HiTZ Zentroa at TestLink IberLEF 2023*

Adrian Cuadrón*†[1], Aimar Sagasti*†[1], Ander Barrena[1] and Aitziber Atutxa[1]

[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU

Abstract

This paper describes the systems presented by the HiTZ Center team for the TestLink track of IberLEF 2023 at SEPLN 2023. TestLink2023 encompasses a relation extraction task that is subdivided into two distinct subtasks. The initial subtask involves the identification of laboratory test results and measurements within clinical case documents, while the second subtask focuses on accurately establishing relations between the identified results and their respective measurements. Our team successfully tackled both subtasks by employing language model-based systems, though the result was worse than expected. Our best system achieved an F-Score of 0.17 in the combined task, which is below the presented baselines. However, we have identified that the main weakness of the model lies in the first subtask. This finding indicates that we have the potential to improve upon the presented solution.

Keywords

Entity Recognition, Relation Extraction, Clinical cases, Laboratory tests

1. Introduction

Accurate identification of substances and measurements in laboratory results is highly significant due to various reasons. From the patient's point of view, accessing laboratory test results is a common activity on patient portals, nevertheless they often struggle to comprehend the meaning of these results ([1]). Consequently, Developing tools that aid patients in better understanding laboratory results and formulating relevant questions for physician consultations is very relevant and to that end substance and measurement identification is crucial.

The incorporation of laboratory test entities and measurements has demonstrated its utility and positive impact on several tasks. For instance, during the peak of the pandemic, numerous studies recognized the importance of integrating blood test information to aid in identifying COVID-19 cases when PCR tests were unavailable or had prolonged turnaround times ([2], [3], [4],[5]. Moreover, laboratory test information plays a relevant role in the early detection of certain pathologies such as sepsis, contributing to accurate predictive outcomes ([6]).

Over the years, several methods have been proposed to address this task, ranging from symbolic to traditional Machine Learning methods, and Language Model based methods. [7] compared symbolic systems to traditional machine learning methods (CRFs, SVMs, HMMs).

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

†These authors contributed equally.

acuadron001@ikasle.ehu.eus (A. Cuadrón*†); asagasti036@ikasle.ehu.eus (A. Sagasti*†); ander.barrena@ehu.eus (A. Barrena); aitziber.atucha@ehu.eus (A. Atutxa)

0000-0003-2024-0362 (A. Barrena); 0000-0003-4512-8633 (A. Atutxa)

They concluded that although CRF performed better in extracting units of measures, general machine learning systems did not outperform tailored symbolic method. In the last years, several medical domain Language Models ([8],[9],[10]) have outperformed traditional machine learning methods widely improving previous Named Entity Recognition (NER) and relation extraction results in the biomedical domain. Precisely in laboratoty test NER works like [11] reporting a result of 0.89 F-Score using different medical domain language models, such as BioBERT, ClinicalBERT, RoBERTa_trial and PubMedBERT. To our knowledge, most of the works attempt to solve the task on English data. Thanks to TestLink [12] track of IberLEF 2023 [13], data in Spanish and Basque is now available to test models in languages other than English. In this work, we present a Language Model fine-tuning approach for the Spanish dataset.

2. Materials and Methods

2.1. Dataset

The E3C Corpus [14] is a valuable resource for Developing and evaluating information extraction systems in the medical domain. This multilingual corpus covers English, French, Italian, Spanish, and Basque, containing medical reports that facilitate training and analysis of such systems. The corpus consists of separate documents (each representing a specific clinical case) annotated with clinical entities (pathologies, symptoms, procedures, and body parts, according to standard clinical taxonomies) and temporal information (events, time expressions and temporal relations, according to the THYME TimeML standard).

Regarding the evaluation and benchmarking, the E3C Corpus is divided into train and Test sets by the competition organizers. The train set consists of 81 documents, containing 28,815 tokens and 597 annotated relations. The relations consist of two elements: a RML(results, measurements and lab test results) and an EVENT (actions, states, and circumstances that are relevant to the clinical history of a patient). Additionally, the Test set comprises 80 documents. The train and Dev data splits were generated randomly for each run using a random seed. As a result, we do not know the exact distribution of the RML and EVENT entities between the train and dev sets. We allocated 80% of the data for training purposes and reserved 20% for the Dev set in our experiments. Consequently, approximately 64 documents were included in the train set, while 17 documents were assigned to the Dev set.

Dataset	Documents	Sentences	RML	EVENT	OTHERS
Train (80%)	64	907	325	400	21712
Dev (20%)†	17	227	130	159	5403

Table 1
Distribution of train and Dev datasets. † represents the random generation of Dev sets and approximate values for each generation.

The dataset format follows a tab-delimited text file structure, providing an organized and accessible format for research purposes. Here is the structure (Table 2) of the train dataset along

with an example (Table 3):

<DOCID>|t|<TEXT>
<DOCID> REL <RML_START>-<RML_END> <EVENT_START>-<EVENT_END> <RML_TEXT> <EVENT_TEXT>
<DOCID> REL <RML_START>-<RML_END> <EVENT_START>-<EVENT_END> <RML_TEXT> <EVENT_TEXT>

Table 2
Structure of the dataset

100177|t|Varón de 74 años Antecedentes personales: HTA esencial previa (En tratamiento con Enalapril), Hipercolesterolemia. Litiasis cálcica de 1 cm en infundíbulo de cáliz superior y otra de las mismas dimensiones en infundíbulo de cáliz inferior de riñón izquierdo. Leoch de litiasis superior (3.900 impactos, a intensidad 3-4). Ecografía renal a la terminación de la litotricia normal, no hematoma detectable. Ocho horas después de la Leoch, dolor lumbar intenso y nauseas que requiere analgesia intensa. Diez horas más tarde abombamiento lumbar y hematoma cutáneo diagnosticándose con ecografía y Tac de hematoma subcapsular con rotura capsular. Palidez cutaneomucosa intensa con Hb de 7,6 y Hto 22, pero hemodinámicamente estable. Se administraron 3 concentrados de hematíes. Permaneció ingresado en hospital 13 días. Se optó por tratamiento conservador. Cuatro meses después el Tac de control es semejante al inicial. Y un año después persiste el hematoma aunque de menor tamaño.
100177 REL 137-141 117-125 1 cm Litiasis
100177 REL 375-381 325-334 normal Ecografía
100177 REL 685-688 679-681 7,6 Hb
100177 REL 695-697 691-694 22 Hto

Table 3
Example of a document in the dataset

In addition to the original dataset, the organizers also provided a tokenized version of the corpus. This tokenization process was performed to assist in training our models and accurately retrieve the indices of each word in the text. For each document in the original dataset, a separate file was created with the tokenized representation. The format of each file followed a tab-delimited structure, ensuring consistency and compatibility with the original dataset.

This brief example (Table 4) provides a clear illustration. The numbers on the first column indicate the sentence number and the number of the word inside the sentence. For instance, 1-1 means the first word of the first sentence. The second column indicates the number of characters where the word begins and ends. In the third column we can find the word itself.

2.2. Models

BERT models [15] have emerged as a groundbreaking approach for information extraction tasks. Leveraging the power of transformer-based architectures. Specifically, BERT models can

sent-idx	offset	word
1-1	0-5	Varón
1-2	6-8	de
1-3	9-11	74
1-4	12-16	años
1-5	18-30	Antecedentes
...	...	...

Table 4
Example of a tokenized document

efficiently capture both the left and right context of a given word or phrase. This ability makes BERT models exceptionally effective for tasks such as named entity recognition, document classification and relation extraction. By pre-training on large corpora and subsequently fine-tuning on specific downstream tasks, BERT models have demonstrated remarkable performance in various domains, surpassing previous state-of-the-art methods. These models have significantly advanced the field of information extraction.

Below, we briefly present the three different models we have fine-tuned to carry out both the NER and the document classification tasks, which we will explain in more detail later on.

- BETO [16] model is a BERT model that was pretrained on a big Spanish corpus, containing more than 300M lines of information gathered from different sources.
- BSC-bio-ehr-es [17] is RoBERTa-based, and was pre-trained on biomedical language by the Barcelona Supercomputing Center. According to the official documentation, the bsc-bio-ehr-es model can be used for EHR documents and clinical notes.
- BSC-bio-ehr-es-Pharmaconer is a variant of the previously mentioned model, fine-tuned for the NER task using the PharmaCoNER dataset, which is according to the documentation the largest available biomedical corpus to date.

2.3. Methods

We have divided the TESTLINK task of identifying relations in medical tests in two subtasks. Task 1 involves identifying entities, namely RML and EVENT, in the medical texts. Task 2 will then focus on detecting the existing relationships between those entities. Figure 1 provides a representation of the system we have designed to participate in the competition.

As we have mentioned, the first task aims to extract the entities from the given medical texts. To accomplish this, we have fine-tuned three different pre-trained models, BSC-bio-ehr-es, BSC-bio-ehr-es-Pharmaconer, and BETO, to perform the NER task. NER (Named Entity Recognition) is a natural language processing task consisting in identifying and classifying entities with specific names in a text. These entities can include names of people, organizations, locations, dates, quantities, among others. In our case, the entities are RML and EVENT.

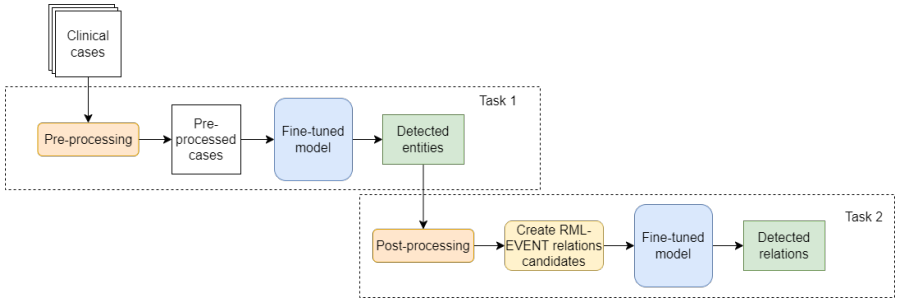

Figure 1: The system pipeline is represented as follows: Given a clinical case, we preprocess the data into BIO format. Next, we fine-tune a BERT-based model to detect the entities. Subsequently, we postprocess the output to further refine the relation extraction module. The pipeline proposed for this task combines the outputs of both models.

Before fine-tuning, we initially had to preprocess the available training data. We utilized both the annotated documents and the tokenized documents to create a dataset in the CoNLL2003 format, containing a word and its corresponding BIO tag per line. This dataset served as the input data for the fine-tuning of the three models.

Second, Relation Extraction enables determining the veracity of relations, in our case between EVENTs and RMLs, in each document. It can assign a label like positive, negative or neutral to a sequence of text. In this case, we will only be interested in positive (1) and negative (0) labels. To reach this, we make use of two pre-trained models, namely BSC and BSC-PharmaCoNER.

Preceding the fine-tuning stage, certain preprocessing steps must be performed on the training data. It is imperative to annotate each relation within the document to identify positive relations accurately. Moreover, to ensure a robust training process, it is essential to introduce artificially generated negative relations. To achieve this, we combine some RML with events that were not actual relations in the training dataset. Subsequently, each relation is assigned a value of 1 for positive instances and 0 for negative instances within the same line separated with a tabulation. This deliberate approach is indispensable for the model's comprehension and training effectiveness. The output of each document in the dataset is assigned a label of either positive or negative, indicating whether the relation should be considered or not.

After obtaining the top models for each task, we proceeded to combine the two parts as a pipelined system for the test, so we could input the documents of the test dataset and obtain as a result the annotated documents, as shown in Figure 1. However, we realized that the output of the first model did not meet the desired quality. In this case, the extracted relations were separated by the model's tokens, and each word was divided into multiple parts. Consequently, when inputting this output into the second model, the results were unsatisfactory as we obtained fragmented relations.

A post-processing step was necessary to refine the initial result. In Figure 1, this step is considered as the pre-processing phase of the second task. During the processing of the initial task's

Model name	Training hyperparameters	Accuracy	Precision	Recall	**F-Score**
Pharmaconer	learning rate: 7.5e-05 batch size:32 epochs: 30	97.58	75.16	69.01	**71.95**
BSC	learning rate: 6e-05 batch size:32 epochs: 32	96.93	66.50	64.22	**65.34**
BETO	learning rate: 5e-05 batch size:32 epochs: 32	97.27	59.90	65.41	**62.53**

Table 5
Results on the Dev set for Task 1. We report the best hyper parameter combination, Accuracy, Precision, Recall and F-Score.

Model name	Training hyperparameters	Accuracy	Precision	Recall	**F-Score**
BSC	learning rate: 2.5e-05 batch size:16 epochs: 2	98.62	98.59	98.62	**98.60**
Pharmaconer	learning rate: 5.5e-05 batch size:16 epochs: 3	97.14	97.04	97.14	**97.08**

Table 6
Results on the Dev set for Task 2. We report the best hyper parameter combination, Accuracy, Precision, Recall and F-Score.

result, we merged the fragmented segments of each word, resulting in a clean file suitable for inputting into the second task. Subsequently, the final outcome met the anticipated expectations.

The results of the models' training for Task 1 (NER) are displayed in Table 5. The numbers in the table represent the performance of the best model achieved, reaching 0.72 of F-Score in the best case. Regarding Task 2 (Relation Extraction), the training results are presented in Table 6. The systems use the entities of the gold standard to predict the relations, and they would extract relations with a 0.98 of F-Score. We have selected the BSC model for Task 2 due to its performance, as indicated by the remarkably high F-Score. Apart from that, there is minimal difference compared to the other model.

Considering these results, we have opted to utilize the Pharmaconer and BETO models for the competition, as we were only able to select two models.

3. Results and Discussion

3.1. Results

From the three models we have trained, we submitted two runs: BETO and Pharmaconer. In Table 7, we present the results obtained by our models on the Test set, as well as the baseline provided by the organization.

Model name	Precision	Recall	**F-Score**
mBERT BASELINE	61.13	60.03	**60.57**
vocabulary transfer BASELINE	17.41	30.24	**22.10**
BETO	43.27	11.08	**17.64**
Pharmaconer	34.84	11.53	**17.32**

Table 7
Precision, Recall and F-Score results for baseline and our models on the Test set.

The mBert baseline approach is similar to our approach, employing two fine-tuned mBert models. The first model is a multilingual BERT model fine-tuned on the entities using the provided training data, enabling it to recognize RMLs and EVENTs. The second model is another multilingual BERT model fine-tuned on relations to identify the relationships between entities. As for the vocabulary transfer baseline approach, it recognizes references to tests by identifying entities from the training data and incorporates regular expressions. For relation extraction, a relation is created for each pair of EVENT and RML that appear together within the same sentence.

The results from both runs we have submitted exhibit a high degree of similarity, with the primary distinction lying in precision, where the BETO model demonstrates a slightly superior performance. Conversely, recall and F-Score show substantial similarity. It is evident that the F-Score of our models falls significantly below the provided baselines. This discrepancy primarily stems from a notably low recall. Despite the precision surpassing that of the vocabulary transfer run and being lower than the mBERT baseline, the limited recall adversely impacts the overall F-Score.

3.2. Discussion

As we have mentioned, our approach is the same as the mBert baseline, so our model was expected to achieve a similar result. To identify the root cause of the low F-Score in both of our runs, we need to examine the performance of both tasks individually. It is possible that the underperformance could stem from either the first task, the second task, or potentially both. Therefore, firstly we analyzed the results from the first task to determine whether the entity extraction process is the underlying cause.

Based on these results, it is evident that the performance of both models in the first task is significantly lower than anticipated. During training, we achieved results exceeding 0.60 (Table 5); however, when evaluating on the Test set, the obtained results are not even half of that

Model name	Type	Precision	Recall	**F-Score**
BETO	ALL	56.78	19.28	**28.79**
	Event	67.48	17.13	**27.33**
	RML	49.36	21.89	**30.33**
Pharmaconer	ALL	54.44	24.57	**33.86**
	Event	62.56	22.12	**32.68**
	RML	48.34	27.55	**35.10**

Table 8
Results obtained on the Test set for Task 1

benchmark. Consequently, we can conclude that the performance of both models in the first task is notably poorer than expected.

In the case of task 2, the overall results are remarkably similar to those obtained in task 1. Consequently, it can be concluded that task 2 does not have a significant negative impact on the overall outcome. Additionally, the F-Score achieved during training for task 2 is notably high, as indicated in Table 6.

In conclusion, in order to enhance the results, it is crucial to focus on improving the performance of Task 1 as a priority. This can be achieved by either searching for the optimum hyper-parameter combination or alternative models as potential solutions.

Acknowledgments

This work has been partially funded by the Basque Government (IXA excellence research group (IT1343-19) and the projects DOTHHEALTH and ANTIDOTE.

References

[1] Z. Zhang, D. Citardi, A. Xing, X. Luo, Y. Lu, Z. He, Patient challenges and needs in comprehending laboratory test results: Mixed methods study, Journal of Medical Internet Research 22 (2020) e18725. doi:10.2196/18725.

[2] A. Batista, J. Miraglia, T. Donato, A. Filho, Covid-19 diagnosis prediction in emergency care patients: a machine learning approach (2020). doi:10.1101/2020.04.04.20052092.

[3] L. Muhammad, M. M. Islam, S. S. Usman, S. I. Ayon, Predictive data mining models for novel coronavirus (covid-19) infected patients' recovery, SN Computer Science 1 (2020) 206.

[4] Q. Ruan, K. Yang, W. Wang, L. Jiang, J. Song, Clinical predictors of mortality due to covid-19 based on an analysis of data of 150 patients from wuhan, china, Intensive Care Medicine 46 (2020). doi:10.1007/s00134-020-05991-x.

[5] S. Cappabianca, F. Roberta, d. Angela, P. Cesare, A. Clemente, G. Gagliardi, L. Giulio, G. Giacobbe, G. M. Russo, M. P. Belfiore, F. Urraro, G. Roberta, F. Beatrice, V. Miele, Clinical and laboratory data, radiological structured report findings and quantitative evaluation

of lung involvement on baseline chest ct in covid-19 patients to predict prognosis, La radiologia medica 126 (2020) 1–11. doi:10.1007/s11547-020-01293-w.

[6] M. Y. Yan, L. T. Gustad, Nytrø, Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review, Journal of the American Medical Informatics Association 29 (2021) 559–575. doi:10.1093/jamia/ocab236.

[7] Y. Kang, M. Kayaalp, Extracting laboratory test information from biomedical text, Journal of pathology informatics 4 (2013) 23. doi:10.4103/2153-3539.117450.

[8] 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 36 (2019) 1234–1240. doi:10.1093/bioinformatics/btz682.

[9] K. Huang, J. Altosaar, R. Ranganath, Clinicalbert: Modeling clinical notes and predicting hospital readmission, arXiv:1904.05342 (2019).

[10] 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, 2020. arXiv:arXiv:2007.15779.

[11] Z. He, A. Erdengarsileng, S. Tian, K. Hanna, Y. Gong, X. Luo, Z. Zhang, M. L. A. Lustria, Towards semi-automated construction of laboratory test result comprehension knowledge-base for a patient-facing application, in: medRxiv, 2022.

[12] B. Altuna, R. Agerri, L. Salas-Espejo, J. J. Saiz, R. Zanoli, M. Speranza, B. Magnini, A. Lavelli, G. Karunakaran, Overview of TESTLINK at IberLEF 2023: Linking Results to Clinical Laboratory Tests and Measurements, Procesamiento del Lenguaje Natural 71 (2023).

[13] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org, 2023.

[14] B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli, 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 (2020).

[15] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[16] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[17] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in Spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: https://aclanthology.org/2022.bionlp-1.19. doi:10.18653/v1/2022.bionlp-1.19.