gutiérrez-Fandiño A, Armengol-Estapé J, Silveira-Ocampo J, et al. Pretrained Biomedical Language Models for Clinical NLP in Spanish. Proceedings of the 21st Workshop on Biomedical Language Processing. Dublin, Ireland: Association for Computational Linguistics; 2022. p. 193–9. doi:10.18653/v1/2022.bionlp-1.19.

47. Li X, Zhang H, Zhou XH. Chinese clinical named entity recognition with variant neural structures based on BERT methods. J Biomed Inform 2020;107:103422. doi:10.1016/j.jbi.2020.103422.

48. Carrino CP, Armengol-Estapé J, Gutiérrez-Fandiño A, Llop-Palao J, Pàmies M, Gonzalez-Agirre A, et al. Biomedical and Clinical Language Models for Spanish: On the Benefits of Domain-Specific Pretraining in a Mid-Resource Scenario. arXiv; 2021. Available at: http://arxiv.org/abs/2109.03570.

49. de Vries W, van Cranenburgh A, Bisazza A, Caselli T, van Noord G, Nissim M. BERTje: A Dutch BERT Model. arXiv; 2019. Available at: https://arxiv.org/abs/1912.09582.

50. Tanvir H, Kittask C, Eiche S, Sirts K. Est-BERT: A Pretrained Language-Specific BERT for Estonian. Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa). Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden; 2021. p. 11–9. Available at: https://aclanthology.org/2021.nodalida-main.2.

51. Olthof AW, van Ooijen PM, Cornelissen LJ. The natural language processing of radiology requests and reports of chest imaging: Comparing five transformer models' multilabel classification and a proof-of-concept study. Health Informatics J 2022 Dec;28(4):14604582221131198. doi:10.1177/14604582221131198.

52. Grancharova M, Dalianis H. Applying and sharing pre-trained BERT-models for named entity recognition and classification in Swedish electronic patient records. Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa). 2021. p. 231–9. Available at: https://aclanthology.org/2021.nodalida-main.23/.

53. Bailly A, Blanc C, Guillotin T. Classification multi-label de cas cliniques avec CamemBERT (Multi-label classification of clinical cases with CamemBERT). Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles Atelier DÉfi Fouille de Textes (DEFT). 2021. p. 14–20. Available at: https://aclanthology.org/2021.jeptalnrecital-deft.2/.

54. Schneider ETR, de Souza JVA, Knafou J, Silva e Oliveira LE, Copara J, Gumiel YB, et al. BioBERTpt - A Portuguese Neural Language Model for Clinical Named Entity Recognition. Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: Association for Computational Linguistics; 2020. p. 65–72. doi:10.18653/v1/2020.clinicalnlp-1.7.

55. Bitton Y, Cohen R, Schifter T, Bachmat E, Elhadad M, Elhadad N. Cross-lingual Unified Medical Language System entity linking in online health communities. J Am Med Inform Assoc 2020 Sep 10;27(10):1585–92. doi:10.1093/jamia/ocaa150.

56. Johnson A, Karanasou P, Gaspers J, Klakow D. Cross-lingual Transfer Learning for Japanese Named Entity Recognition. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers). 2019. p. 182–9. doi:10.18653/v1/N19-2023.

57. Wang C, Wang H, Zhuang H, Li W, Han S, Zhang H, et al. Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree. J Biomed Inform 2020 Nov;111:103583. doi:10.1016/j.jbi.2020.103583.

58. Shi B, Zhang L, Huang J, Zheng H, Wan J, Zhang L. MDA: An Intelligent Medical Data Augmentation Scheme Based on Medical Knowledge Graph for Chinese Medical Tasks. Appl Sci 2022;12(20):10655. doi: 10.3390/app122010655.

59. Kim YM, Lee TH. Korean clinical entity recognition from diagnosis text using BERT. BMC Med Inform Decis Mak 2020 Sep 30;20(7):242.

doi:10.1186/s12911-020-01241-8.

60. Kawazoe Y, Shibata D, Shinohara E, Aramaki E, Ohe K. A clinical specific BERT developed using a huge Japanese clinical text corpus. PLoS One 2021 Nov 9;16(11):e0259763. doi:10.1371/journal.pone.0259763.

61. de Souza JVA, Schneider ETR, Cezar JO, Silva LE, Gumiel YB, Paraiso EC, et al. A multilabel approach to Portuguese clinical named entity recognition. J Health Inform 2020;12.

62. Mitrofan M, Pais V. Improving Romanian BioNER Using a Biologically Inspired System. Proceedings of the 21st Workshop on Biomedical Language Processing. Association for Computational Linguistics; 2022. p. 316–22. doi:10.18653/v1/2022.bionlp-1.30.

63. Kaplar A, Stošović M, Kaplar A, Brković V, Naumović R, Kovačević A. Evaluation of clinical named entity recognition methods for Serbian electronic health records. Int J Med Inf 2022 Aug;164:104805. doi:10.1016/j.ijmedinf.2022.104805.

64. Frei J, Frei-Stuber L, Kramer F. GERNERMED++: Transfer Learning in German Medical NLP. ArXiv; 2022. Available at: https://arxiv.org/abs/2206.14504.

65. Wajsbürt P, Sarfati A, Tannier X. Medical concept normalization in French using multilingual terminologies and contextual embeddings. J Biomed Inform 2021 Feb 1;114:103684. doi:10.1016/j.jbi.2021.103684.

66. Budiarti RPN, Sukaridhoto S, Al-Hafidz IA, Satrio NA. Symptoms identification of ICD-11 based on clinical NLP mobile apps for diagnosing the disease (ICD-11). Bali Med J 2022;11(3):1162–7.

67. French E, McInnes BT. An overview of biomedical entity linking throughout the years. J Biomed Inform 2022 Dec 2;104252. doi:10.1016/j.jbi.2022.104252.

68. Névéol A, Grouin C, Leixa J, Rosset S, Zweigenbaum P. The Quaero French Medical Corpus: A Resource for Medical Entity Recognition and Normalization. Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing. 2014. p. 24–30.

69. Gonzalez-Agirre A, Marimon M, Intxaurrondo A, Rabal O, Villegas M, Krallinger M. PharmaCoNER: Pharmacological Substances, Compounds and proteins Named Entity Recognition track. Proceedings of the 5th Workshop on BioNLP Open Shared Tasks. Hong Kong, China: Association for Computational Linguistics; 2019. p. 1–10. doi:10.18653/v1/D19-5701.

70. Miranda-Escalada A, Farré E, Krallinger M. Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish, Corpus, Guidelines, Methods and Results. Proceedings of the Iberian Languages Evaluation Forum. 2020;303–23.

71. Liu F, Vuli I, Korhonen A, Collier N. Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2021. p. 565–74. doi:10.18653/v1/2021.acl-short.72.

72. Shaitarova A, Rinaldi F. Negation typology and general representation models for cross-lingual zero-shot negation scope resolution in Russian, French, and Spanish. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. 2021. p. 15–23. doi:10.18653/v1/2021.naacl-srw.3.

73. Jiménez-Zafra SM, Morante R, Martín-Valdivia MT, Ureña-López LA. Corpora Annotated with Negation: An Overview. Comput Linguist 2020;46(1):1–52. doi:10.1162/coli_a_00371.

74. Mahany A, Khaled H, Elmitwally NS, Aljohani N, Ghoniemy S. Negation and Speculation in NLP: A Survey, Corpora, Methods, and Applications. Appl Sci 2022 Jan;12(10):5209. doi:10.3390/app12105209.

75. Marimon M, Vivaldi J, Bel N. Annotation of negation in the IULA Spanish Clinical Record Corpus. Proceedings of the Workshop Computational Semantics Beyond Events and Roles. Association for Computational Linguistics; 2017. p. 43–52. doi:10.18653/v1/W17-1807.

76. Dalloux C, Claveau V, Grabar N. Speculation and Negation detection in French biomedical corpora. Proceedings of the International Conference on Recent Advances in Natural Language Processing. INCOMA Ltd.; 2019. p. 223–32. doi:10.26615/978-954-452-056-4_026.

77. Pabón OS, 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 Comput Sci 2022 Mar 7;8:e913. doi:10.7717/peerj-cs.913.

78. Lima Lopez S, Perez N, Cuadros M, Rigau G. NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. Proceedings of the Twelfth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association; 2020. p. 5772–81. Available at: https://aclanthology.org/2020.lrec-1.708.

79. Oliveira LESE, Peters AC, da Silva AMP, Gebeluca CP, Gumiel YB, Cintho LMM, et al. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. J Biomed Semant 2022 May 8;13(1):13. doi:10.1186/s13326-022-00269-1.

80. Shaitarova A, Furrer L, Rinaldi F. Cross-lingual transfer-learning approach to negation scope resolution. Proceedings of the Swiss Text Analytics Conference & Conference on Natural Language Processing. 2020 Jun 25; doi:10.5167/UZH-197355.

81. Mirzapour M, Abdaoui A, Tchechmedjiev A, Digan W, Bringay S, Jonquet C. French FastContext: A publicly accessible system for detecting negation, temporality and experiencer in French clinical notes. J Biomed Inform 2021 May 1;117:103733. doi:10.1016/j.jbi.2021.103733.

82. Santiso S, Pérez A, Casillas A, Oronoz M. Neural negated entity recognition in Spanish electronic health records. J Biomed Inform 2020 May;105:103419. doi:10.1016/j.jbi.2020.103419.

83. Dalloux C, Claveau V, Grabar N, Oliveira LES, Moro CMC, Gumiel YB, et al. Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. Nat Lang Eng 2020;1–21. doi:10.1017/S1351324920000352.

84. Funkner A, Balabaeva K, Kovalchuk S. Negation Detection for Clinical Text Mining in Russian. Digit Pers Health Med 2020;342–6. doi:10.3233/SHTI200179.

85. Hartmann M, Søgaard A. Multilingual Negation Scope Resolution for Clinical Text. Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis. Association for Computational Linguistics; 2021. p. 7–18. Available at: https://aclanthology.org/2021.louhi-1.2.

86. Rivera Zavala 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 Dec 3;8(12):e18953. doi:10.2196/18953.

87. Gasco L, Estrada-Zavala D, Farré-Maduell E, Lima-López S, Miranda-Escalada A, Krallinger M. Overview of the SocialDisNER shared task on detection of diseases mentions from healthcare related and patient generated social media content: methods, evaluation and corpora. Proceedings of the Seventh Social Media Mining for Health (# SMM4H) Workshop and Shared Task. 2022. Available at: https://aclanthology.org/2022.smm4h-1.48/.

88. Lima-López S, Farré-Maduell E, Miranda-Escalada A, Brivá-Iglesias V, Krallinger M. NLP applied to occupational health: MEDDOPROF shared task at IberLEF 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. Procesamiento del Lenguaje Natural 2021;67:243–56. doi:10.26342/2021-67-21.

89. Chapman WW, Nadkarni PM, Hirschman L, D'avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc 2011 Sep-Oct;18(5):540-3. doi:10.1136/amiajnl-2011-000465.

90. Deleger L, Li Q, Lingren T, Kaiser M, Molnar K, Stoutenborough L, et al. Building gold standard corpora for medical natural language processing tasks. AMIA Annu Symp Proc 2012:144-53.

91. Wissler L, Almashraee M, Díaz DM, Paschke A. The Gold Standard in Corpus Annotation. Proceedings of the 5th IEEE Germany Student Conference. 2014;21. doi:10.13140/2.1.4316.3523.

92. Miranda-Escalada A, Gonzalez-Agirre A, Armengol-Estapé J, Krallinger M. Overview of Automatic Clinical Coding: Annotations, Guidelines, and Solutions for non-English Clinical Cases at CodiEsp Track of CLEF eHealth 2020. CLEF (Working Notes). 2020.

93. Costa J, Lopes I, Carreiro AV, Ribeiro D, Soares C. Fraunhofer AICOS at CLEF eHealth 2020 Task 1: Clinical Code Extraction From Textual Data Using Fine-Tuned BERT Models. CLEF (Working Notes). 2020.

94. Perea-Ortega JM, López-Úbeda P, Díaz-Galiano

MC, Valdivia MTM, López LAU. SINAI at CLEF eHealth 2020: Testing Different pre-trained Word Embeddings for Clinical Coding in Spanish. CLEF (Working Notes). 2020.

95. Cossin S, Jouhet V. IAM at CLEF eHealth 2020: Concept Annotation in Spanish Electronic Health Records. CLEF (Working Notes). 2020.

96. Mayya V, Kamath SS, Sugumaran V. LAT A − Label Attention Transformer Architectures for ICD-10 Coding of Unstructured Clinical Notes. Proceedings of the IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology. 2021. p. 1–7.

97. García-Santa N, Cetina K, Cappellato L, Eickhoff C, Ferro N, Nevéol A. FLE at CLEF eHealth 2020: Text Mining and Semantic Knowledge for Automated Clinical Encoding. CLEF (Working Notes). 2020.

98. de la Iglesia I, Martínez-Puente M, Platas A, San Miguel I, Atutxa A, Gojenola K. Media team: Clef-2020 ehealth task 1: Multilingual information extraction-codiesp. CLEF (Working Notes). 2020.

99. Cotik V, Alemany LA, Filippo D, Luque F, Roller R, Vivaldi J, et al. Overview of CLEF eHealth Task 1-SpRadIE: A challenge on information extraction from Spanish Radiology Reports. CLEF (Working Notes). 2021. p. 732–50.

100. Solarte-Pabón O, Montenegro O, Blazquez-Herranz A, Saputro H, Rodriguez-González A, Menasalvas E. Information extraction from Spanish radiology reports using multilingual BERT. CLEF Ehealth. 2021. p. 834–45.

101. Fabregat H, Duque A, Araujo L, Martínez-Romo J. LSI_UNED at CLEF eHealth2021: Exploring the effects of transfer learning in negation detection and entity recognition in clinical texts. CLEF (Working Notes). 2021. p. 780–93.

102. Ruas P, Neves A, Andrade VD, Couto FM, Aragón ME. LasigeBioTM at CANTEMIST: Named Entity Recognition and Normalization of Tumour Morphology Entities and Clinical Coding of Spanish Health-related Documents. Proceedings of the Iberian Languages Evaluation Forum. 2020. p. 422–37.

103. Osborne JD, O'Leary T, Del Monte J, Sasse K. Identification of Cancer Entities in Clinical Text Combining Transformers with Dictionary Features. Proceedings of the Iberian Languages Evaluation Forum. 2020. p. 458–67.

104. Han JC, Tsai RTH. NCU-IISR: Pre-trained Language Model for CANTEMIST Named Entity Recognition. Proceedings of the Iberian Languages Evaluation Forum. 2020. p. 347–51.

105. López-Úbeda P, Díaz-Galiano MC, Martín-Valdivia MT, López LAU. Extracting Neoplasms Morphology Mentions in Spanish Clinical Cases through Word Embeddings. Proceedings of the Iberian Languages Evaluation Forum. 2020. p. 324–34.

106. de Vargas Romero G, Segura-Bedmar I. Exploring Deep Learning for Named Entity Recognition of Tumor Morphology Mentions. Proceedings of the Iberian Languages Evaluation Forum. 2020. p. 396–411.

107. Vunikili R. Clinical NER using Spanish BERT Embeddings. Proceedings of the Iberian Languages Evaluation Forum. 2020. p. 505–11

108. García-Pablos A, Perez N, Cuadros M, Zotova E. Vicomtech at eHealth-KD Challenge 2020: Deep End-to-End Model for Entity and Relation Extraction in Medical Text. Proceedings of the Iberian Languages Evaluation Forum. 2020. p. 102–11.

109. Pavanelli L, Schneider ETR, Gumiel YB, Ferreira TC, de Oliveira LFA, de Souza JVA, et al. PU-CRJ-PUCPR-UFMG at eHealth-KD Challenge 2021: A Multilingual BERT-based System for Joint Entity Recognition and Relation Extraction. Proceedings of the Iberian Languages Evaluation Forum. 2021. p. 683–91.

110. Balouchzahi F, Sidorov G, Shashirekha HL. ADOP FERT-Automatic Detection of Occupations and Profession in Medical Texts using Flair and BERT. Proceedings of the Iberian Languages Evaluation Forum. 2021. p. 747–57.

111. Harkawat J, Vaidhya T. Spanish Pre-Trained Language Models for HealthCare Industry. Proceedings of the Iberian Languages Evaluation Forum. 2021. p. 796–802.

112. Schwarz M, Chapman K, Häussler B. Multilingual Medical Entity Recognition and Cross-lingual Zero-Shot Linking with Facebook AI Similarity Search. Proceedings of the Iberian Languages Evaluation Forum. 2022.

113. Avram AM, Mitrofan M, P is V. Species Entity Recognition Using a Neural Inhibitory Mechanism. Proceedings of the Iberian Languages Evaluation Forum. 2022.

114. Tamayo A, Burgos D, Gelbukh A. ParTNER: Paragraph Tuning for Named Entity Recognition on Clinical Cases in Spanish using mBERT+ Rules. Proceedings of the Iberian Languages Evaluation Forum. 2022.

115. Piad-Morffis A, Gutiérrez Y, Canizares-Diaz H, Estevez-Velarde S, Muñoz R, Montoyo A, et al. Overview of the ehealth knowledge discovery challenge at IberLEF 2020. Proceedings of the Iberian Languages Evaluation Forum. 2020. p. 71–84.

116. Miranda-Escalada A, Gascó L, Lima-Lopez S, Farré-Maduell E, Estrada D, Nentidis A, et al. Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. CLEF (Working Notes). 2022; p. 179–203.

117. Chizhikova M, Collado-Montañez J, López-Úbeda P, Díaz-Galiano MC, Ureña-López LA, Martín-Valdivia MT. SINAI at CLEF 2022: Leveraging biomedical transformers to detect and normalize disease mentions. CLEF (Working Notes). 2022. p. 265–73.

118. Neves A. Unicage at DISTE.MIST-Named Entity Recognition system using only Bash and Unicage tools. CLEF (Working Notes); 2022. p. 325–34

119. Borchert F, Schapranow MP. HPI-DHC@ BioASQ DisTEMIST: Spanish Biomedical Entity Linking with Pre-trained Transformers and Cross-lingual Candidate Retrieval. CLEF (Working Notes). 2022. p. 244–58.

120. Cardon R, Grabar N, Grouin C, Hamon T. Présentation de la campagne d'évaluation DEFT 2020: similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques. DEFT; 2020. Available at: https://

aclanthology.org/2020.jeptalnrecital-deft.1/.

121. Copara Zea JL, Knafou JDM, Naderi N, Moro C, Ruch P, Teodoro D. Contextualized French Language Models for Biomedical Named Entity Recognition. Actes de la 6e conférence conjointe Journées d'Études sur la Parole, Traitement Automatique des Langues Naturelles, Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues, Atelier DÉfi Fouille de Textes (DEFT). 2020. p. 36–48. Available at: https://aclanthology.org/2020.jeptalnrecital-deft.4/.

122. Naderi N, Knafou J, Copara J, Ruch P, Teodoro D. Ensemble of deep masked language models for effective named entity recognition in health and life science corpora. Front Res Metr Anal 2021;6. doi:10.3389/frma.2021.689803.

123. Grouin C, Grabar N, Illouz G. Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne DEFT 2021. Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles, Atelier DÉfi Fouille de Textes (DEFT). 2021. p. 1–13. Available at: https://aclanthology.org/2021.jeptalnrecital-deft.1.

124. Mannion A, Chevalier T, Schwab D, Goeuriot L. Identification de profil clinique du patient : Une approche de classification de séquences utilisant des modèles de langage français contextualisés. Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. 2021. p. 54–62. Available at: https://aclanthology.org/2021.jeptalnrecital-deft.6/.

125. Dou Z, Yamamoto T. Overview of NTCIR-16. Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies. 2022. p. 3–7.

126. Tamayo A, Gelbukh A, Burgos DA. NLP-CIC-WFU at SocialDisNER: Disease mention extraction in Spanish tweets using transfer learning and search by propagation. Proceedings of The Seventh Workshop on Social Media Mining for Health Applications. 2022. p. 19–22. Available at: https://aclanthology.org/2022.smm4h-1.6.

127. Cetina K, García-Santa N. FRE at SocialDisNER: Joint learning of language models for named entity recognition. Proceedings of The Seventh Workshop on Social Media Mining for Health Applications. 2022. p. 68–70. Available at: https://aclanthology.org/2022.smm4h-1.20.

128. Verma H, Bagherzadeh P, Bergler S. CLaCLab at SocialDisNER: Using medical gazetteers for named-entity recognition of disease mentions in Spanish tweets. Proceedings of The Seventh Workshop on Social Media Mining for Health Applications. 2022. p. 68–70. Available at: https://aclanthology.org/2022.smm4h-1.16.

129. Montañés-Salas R, López-Bosque I, García-Garcés L, del-Hoyo-Alonso R. ITAINNOVA at SocialDisNER: A Transformers cocktail for disease identification in social media in Spanish. Proceedings of The Seventh Workshop on Social Media Mining for Health Applications. 2022. p. 71–4. Available at: https://aclanthology.org/2022.smm4h-1.21.

130. Fu J, Li S, Yuan HM, Li Z, Gan Z, Chen Y, et al. CASIA@SMM4H'22: A uniform health information mining system for multilingual

social media texts. Proceedings of The Seventh Workshop on Social Media Mining for Health Applications. p. 143–7. Available at: https://aclanthology.org/2022.smm4h-1.39.

131. Conneau A, Khandelwal K, Goyal N, Chaudhary V, Wenzek G, Guzmán F, et al. Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020. p. 8440–51. doi:10.18653/v1/2020.acl-main.747.

132. Canete J, Chaperon G, Fuentes R, Ho JH, Kang H, Pérez J. Spanish pre-trained BERT model and evaluation data. Practical ML for Developing Countries Workshop. 2020:1–10.

133. López-García G, Jerez JM, Ribelles N, Alba E, Veredas FJ. Detection of Tumor Morphology Mentions in Clinical Reports in Spanish Using Transformers. Proceedings of the International Work-Conference on Artificial Neural Networks. Springer; 2021. p. 24–35.

134. Martin L, Muller B, Suárez PJO, Dupont Y, Romary L, de la Clergerie ÉV, et al. CamemBERT: a Tasty French Language Model. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2020. p. 7203–19. doi:10.18653/v1/2020.acl-main.645.

135. Le H, Vial L, Frej J, Segonne V, Coavoux M, Lecouteux B, et al. FlauBERT: Unsupervised Language Model Pre-training for French. ArXiv; 2020; Available at: http://arxiv.org/abs/1912.05372.

136. Otegi A, Agirre A, Campos JA, Soroa A, Agirre E. Conversational question answering in low resource scenarios: A dataset and case study for Basque. Proceedings of the 12th Language Resources and Evaluation Conference. 2020. p. 436–42. Available at: https://aclanthology.org/2020.lrec-1.55/.

137. Gutiérrez Fandiño A, Armengol Estapé J, Pàmies M, Llop Palao J, Silveira Ocampo J, Pio Carrino C, et al. MarIA: Spanish language models. Procesamiento del Lenguaje Natural 2022;68:39–60. doi:10.26342/2022-68-3.

138. López-García G, Jerez JM, Ribelles N, Alba E, Veredas FJ. Transformers for Clinical Coding in Spanish. IEEE Access 2021;9:72387–97. doi:10.1109/ACCESS.2021.3080085.

139. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. Proceedings of the Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2020. p. 1877–901. Available at: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

140. Biswas S. ChatGPT and the Future of Medical Writing. Radiology 2023 Apr;307(2):e223312. doi:10.1148/radiol.223312.

141. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. Appl Sci 2021 Jan;11(14):6421. doi:10.3390/app11146421.

142. Andrade VDT, Ruas P, Couto FM. Named Entity Recognition and Linking: a Portuguese and Spanish Oncological Parallel Corpus. bioRxiv; 2021. p. 2021.09.16.460605.

doi:10.1101/2021.09.16.460605.

143. Ma H, Yang F, Ren J, Li N, Dai M, Wang X, et al. ECCParaCorp: a cross-lingual parallel corpus towards cancer education, dissemination and application. BMC Med Inform Decis Mak 2020 Jul 9;20(3):122. doi:10.1186/s12911-020-1116-1.

144. Mititelu VB, Mitrofan M. The Romanian medical treebank-SiMoNERo. Proceedings of the 15th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing–ConsILR. 2020. p. 7–16.

145. Blinov P, Reshetnikova A, Nesterov A, Zubkova G, Kokh V. RuMedBench: A Russian Medical Language Understanding Benchmark. Proceedings of the International Conference on Artificial Intelligence in Medicine. 2022. p. 383–92.

146. Campillos-Llanos L, Valverde-Mateos A, Capllonch-Carrión A. A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. BMC Med Inform Decis Mak 2021;21(69). doi:10.1186/s12911-021-01395-z.

Correspondence to:
Anastassia Shaitarova
Department of Computational Linguistics
University of Zurich
Andreasstrasse 15
8050 Zürich
Switzerland
Tel: +41 78 614 7886
E-mail: anastassia.shaitarova@uzh.ch