SINAI at LivingNER Shared Task 2022: Species Mention Recognition and Normalization Using Transfer Learning and String Matching Techniques

Mariia Chizhikova[1,*,†], Jaime Collado-Montañez[1,†], Pilar López-Úbeda[2], Manuel Carlos Díaz-Galiano[1], Luis Alfonso Ureña-López[1] and María Teresa Martín-Valdivia[1]

[1]University of Jaén, Campus Las Lagunillas s/n, 23071, Jaén, Spain
[2]R+D+I department, HT medica, Carmelo Torres nº2, 23007, Jaén, Spain

Abstract

This paper describes the systems presented by the SINAI team for two LivingNER tracks on the IberLEF 2022 at SEPLN 2022. These shared tasks focus on the recognition and normalization of human and non-human species mentions in electronic health records written in Spanish. In order to solve the proposed problem, our team presents a multi-label named entity recognition system based on fine-tuned RoBERTa architecture model, which scored 0.9457 micro-average F1-score when officially tested on the gold-standard corpus. Based on this system's prediction, our entity normalization system combines exact matching, TF-IDF fuzzy matching, and Levenshtein distance-based matching with a 0.8629 micro-averaged F1-score.

Keywords

Clinical entity recognition, Clinical entity linking, Biomedical Natural Language Processing, RoBERTa language model

1. Introduction

The introduction of Electronic Health Records (EHRs) as the main way of registering patients' medical history created new opportunities to impulse research and development of biomedical technologies to bring knowledge and evidence to practice[1]. Structured clinical information translated to machine-readable format is considered an essential part of Clinical Decision Support Systems (CDSS), which aim to increase efficiency and quality of healthcare [2]. However, one of the most important parts of EHRs, free-text clinical narrative is highly expressive but difficult to handle automatically [3]. This fact makes the extraction of clinically meaningful features from raw text data an essential step to leverage the knowledge stored in EHRs.

One of the ways of making relevant concepts in free-text clinical narrative machine-readable

IberLEF 2022, September 2022, A Coruña, Spain.
*Corresponding author.
†These authors contributed equally.
mc000051@red.ujaen.es (M. Chizhikova); jcollado@ujaen.es (J. Collado-Montañez); p.lopez@htmedica.com (P. López-Úbeda); mcdiaz@ujaen.es (M. C. Díaz-Galiano); laurena@ujaen.es (L. A. Ureña-López); maite@ujaen.es (M. T. Martín-Valdivia)

is clinical coding. This term refers to the process of mapping concepts from patients' EHR to codes from a relevant controlled vocabulary[4]. At the present, the interpretation of EHRs and structuring this information is performed manually and can even be an additional responsibility of physicians. This process is time-consuming and entails errors attributed to the human factor. For these reasons the development of systems that could perform accurate automated clinical coding has great potential to save resources spent on manual annotation, improve oversight of patient care and accelerate evidence-based biomedical research [3].

Biomedical Natural Language processing (Bio-NLP) aims to give a response to the need for automated extraction of evidence from a retrospective analysis of EHR data by Named Entity Recognition (NER) and Named Entity Normalization (NEN), which constitute two main phases of the clinical coding process. Biomedical named entities conform a large and heterogeneous group of concepts that can include gene, disease, chemicals, and drug names [5]. However, not all of this entities types received equal attention from the Bio-NLP community: while detection and normalization of the cited were targeted by several shared tasks [6, 7, 4], organisms and species mentions have been scarcely featured in Bio-NLP domain. For instance, LINNAEUS tool [8] combines rule-based and pattern matching approaches for species entity recognition and normalization in English clinical texts. One of the disadvantages of this system is that the described approach makes it difficult to adapt for other languages and to keep it updated apace with taxonomy extensions.

By providing an exhaustively annotated corpus of clinical case reports, LivingNER shared task [9] at Iberian Languages Evaluation Forum (IberLEF) 2022 brings the community effort to designing clinical coding systems capable of correctly identifying both human and non-human species mentions in Spanish clinical narratives and normalizing these mentions by assigning each one its corresponding code from National Center for Biotechnology Information Taxonomy (NCBITax) [10]. As of June 2022 NCBITax includes 738,683 species with formal names structured according to the evolution history [10].

In this paper, we describe the systems presented by the SINAI team to tackle LivingNER-Species NER and LivingNER-Species Norm tracks.

2. Data

LivingNER Corpus is a collection of 2,000 clinical case reports annotated with quite diversified species mentions such as humans, animals, plants, and microorganisms [11, 9]. The fact that the texts come from 20 different medical disciplines, such as cardiology, neurology, oncology, otolaryngology or dentistry, contributes to the variety of the dataset. The corpus annotations considered not only scientific names (e.g. *Homo sapiens*), but also common names of the cited concept types (e.g. *Person*).

The entire collection was randomly split by the task's organizers into training, validation, and test sets that comprised 1000, 500, and 500 texts respectively. Additionally to the test set, a larger background set was released to prevent participating teams from manually correcting their predictions and to generate a Silver Standard. Training texts were annotated with 3,006 unique entity mentions that were mapped to 887 unique NCBITax identifiers, while the validation dataset contained 1,704 unique mentions mapped to 559 different NCBITax annotations. Figure 1

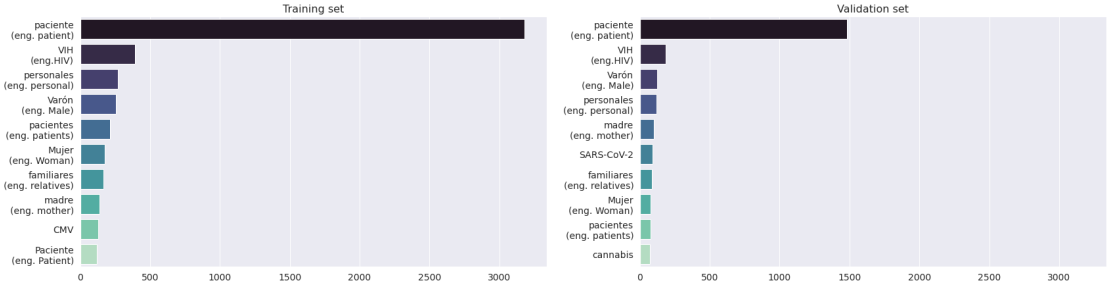

Figure 1: 10 most common entity spans among the dataset. *Translation made only to ease the reading*

Figure 2: Entity label frequency comparison.

shows the 10 most common entity spans in training and validation sets. Considering that the main objective of the LivingNER-Species NER is retrieving both human and non-human species mentions from EHRs, corpus annotations also specify a label for each entity which can be 'SPECIES' or 'HUMAN', where entities with this last label are always linked to the same code in the taxonomy (9606, *Homo sapiens*). As shown in Figure 2, the dataset is quite balanced in this regard.

As for the NCBITax identifiers, it is important to note that not all entities were assigned a single NCBITax code since there can be found composite mentions, used for entities that required several codes to be described ('2|2759|10239' for 'patógenos respiratorios' - eng. *respiratory pathogens*), and terminology codes with '|H' modifiers which refer to codes that are more general than the annotated mention ('2|H' for 'baciloscopia' - eng. *bacilloscopy*). As can be seen in Table 1, composite mentions and general codes are among the 10 most frequently appearing codes in both training and validation subsets of the LivingNER corpus.

Another highly relevant characteristic of the LivingNER corpus is the length of the texts. Tokenization of each one with the RoBERTa byte-level Byte-Pair-Encoding tokenizer [12] revealed that 595 texts from the training set and 257 from the validation set exceeded the maximum length of input for the RoBERTa model chosen as the core part of our system, which

Training set			Validation set		
Code	**Description**	**Count**	**Code**	**Description**	**Count**
9606	homo sapiens	7,007	9606	homo sapiens	3,289
12721	hiv	460	12721	hiv	193
2	bacteria	370	10239	viruses	163
10239	viruses	318	2	bacteria	151
2|2759|10239	bacteria | eukaryota | viruses	280	10407	hepatitis b virus	133
4751	fungi	242	4751	fungi	110
2|H	bacteria (general)	234	10358	cytomegalovirus	109
10407	hepatitis b virus	232	2|H	bacteria (general)	100
160	treponema pallidum	226	2697049	sars-cov2	98
10358	cytomegalovirus	214	2|2759|10239	bacteria | eukaryota | viruses	90

Table 1
10 most frequent NCBITax codes in training and validation datasets.

is set to 512 tokens. This fact brought us to adopt a sentence-level NER approach, thus text preprocessing consisted in splitting the texts into sentences using the *SentenceRecognizer* from the SpaCy processing pipeline [13]. SpaCy's *SentenceRecognizer* relies on the `es_core_news_sm` pre-trained language model which was employed to identify tokens that constitute the start of each sentence. Table 2 summarizes basic statistics of the dataset.

	Training set			Validation set		
	Entities	**Tokens**	**Sentences**	**Entities**	**Tokens**	**Sentences**
max	90	2,744	138	110	2,534	97
min	1	81	3	1	95	3
avg	16.10	728.53	28.34	14.21	671.89	25.69
SD	14.58	457.84	18.34	13.26	434.98	17.22

Table 2
Training and validation corpus statistics.

3. System Description

In this section, we describe the systems developed for the subtasks we participated. Figure 3 provides an overview of the proposed system's architecture.

3.1. LivingNER-Species NER

LivingNER-Species NER track consisted in providing character offsets of all human and non-human species mentions given a collection of plain text clinical report documents. As we mentioned earlier, many texts from both training and validation sets exceeded the maximum length of input of transformer models we were planning to work with, we opted for developing

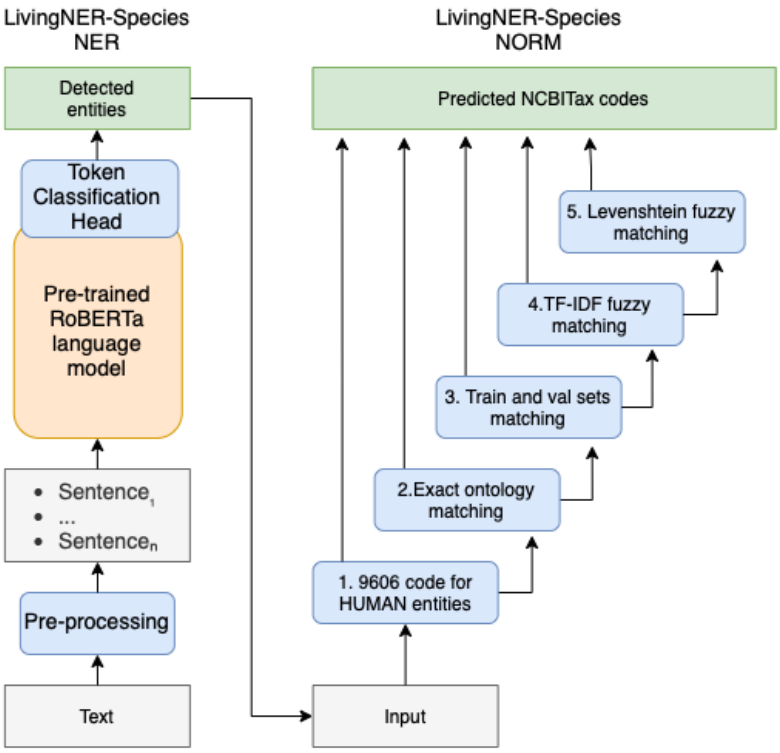

Figure 3: Overview of system architecture.

a sentence-level NER system based on fine-tuning of pre-trained language models. With the objective of measuring the impact of domain-specific training that has frequently been pointed at since the release of the first Bio-BERT [14], we carried out three experiments testing RoBERTa models pre-trained on general, biomedical, and biomedical-clinical corpora.

For the general-domain model we selected RoBERTa-base-BNE [15]. This model was pre-trained on more than 200M documents assembled from crawls of all .es domains performed by the National Library of Spain (Biblioteca Nacional de España or BNE) between 2009 and 2019 that were subsequently subjected to a complex filtering and deduplication process [15]. In its original paper, the authors provide benchmarks for RoBERTa-base performance on several downstream tasks such as NER, text classification, question answering, and others, making use of general domain corpora, such as the Spanish portion of MLDoc [16] or CAPITEL-NER [17]. Unlike this, using this model on clinical text constitute a cross-domain experiment.

Two other models chosen for experimentation have the same RoBERTa architecture and differ in terms of the pre-training data. The medical model was trained on data gathered from a variety of sources, such as Medical Crawler [18], SciELO-Spain-Crawler [19], and others. Language variety used in scientific medical texts, despite being semantically closer to the jargon used in EHRs, still differs from it in terms of syntax and vocabulary, a fact that made relevant introduction of clinical narratives into the pre-training corpus and led to the creation of clinical

model [20].

All three (medical, clinical, and general) models were fine-tuned for the token classification task by adding a linear classifier layer preceded by a 0.1 dropout layer on top of the original architecture.

3.2. LivingNER-Species NORM

This task aims to assign each mention found in the first task a code from the NCBI Taxonomy. In order to tackle this, we developed a sequential system that combines exact and fuzzy string matching between entities and taxonomy terms.

All mentions in the train set labeled as HUMAN are linked to the *Homo sapiens* term in the taxonomy for which code is 9606, therefore, our system first assigns this code to every entity containing such label as detected in the previous task (step 1 in Figure 3). Then, regarding the mentions labeled as SPECIES, we first looked for exact string matching between the entities and both the original taxonomy descriptions and their Spanish translated version (step 2). In the third step, the system performs another exact string matching, but this time against the entities present in the sets provided by the organizers (i.e. training and validation). The remaining entities to be linked are then encoded into numerical vectors with TF-IDF in order to get the most similar terms by calculating the cosine similarity with every TF-IDF encoded taxonomy description (step 4). Finally in the fifth step, for the entities that do not have any similar term in the ontology - the normalized cosine similarity is less than 0.75 -, we repeat this fuzzy string matching process, but now calculate the Levenshtein distance between entities and taxonomy descriptions. The entities for which this last normalized distance metric is not greater than 0.9 will be coded as _NOCODE_, which is the code for those mentions that are out of the NCBI taxonomy scope.

4. Experimental Setup

All the transformer models were fine-tuned on a single NVIDIA Ampere A100 GPU by making use of the Hugging Face's transformers Python library [21].

To yield better performance out of our systems, we carried out a hyperparameter optimization powered by the Optuna Framework [22]. It incorporates an efficient implementation of both searching and trial pruning strategies. During the optimization, Optuna infers concurrence relations between the searched parameters to switch from independent sampling to concurrence sampling after a few trials. In addition, a pruning algorithm monitors intermediate training results and stops unpromising trials at an early stage.

The hyperparameter space for the optimization was defined as follows:

- Learning rate: float value between $3e-5$ and $5e-5$
- Number of training epochs: integer value between 3 and 10
- Training batch size: 8, 16 and 32
- Weight decay: float value between $1e-12$ and $1e-1$
- AdamW optimizer epsilon: float value between $1e-10$ and $1e-6$
- Warmup steps: integer value between 0 and 1000

	Medical model	Clinical model	General model
Learning rate	3.6e-5	4.8e-5	3.1e-5
Training epochs	10	4	9
Batch size	16	8	32
Weight decay	1.4e-11	3.2e-6	2.4e-11
AdamW epsilon	3.9e-10	1.4e-8	4.4e-100
Warmup steps	622	440	276

Table 3
Hyperparameters selected for each model.

Finally, Table 3 summarizes hyperparameters selected for each model after optimization trials.

5. Results

Our team submitted a total of three runs corresponding to each of the above-described systems. In this section, we present the results obtained in LivingNER-Species NER and LivingNER-Species NORM sub-tracks during the official evaluation. Metrics selected by the organization committee to measure the performance for both are those commonly used in token classification tasks, namely micro-averaged precision (P), recall (R), and F1-score (F1).

Table 4 summarises the results obtained by the three NER systems presented on the LivingNER-Species NER subtrack. Metrics prove that the RoBERTa model pre-trained on general domain corpora has a remarkable cross-domain potential as its performance, despite being the lowest of all our systems is still promising and close to the mean values among all submitted runs. The general model is also able to most accurately identify the relevant human mentions, as it has shown the highest recall (0.9803) in the 'Only HUMAN' category.

The highest F1-scores were shown by the medical model (0.9457 overall, 0.9198 for 'Only SPECIES' and 0.9805 for 'Only HUMAN'). Nevertheless, the performance of this pair is very similar to each other.

	LivingNER NER			Only SPECIES			Only HUMAN		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Medical model	0.9571	**0.9346**	**0.9457**	0.939	**0.9014**	**0.9198**	**0.9815**	0.9794	**0.9805**
Clinical model	**0.9574**	0.905	0.9305	**0.9421**	0.8572	0.8976	0.9776	0.9695	0.9735
General model	0.8995	0.9295	0.9142	0.8643	0.8918	0.8779	0.9502	**0.9803**	0.965
MEAN	0.8763	0.8077	0.8239	0.8112	0.7579	0.7781	0.9312	0.875	0.8849
SD	0.1542	0.2465	0.2371	0.249	0.2565	0.251	0.1156	0.2388	0.223

Table 4
Official results obtained in the LivingNER-Species NER track.

The results achieved in the second subtask (see Table 5) are highly related to the ones in the LivingNER-Species NER subtask since the taxonomy codes are linked to the entities detected in

this first track. This explains why the 'Only HUMAN' results in both tasks are identical (all entities with this label are linked to the *Homo sapiens* code, therefore, the relative accuracy of our system for this label in the second subtask is 100%). Regarding the 'Only SPECIES' metrics, our best model achieved micro-averaged scores of 0.7905 for precision, 0.7589 for recall, and 0.7744 for F1. Finally, the general scores achieved by our best system are 0.8733, 0.8527, and 0.8629 for precision, recall, and F1 respectively which are all higher than the mean metrics achieved by all the participants.

	LivingNER Norm			**Only SPECIES**			**Only HUMAN**		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
medical-model	0.8733	**0.8527**	**0.8629**	0.7905	**0.7589**	**0.7744**	**0.9815**	0.9794	**0.9805**
clinical-model	**0.8738**	0.826	0.8492	**0.7908**	0.7196	0.7535	0.9776	0.9695	0.9735
general-model	0.823	0.8504	0.8365	0.731	0.7542	0.7424	0.9502	**0.9803**	0.965
MEAN	0.8488	0.8068	0.8267	0.7598	0.6917	0.7228	0.9588	0.9621	0.9604
SD	0.1577	0.1476	0.1508	0.2893	0.2582	0.2704	0.0193	0.014	0.0151

Table 5
Official results obtained in the LivingNER-Species Norm track.

5.1. Conclusions an future work

This paper covers the participation of the SINAI research group in the LivingNER shared task at the IberLEF 2022 workshop. In particular, we took part in two subtasks: one consisting of species mention detection (LivingNER-Species NER) and the other focusing on mapping the previously detected entities to NCBITax codes (LivingNER-Species NORM).

The approach proposed to tackle the species mention recognition task (first subtask) has in its basis fine-tuning of transformer models pre-trained on different corpora. To measure the impact of domain-specific pre-training of large language models we tested three systems with different cores: one relied on a general-domain model, the second on a medical model, and the third, on a clinical one. All pre-trained models had the same RoBERTa transformer architecture and differed in terms of pre-training corpora. All three proposed systems achieved encouraging results, being the medical model the best performing one with 0.9457 overall micro-average F1-score.

Despite our initial hypothesis of being the clinical model more suitable for LivingNER Corpus, it showed very similar but slightly lower performance (0.9305 overall F1-score) compared to the medical model. The reason for this could be the fact that, in absence of large-scale clinical corpora, RoBERTa clinical was pre-trained on a concatenation of medical scientific publications with clinical case reports, and clinical cases accounted for only 14,96% of the total number of documents in the resulting collection [20]. Another interesting finding of our experimentation is the remarkable cross-domain potential the general model has shown since when fine-tuned on clinical texts it achieved 0.9142 overall micro-average F1-score which is higher than the mean value among all submitted predictions.

Regarding our approach for the second subtask, we developed a sequential system including

exact and fuzzy string matching to find the closest NCBITax codes for the entities found in the previous subtask. The results obtained suggest the strength of this approach as our system achieved 0.8629 F1-score, which is almost four percent higher than all participants' mean score.

As future work, an exhaustive error analysis would help to identify our systems' weaknesses and improve their performance. Furthermore, we plan to test the emerging weak-supervision techniques that leverage so-called 'weak' labels obtained by applying different matching techniques or by making predictions with another previously fine-tuned model to enhance the resulting system performance.

Acknowledgments

This work has been partially supported by Big Hug project (P20_00956, PAIDI 2020) and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, LIVING-LANG project (RTI2018-094653-B-C21) funded by MCIN/AEI/10.13039/501100011033

References

[1] J. M. McGinnis, D. Aisner, L. Olsen, et al., The learning healthcare system: workshop summary, National Academies Press, 2007.

[2] R. Williams, E. Kontopantelis, I. Buchan, N. Peek, Clinical code set engineering for reusing ehr data for research: a review, Journal of Biomedical Informatics 70 (2017) 1–13.

[3] Towards automated clinical coding, International Journal of Medical Informatics 120 (2018) 50–61. URL: https://www.sciencedirect.com/science/article/pii/S1386505618304039. doi:https://doi.org/10.1016/j.ijmedinf.2018.09.021.

[4] A. Miranda-Escalada, A. Gonzalez-Agirre, J. Armengol-Estapé, M. Krallinger, Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2020.

[5] H. Cho, H. Lee, Biomedical named entity recognition using deep neural networks with contextual information, BMC bioinformatics 20 (2019) 1–11.

[6] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results., IberLEF@ SEPLN (2020) 303–323.

[7] C. Arighi, L. Hirschman, T. Lemberger, S. Bayer, R. Liechti, D. Comeau, C. Wu, Bio-id track overview, in: Proc. BioCreative Workshop, volume 482, 2017, p. 376.

[8] M. Gerner, G. Nenadic, C. M. Bergman, Linnaeus: a species name identification system for biomedical literature, BMC bioinformatics 11 (2010) 1–17.

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

[10] S. Federhen, The ncbi taxonomy database, Nucleic acids research 40 (2012) D136–D143.

[11] A. Miranda-Escalada, E. Farré-Maduell, G. González Gacio, M. Krallinger, LivingNER corpus: Named entity recognition, normalization & classification of species, pathogens and food, 2022. URL: https://doi.org/10.5281/zenodo.6642852. doi:10.5281/zenodo.6642852, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv:1907.11692 [cs] (2019). URL: http://arxiv.org/abs/1907.11692, arXiv: 1907.11692.

[13] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python (2020). doi:10.5281/zenodo.1212303.

[14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics (2019). URL: https://doi.org/10.1093%2Fbioinformatics%2Fbtz682. doi:10.1093/bioinformatics/btz682.

[15] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022) 39–60. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405.

[16] H. Schwenk, X. Li, A corpus for multilingual document classification in eight languages, arXiv preprint arXiv:1805.09821 (2018).

[17] J. Porta-Zamorano, L. Espinosa-Anke, Overview of capitel shared tasks at iberlef 2020: Named entity recognition and universal dependencies parsing, arXiv preprint arXiv:2011.05932 (2020).

[18] C. P. Carrino, J. Armengol-Estapé, O. de Gibert Bonet, A. Gutiérrez-Fandiño, A. Gonzalez-Agirre, M. Krallinger, M. Villegas, Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models, 2021. arXiv:2109.07765.

[19] A. Intxaurrondo, Scielo-spain-crawler, 2019. URL: https://doi.org/10.5281/zenodo.2541681. doi:10.5281/zenodo.2541681, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL), the ICTUSnet INETRRREG SUdooe project and "La MARATO de TV3".

[20] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, M. Villegas, Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, 2021. arXiv:2109.03570.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Huggingface's transformers: State-of-the-art natural language processing, 2019. URL: https://arxiv.org/abs/1910.03771. doi:10.48550/ARXIV.1910.03771.

[22] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019.