SINAI at CLEF 2022: Leveraging biomedical transformers to detect and normalize disease mentions

Mariia Chizhikova[1], Jaime Collado-Montañez[1], Pilar López-Úbeda[2], Manuel C. Díaz-Galiano[1], L. Alfonso Ureña-López[1] and M. 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 n°2, 23007, Jaén, Spain

Abstract

This paper presents the system developed by SINAI team for Disease Text Mining and Indexing Shared Task at CLEF 2022 BioASQ workshop. This task brings the community effort to development of automatic disease mention detection and semantic indexing systems for electronic health records written in Spanish. Our proposal is based on a deep learning RoBERTa architecture model fine-tuned for the named entity recognition task, which achieved 0.75 micro-average F1-score during the official evaluation. For the entity linking task, we introduce a system based on a combination of term matching and embedding similarity calculation which best micro-average F1-score is 0.41.

Keywords

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

1. Introduction

Clinical coding stands for transforming medical information from patient's Electronic Health Records (EHR) into alphanumeric codes using a classification standard [1]. Nowadays the interpretation of EHRs and the assignation of standardised codes lies on human coders or even on physicians themselves. However, the massive amount medical of data that increases with each patient's visit has become an excessive burden for human annotators [2]. This led to a rise in demand for development and improvement of the automatic curation systems capable to handle massive amounts of EHRs.

Natural Language Processing (NLP) aims to address the need of managing unstructured data in order to extract relevant information that makes Information Retrieval (IR) more efficient or can serve as input for such application as Clinical Decision Support Systems (CDSS), for example [3].

Search queries that mention disorders (this refers to diseases, abnormalities, dysfunctions, syndromes, injuries, etc.) constitute the first most frequent non-bibliographic query type among

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
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)

PubMed users [4]. The relevance of this category in clinical texts is also very high, which emphasises the need of creating accurate Named Entity Recognition (NER) and Named Entity Normalization (NEN) systems to improve information retrieval. This kind of systems would automatically detect disorder mentions in both scientific and clinical texts subsequently mapping them to codes in a relevant controlled vocabulary.

With Bidirectional Encoder Representation from Transformers (BERT)[5], large pre-trained neural language models of transformer architecture became an essential building block for many NLP tasks, such as text classification, NER, text summarization, etc. Nevertheless, transfer-learning capacities of these models depend, among many diverse factors, on the language variety differences between the pre-training corpora and task's data. Considering that the vocabulary and syntax of medical jargon differs from general-domain language, continual pre-training of a general-domain model was proposed as a way of improving its performance in biomedical NLP [6]. Despite being beneficial, continual pre-training does not extend the vocabulary of the original model, which maintains unrepresentative of domain-specific texts. This fact led to the proposal of domain-specific pre-training from scratch that was proved to be more efficient than continual pre-training [7].

Disease Text Mining and Indexing Shared Task (DISTEMIST) at CLEF 2022 BioASQ workshop brings the community effort to design systems capable of making disorder mention in clinical text accessible for search systems by identifying them and mapping each one to a code from the Systematized Nomenclature of Medicine – Clinical Terms (Snomed-CT)[1]. Snomed-CT is a integral multilingual clinical terminology that contains almost 800,000 descriptions, including synonyms that can be used to refer to a concept, that are linked with semantic relationships. Moreover, Snomed-CT is called the most comprehensive clinical healthcare terminology in the world[2].

In this paper we describe the approach followed by the SINAI team to tackle both NER and NEN DISTEMIST subtasks. The success of biomedical domain-specific pre-training of large transformer language models [6] brought us to test two models of the same architecture that were pre-trained on different corpora [8] to evaluate its performance on NER task. For the DISTEMIST-linking sub-task we propose a multi-step approach that combines embedding similarity calculation and term matching.

2. Data

Both DISTEMIST subtasks challenge researchers with real-world datasets, promoting the improvement of the state-of-the-art NLP systems for clinical coding[9]. The gold-standard corpus provided by workshop organization committee is a collection of 1,000 clinical cases in Spanish from different medical specialities such as cardiology, oncology, otorhinolaryngology, dentistry, pediatrics, primary care, allergology, radiology, psychiatry, ophthalmology, and urology annotated with disease mentions [10].

DISTEMIST organizers provided a collection of 750 clinical cases for the NER sub-track, 583 of which formed the training set for the NEN sub-track. This training set was annotated

[1] https://www.snomed.org

[2] https://www.snomed.org/snomed-ct/five-step-briefing

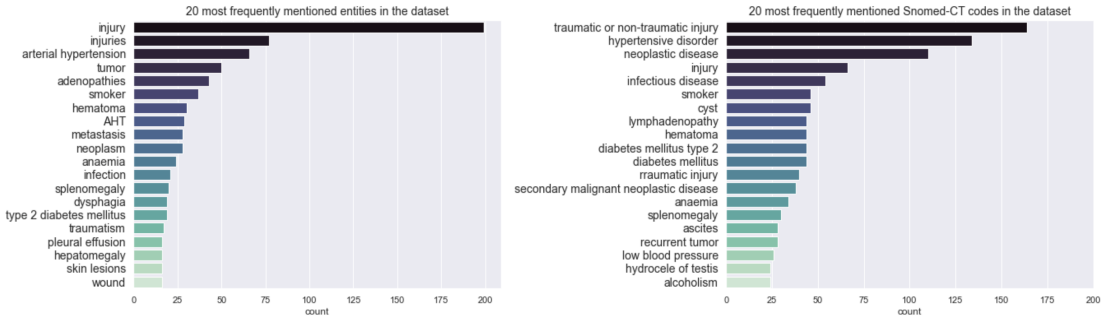

Figure 1: Descriptions of the 20 most frequent entities and Snomed-CT codes in the corpus (English translation of entities and code descriptions made only to ease the reading).

	Entities	Tokens	Sentences
max	57	1,486	132
min	1	98	6
avg	10.75	457	33.96
SD	6.21	218.28	16.39

Table 1
Corpus statistics.

with 5,348 unique entity mentions and 1,844 unique Snomed-CT codes, being 57 the maximum number of disease annotations per text. Figure 1 shows descriptions of the 20 most frequently mentioned entities and Snomed-CT codes in the DISTEMIST corpus.

One peculiarity of the provided annotations is the existence of nested disease mentions. With this we refer to complex expressions like "loss of kidney graft from chronic nephropathy" which is a disorder mention that contains another one, namely "chronic nephropathy". In the DISTEMIST Corpus such entities appear as separate annotations and the total count of this mentions is 411. During the pre-processing, we resolve nested disorder mentions in favour of the longest one.

The text length measures obtained by tokenizing each text with RoBERTa byte-level Byte-Pair-Encoding tokenizer [11] showed that the longest text contained 1,486 tokens and, most importantly, 248 texts from the training set exceeded the maximum length of input for the RoBERTa model selected as core of our system, which is set to 512. This fact brought us to perform sentence-level NER, thus text pre-processing consisted in splitting the texts in sentences using the *SentenceRecognizer* from the SpaCy processing pipeline [3]. SpaCy´s *SentenceRecognizer* relies on `es_core_news_sm` pre-trained language model [4] which was used to predict whether each token of every text starts a sentence or not. Some basic statistics of the dataset are summarized in Table 1.

It is important to mention that we randomly splitted the training set to be able to perform in-house validation during the process of system development. The resulting validation set

[3]https://spacy.io/api/sentencerecognizer
[4]https://spacy.io/models/es

contained 30% of training data.

As for the test set, it consisted of 250 additional cases for both sub-tracks, while the predictions were made on a concatenation of test and background sets with the total number 3,000 documents, which also we subjected to the same pre-processing as the training set.

3. System Description

In this section, we describe the systems developed for DISTEMIST-entities and DISTEMIST-linking sub-tasks.

3.1. Sub-task 1

The DISTEMIST-entities subtrack required automatically finding disease mentions in clinical cases. Taking into account the length of clinical texts in the dataset, as we anticipated in Section 2, we opted for a sentence-level NER approach based on fine-tuning of two pre-trained RoBERTa language models [11].

Our first system is based on a fine-tuned biomedical-clinical model[5], trained on a combination of biomedical and clinical texts that, hypothetically, suits better for the proposed task, due to the particular syntax and vocabulary that clinical texts present, comparing to medical scientific literature.

The second system developed for NER subtask leveraged medical domain-specific model [6] pre-trained on a the medical crawler collection [12] extended with data from other sources, such as SciELO-Spain-Crawler [13] and other.

The two models were fine-tuned for the token classification task by adding a linear classifier layer preceded with a 0.1 dropout layer on top of the original architecture.

3.2. Sub-task 2

This task aims to assign each mention found in the DISTEMIST-entities track a code from the list of relevant terms from Snomed-CT provided by the competition organizers [14]. To address this, we suggest a three step approach. First, we calculate the embeddings for all the entity spans detected in subtask 1 and for every term in the ontology with RoBERTa models that we fine-tuned for the previous subtask. We achieve this by mean pooling the last hidden state of the model's output. Then, we link each entity to the closest term in the ontology by calculating the cosine similarity between the resulting embeddings. The second step of our approach relies on looking for perfect matches between the mentions found and the ontology terms. In this phase, the system replaces the Snomed-CT codes assigned in the previous step if the mention's string is exactly the same as any ontology's term. 14429 entities, out of which 2618 are unique, have been found in this step. Lastly, we follow the same approach, but in this case we look for direct matches in the training set provided by the organizers where 6246 additional entities are found - 633 unique-. Therefore, exact string matching finds 20675 out of the 48699 entities detected in the previous subtask. Figure 2 illustrates architecture of the proposed system.

[5] https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es
[6] https://huggingface.co/PlanTL-GOB-ES/bsc-bio-es

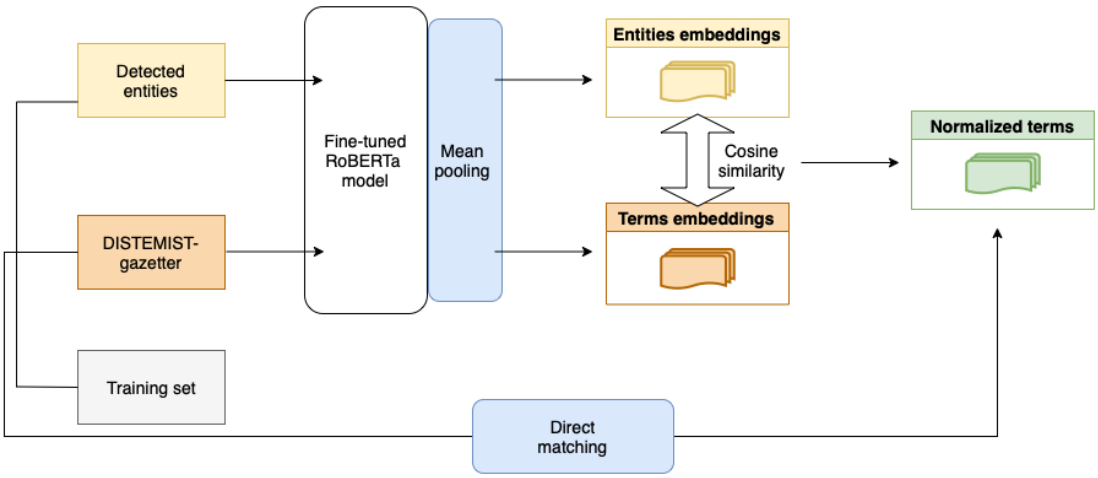

Figure 2: Named entity normalization system architecture.

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 [15].

In order to maximize the resulting performance of our systems we carried out a hyperparameter optimization powered by Optuna Framework [16]. The cited framework incorporates efficient implementation of both searching and pruning strategies. During the optimization, Optuna infers concurrence relations between the searched parameters to switch from independent sampling to concurrence sampling after few trials. In addition, a pruning algorithm monitors intermediate training results and terminates unpromising trials.

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

Finally, Table 2 summarizes hyperparameters selected for each model after optimization trials.

5. Results

Metrics selected by DISTEMIST organization team to evaluate system performance on both tracks are micro-average precision (MiP), micro-average recall (MiR) and micro-average F1-score (MiF1) - those are very commonly used for text and token classification tasks. Table 3

	RoBERTa biomedical	RoBERTa clinical
Learning rate	4e-5	5e-5
Training epochs	10	10
Batch size	8	16
Weight decay	1e-6	3e-6
AdamW epsilon	2.6e-9	1e-8
Warmup steps	73	440

Table 2
Hyperparameters selected for each model.

Subtask	System	MiP	MiR	MiF1
DISTEMIST-entities	RoBERTa biomedical	0.7520	0.7259	**0.7387**
	RoBERTa clinical	0.7519	0.7221	0.7367
	MEAN	0.6502	0.6079	0.6221
	SD	0.1633	0.1475	0.1585
DISTEMIST-linking	RoBERTa biomedical	0.4134	0.4069	0.4101
	RoBERTa clinical	0.4163	0.4081	**0.4122**
	MEAN	0.3965	0.335	0.3588
	SD	0.1381	0.1202	0.127

Table 3
Official results obtained by the SINAI team in DISTEMIST-entities and DISTEMIST-linking subtasks along with the mean (MEAN) and standard deviation (SD) of all participants' submissions.

summarises the results obtained by the SINAI team during the official evaluation carried out by the organizers.

The evaluation demonstrated that the systems pairs presented on both sub-tracks achieve very similar results despite the fact of being based on two different pre-trained models. Using the biomedical model on EHRs can be considered as cross-domain experiment and the fact that our biomedical system exhibits encouraging results (0.7387 MiF1) on the NER task highlights the existence of domain transfer potential between biomedical and clinical fields. The clinical model also performed well on the first sub-track scoring 0.7367 MiF1 on the test set.

Regarding the results obtained in the second subtask, our best system achieved a MiP of 0.4163, a MiR of 0.4081, and a MiF1 of 0.4122, all of them being higher than the average scores of all the participants. It is important to note that these results are highly dependent on the ones scored in the NER subtrack since the entities used to assign the normalized codes are the ones detected in that task.

5.1. Error analysis

With the objective of forming an opinion about pockets of low performance of our NER system, we conducted an error analysis based on model's performance on custom validation set that consisted in a random 30% split DISTEMIST Corpus, as stated in Section 2. The most frequently observed error type is related to nested entities. The system occasionally either detects a

Entity span	Detected
insuficiencia renal aguda *eng.: acute renal failure*	✓
insuficiencia renal aguda secundaria a administración de aciclovir *eng.: acute renal failure secondary to acyclovir administration*	✗

Table 4
Example 1 of incorrect labelling of a nested entity

Entity span	Detected
epilepsia rolándica izquierda secundaria a cisticercosis sistémica *eng.: left rolandic epilepsy secondary to systemic cysticercosis*	✗
cisticercosis sistémica *eng.: systemic cysticercosis*	✓

Table 5
Example 2 of incorrect labelling of a nested entity

complex mention and, consequently is not able to recognize one that forms part of it, as shown on Table 4, or detects only simple mention without returning the nested one, as illustrates Table 5.

6. Conclusions and future work

In this paper, we present systems developed by the SINAI team for DISTEMIST track at CLEF 2022 BioASQ workshop. This challenge consisted of two sub-tracks: one focused on detection of disease mention in EHRs and the other targeted mapping this mention to codes from the Snomed-CT ontology.

In order to address these two tasks our team followed an approach that takes as its basis fine-tuning of two transformer models pre-trained on biomedical and biomedical-clinical corpora. For the DISTEMIST-entities sub-track we fine-tune both models to perform sentence-level NER with a prior hyperparameter optimization step. For the DISTEMIST-linking sub-track we applied several techniques to find the closest term in the Snomed-CT ontology in order to assign a code to each entity.

The resulting performance of our NER systems revealed the remarkable cross-domain potential that the selected transformer-based model pre-trained on biomedical corpora has when fine-tuned on clinical texts. Our best performing NER system was also made publicly available on HuggingFace Hub [7]. As for the entity linking, calculating embedding distances provided encouraging results for entities that did not appear neither in the ontology nor in the training dataset.

For future work, we plan to address nested entities issue by testing novel approaches such as

[7] https://huggingface.co/chizhikchi/Spanish_disease_finder

Parallel Instance Query Networks (PIQN) [17]. Furthermore, a more in-depth error analysis needs to be carried out in order to infer the reasons of false positives and false negatives in the test test predictions and be able to suggest solutions for these problems. With the object of improving entity linking system performance, we plan on improving both matching and semantic similarity calculation. Having in mind that abbreviations and acronyms are commonly used in medical texts [4] we contemplate including disambiguation of abbreviated terms as a step prior to matching in our processing pipeline. Furthermore, we aim to execute some experiments using Levenshtein distance as the indicator of semantic similarity between text sequences.

7. Acknowledgements

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] 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: CLEF (Working Notes), 2020.

[2] F. Catling, G. P. Spithourakis, S. Riedel, 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.

[3] B. Al-Hablani, The use of automated snomed ct clinical coding in clinical decision support systems for preventive care, Perspectives in health information management 14 (2017).

[4] R. Islamaj Dogan, G. C. Murray, A. Névéol, Z. Lu, Understanding PubMed® user search behavior through log analysis, Database 2009 (2009). URL: https://doi.org/10.1093/database/bap018. doi:10.1093/database/bap018.

[5] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[6] 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.

[7] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare 3 (2022) 1–23. URL: https://doi.org/10.1145%2F3458754. doi:10.1145/3458754.

[8] 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 span-

ish: On the benefits of domain-specific pretraining in a mid-resource scenario, 2021. arXiv:2109.03570.

[9] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farr'e-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2022.

[10] A. Miranda-Escalada, E. Farré, L. Gasco, S. Lima, M. Krallinger, DisTEMIST corpus: detection and normalization of disease mentions in spanish clinical cases, 2022. URL: https://doi.org/10.5281/zenodo.6532684. doi:10.5281/zenodo.6532684, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

[11] 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.

[12] 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.

[13] 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".

[14] L. Gascó, M. Krallinger, Distemist gazetteer, 2022. URL: https://doi.org/10.5281/zenodo.6505583. doi:10.5281/zenodo.6505583, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

[15] 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.

[16] 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.

[17] Y. Shen, X. Wang, Z. Tan, G. Xu, P. Xie, F. Huang, W. Lu, Y. Zhuang, Parallel instance query network for named entity recognition, arXiv preprint arXiv:2203.10545 (2022).