

URJC-Team at MEDDOPLACE 2023: Bi-LSTM and Transformers for Medical Document Place-Related Content Extraction

David Roldán-Álvarez¹, Miguel Ángel Rodríguez-García¹, Soto Montalvo-Herranz¹
and Raquel Martínez-Unanue²

¹Universidad Rey Juan Carlos, Spain

²Universidad Nacional de Educación a Distancia, Spain

Abstract

In the past few years, the exponential increase of clinical information and the usage of electronic medical records motivated the application of automatic processing techniques for information extraction, becoming a hot research topic for the community. In this sense, the MEDDOPLACE track proposes several sub-tasks related to the location, normalization and classification of various kinds of entities in medical documents in the Spanish language. This article describes the system proposed for the two sub-tasks linked to localization and classification. The built system combines Recurrent Neural Network with Language Models to face the sub-tasks. Concretely, we employed a Bidirectional Long Short-Term Memory (BiLSTM) in conjunction with the Transformer model pre-trained RoBERTa for the location sub-task and the pre-trained BERT model for sequence classification on the second sub-task. Although the system got high positions on the leaderboard, the outcomes reveal a not-to-high performance of the built system, exposing a high margin of improvement.

Keywords

Deep Learning, Transformers, Natural Language Processing, Named Entity Recognition

1. Introduction

With the arrival of electronic format in hospitals, clinical text has increased drastically [1]. This text contains a large amount of relevant information like places, facilities, nationalities, patient movement, diseases, drugs, diagnoses, and treatment plans, among others [2, 3]. Hence, clinical text mining is increasingly used in various domains due to its practical applications on diagnosis prediction, reducing the gap between unstructured and structured information to provide more rich datasets, and identifying human health risk assessments [4, 5, 6].

In health care, several studies have demonstrated that narrative text is more expressive, allowing experts to describe patients' stories accurately [7]. Hence, it is demanded to develop tools that process this valuable narrative text and extract helpful knowledge to assist practitioners [8]. In this sense, concept extraction is a subdomain of Natural Language Processing (NLP)


IberLEF 2023, September 2023, Jaén, Spain

✉ david.roldan@urjc.es (D. Roldán-Álvarez); miguel.rodriguez@urjc.es (M. Á. Rodríguez-García);
soto.montalvo@urjc.es (S. Montalvo-Herranz); raquel@lsi.uned.es (R. Martínez-Unanue)

🆔 0000-0001-7049-7460 (D. Roldán-Álvarez); 0000-0001-6244-6532 (M. Á. Rodríguez-García); 0000-0001-8158-7939
(S. Montalvo-Herranz); 0000-0003-1838-632X (R. Martínez-Unanue)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

employed computationally to analyze this unstructured textual data [9]. This technique aims to identify related clinical concepts, which is a keystone to identifying clinical information, organizing data into suitable representation for efficient analysis, and concentrating the useable knowledge dispersed by healthcare professionals into various format files like clinical reports and monitoring sheets [10, 11]. In this context, the MEDDOPLACE proposes the track [12]: a set of shared sub-tasks that demands the development of tools for MEDical DOcument PLAcE-related Content Extraction. In this work, we describe the contribution of the two sub-tasks which aim to locate and classify various types of entities. The proposed system combines three Deep Learning Neural Network architectures, RoBERTa and BERT Transform models, and Bidirectional Long Short-Term Memory (BiLSTM).

The remainder of the manuscript is organised as follows. Section 2, where the MEDDOPLACE's sub-tasks faced are described. Section 3 details the hybrid system proposed for the challenge. Section 4 presents and discusses the results achieved in the challenge. Finally, Section 5 summarises the findings harvested facing this challenge.

2. Task and dataset description

The shared task MEDDOPLACE, organized at IberLEF 2023 workshop, aims to detect different kinds of places and related types of information, such as nationalities or patient movements, in medical documents in Spanish. Specifically, it proposes four challenges. *Sub-task 1: Location Entity Recognition* is an entity recognition task to detect mentions of locations and location-related entities. *Sub-task 2: Geographic Normalization* aims at normalizing found locations in Sub-task 1 by linking them to GeoNames, PlusCodes or SNOMED-CT codes. *Sub-task 3: Entity Classification* is defined as a multi-class classification problem, where each entity located has to be classified into these four classes of clinical relevance: (a) the patient's origin place; (b) the patient's residence's location; (c) a place where the patient has travelled to or from; (d) a place where the patient has received medical attention. Finally, *Sub-task 4: End-to-End Evaluation* is a task where all three tasks above are performed sequentially instead of being evaluated individually. We participated in sub-tasks 1 and 3.

Concerning the MEDDOPLACE corpus is a collection of 1,000 clinical case reports (635,785 tokens) in Spanish from different medical specialties, but only 750 were provided as training sets. The corpus was annotated following the guidelines of the annotation scheme created by clinical and linguistic experts, who reviewed several location-related documents for defining an annotation scheme, detailed enough and specialized to the clinical domain.

3. Methodology

The system presents a modular architecture of four modules: (i) pre-processor responsible for encoding the training dataset into recognizable numbers by Deep Learning architectures. It takes the text and its annotations and changes their representation into a numerical distribution; (ii) two named entity-recognition models for location detection used for sub-task 1; (iii) a BERT model that classifies entities found by the previous module for sub-task 3, (iv) the evaluator, which quantifies the performance of both modules. It uses the test dataset for assessing the

performance of each sub-task module. Figure 1 depicts the system's modular structure built for the sub-tasks 1 and 3.

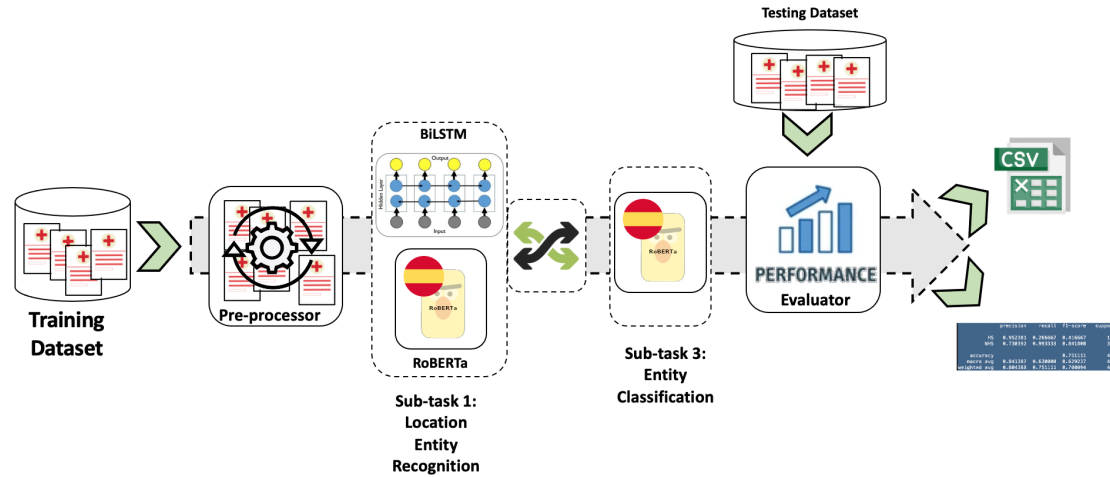


Figure 1: Architecture of the proposed system.

The system has been structured as a pipeline of modules since the sub-tasks proposed in MEDDOPLACE have been configured as a list of connected methods to process medical documents and locate and classify specific entities. Thus, the system works as follows. Firstly, the annotating files are read and a new training dataset that combines the texts with the annotated words is created. Then, the pre-processor module analyses the text and incorporates the annotations following the BIO / IOB format (short for inside, outside, beginning). Next, the resulting text, words and embedded annotations are encoded into a numerical distribution that can be understood by neural networks. This mathematical representation is given to the modified RoBERTa Transformer model for multilabelling and the BiLSTM located in sub-task 1 module to carry out the entity location. The RoBERTa model was configured by the following hyperparameters: 514 for max_position_embeddings, a batch size of 32, 15 epochs, a learning rate of 3e-5 and "adamw_hf" as the optimizer. In the case of BiLSTM, the setting was 250, 256, 25, and "rmsprop" configured by default, in the word embedding length, batch size, number of epochs, and the optimizer, respectively. As a result, each model returns its own list of results, which are combined into one file, dropping the duplicated annotations. Thus, a list of entities is provided in conjunction with their labels and offsets, which specifies the start and end position inside the sentence. Then, each entity found with its sentence is provided to the next module to proceed to classify them. In sub-task 3, we employed the BERT Transformer for sequence classification to categorize the entities found. The setting utilized in this model was: 512 for max_position_embeddings, a batch size of 32, 5 epochs, a learning rate of 1e5 and "adamw_hf". Finally, the evaluation module receives the test dataset and computes the performance of both modules by employing the evaluation metrics.

4. Results

The metrics selected to evaluate the performance of the participant systems are precision, recall and F1-score. The outcomes obtained by the proposed methods is studied independently. Hence, regarding sub-task 1, one of the main limitations of our solution is that, when using transformers, to achieve maximum precision it is needed to have a balanced dataset, it means that, for this particular task, there should be a similar number of annotations for each specific label. The provided annotated dataset did not meet this requirement, since we can find labels such as GPE_NOM, DEPARTAMENTO and FAC_GEN with more than 1500 appearances, while other labels such as IDIOMA and GEO_NOM appear less than 50 times. This situation resulted in accuracies close to 0 in those tags, which ended up reducing the Macro F1 scores. To resolve this limitation, we included a BiLSTM architecture since studies demonstrated its good performance on named entity locations [13]. Therefore, we re-designed the system for that two architectures carried out the sub-task 1 independently. Thus, the resulting entities list is made of a combination of both outcomes. The results for each tag are depicted in Table 1. The columns represent the tag, the precision, the recall, the f-score, the overlapping precision, the overlapping recall and the overlapping f-score.

Table 1
Sub-task 1 scores

Tag	Precision	Recall	F-score	O_precision	O_recall	O_f
DEPARTAMENTO	0.48	0.69	0.56	0.51	0.74	0.60
FAC_GEN	0.37	0.57	0.45	0.39	0.61	0.48
GPE_NOM	0.56	0.53	0.55	0.59	0.56	0.58
FAC_NOM	0.42	0.61	0.49	0.56	0.82	0.67
TRANSPORTE	0.48	0.45	0.46	0.52	0.48	0.50
COMUNIDAD	0.59	0.21	0.31	0.59	0.21	0.31
GPE_GEN	0.22	0.42	0.29	0.34	0.63	0.44
GEO_GEN	0.33	0.58	0.42	0.37	0.65	0.47
GEO_NOM	0.55	0.50	0.53	0.55	0.50	0.53
TOTAL	0.43	0.57	0.49	0.47	0.63	0.53

Concerning the sub-task 3, we followed a similar approach to address sub-task, and faced similar problems as in sub-task 1. After analysing the dataset, we found that the number of classes was unbalanced, finding locations with the class ATENCION a total of 1387 times, while locations with the class LUGAR-NATAL only appeared 333 times. However, for this sub-task, we tried to balance the dataset by adding new sentences generated using ChatGPT (<https://chat.openai.com/>). We oriented the platform to create sentences for the corresponding labels, including the indexes where the word that represent the location is within the sentence and including their corresponding classes for the tags RESIDENCIA, MOVIMIENTO and LUGAR-NATAL, resulting in a final training dataset where every single class had around 1300 sentences

in which they were represented. In order to create the sentences, we needed to orient the platform towards understanding the context of the tasks at hand. Once the context was given, we instructed Chat-GPT to create examples for each tag. Below, it is shown an example created with this generative model.

"I want long medical sentences where locations appear. For each sentence, you need to provide me with information about what the location represents. The locations can represent 5 different type of things: Locations where people are cared for (hospitals, health centres, etc) , locations that represent the residence of the patient (House, chalet, etc) , locations that represent the patients birthplace (Spain, France, Germany, depending on the context) , a location that represents a movement from one place to another and other locations."

Once we reviewed that the provided sentences, locations and classes were correct, we proceeded to tell the platform to provide 1000 sentences for each tag following a .csv format and including the indexes where the word representing the location appeared in each sentence. Below, an example of the prompt generated to query the Chat-GPT is shown and several examples produced are compiled in Table 2.

"Now I need you to give me the data in a .csv format separated by commas. I need these columns: sentence (the full sentence), word (the word that represents the location), class (the class of the location), start (the start index of the location) and, end (the end index of the location)."

Table 2

Chat-GPT output for sub-task 3

sentence	word	class	start	end
Después de un largo día de trabajo, el paciente regresó a su acogedora casa en el tranquilo vecindario.	casa	RESIDENCIA	71	75
"El paciente se desplazó en moto de agua para explorar la costa y sentir la emoción del mar."	"moto"	MOVIMIENTO	27	31
"Hombre de 33 años, natural de Francia, que ha viajado por toda Europa por su trabajo como consultor"	"Francia"	"LUGAR-NATAL"	30	37

The results for each tag in sub-task 3 are presented in Table 3.

5. Conclusions

This article details the system proposed for the shared sub-tasks included in the MEDDOPLACE track, which provides varied resources for identifying, normalizing and classifying location entities in medical texts in the Spanish Language. In particular, the proposed system has been designed for the identification and classification sub-tasks. For the first one, we combined two Deep Learning Neural Network Architectures based on BiLSTM and RoBERTa Transformer model. In the second one, we employed BERT for sequence classification model to classify locations into their corresponding classes. The discovery of the issue of disparity in the classes in the provided datasets motivated us to design a strategy to try to deal with it. This strategy

Table 3
Sub-task 3 scores

Tag	Accuracy
ATENCION	0.33
MOVIMIENTO	0.3
RESIDENCIA	0.39
LUGAR-NATAL	0.41
OTHER	0.29
TOTAL	0.33

was based on integrating the ChatGPT in the resulting system pipeline to generate samples that help us to reduce such disparity. Despite the efforts carried out the performance reached by the system was not too much affected, meaning it is still work to accomplish with the proposed methods and the dataset provided.

In future work, we would like to try more complex Neural Network Architectures to assess their performance. Furthermore, one limitation that breaks the current system’s performance is the reduced number of cases provided on the datasets. Despite we have tried to deal with this problem by employing a generative model without a clear improvement, we believe that employing other augmentation techniques could enhance the systems’ performance, and a more detailed analysis of the data could provide clues about the issues affecting the methods proposed.

6. Acknowledgments

This work has been partially supported by projects DOTT-HEALTH (PID2019-106942RB-C32, MCI/AEI/FEDER, UE), GELP (TED2021-130398B-C21, MCIN/AEI/10.13039/501100011033), “NextGenerationEU”/PRTR (funded by European Union), grant “Programa para la Recualificación del Sistema Universitario Español 2021-2023”, and the project M2297 from call 2022 for impulse projects funded by Rey Juan Carlos University.

References

- [1] D. Ningthoujam, S. Yadav, P. Bhattacharyya, A. Ekbal, Relation extraction between the clinical entities based on the shortest dependency path based lstm, arXiv preprint arXiv:1903.09941 (2019).
- [2] Z. Dai, X. Wang, P. Ni, Y. Li, G. Li, X. Bai, Named entity recognition using bert bilstm crf for chinese electronic health records, in: 2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei), IEEE, 2019, pp. 1–5.
- [3] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng,

- S. Mehrabi, S. Sohn, et al., Clinical information extraction applications: a literature review, *Journal of biomedical informatics* 77 (2018) 34–49.
- [4] R. Weegar, Mining Clinical Text in Cancer Care, Ph.D. thesis, Department of Computer and Systems Sciences, Stockholm University, 2020.
- [5] P. Belloni, G. Boccuzzo, S. Guzzinati, I. Italiano, C. R. Rossi, M. Rugge, M. Zorzi, Staging cancer through text mining of pathology records, in: *Data Science and Social Research II: Methods, Technologies and Applications*, Springer, 2021, pp. 29–46.
- [6] K. De Silva, N. Mathews, H. Teede, A. Forbes, D. Jönsson, R. T. Demmer, J. Enticott, Clinical notes as prognostic markers of mortality associated with diabetes mellitus following critical care: a retrospective cohort analysis using machine learning and unstructured big data, *Computers in Biology and Medicine* 132 (2021) 104305.
- [7] P. Bhatia, B. Celikkaya, M. Khalilia, S. Senthivel, Comprehend medical: a named entity recognition and relationship extraction web service, in: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, IEEE, 2019, pp. 1844–1851.
- [8] M. Y. Yan, L. T. Gustad, Ø. Nytrø, Sepsis prediction, early detection, and identification using clinical text for machine learning: a systematic review, *Journal of the American Medical Informatics Association* 29 (2022) 559–575.
- [9] S. Fu, D. Chen, H. He, S. Liu, S. Moon, K. J. Peterson, F. Shen, L. Wang, Y. Wang, A. Wen, et al., Clinical concept extraction: a methodology review, *Journal of biomedical informatics* 109 (2020) 103526.
- [10] C. Luque, J. M. Luna, M. Luque, S. Ventura, An advanced review on text mining in medicine, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9 (2019) e1302.
- [11] K. Ganesan, S. Lloyd, V. Sarkar, Discovering related clinical concepts using large amounts of clinical notes: supplementary issue: big data analytics for health, *Biomedical Engineering and Computational Biology* 7 (2016) BECB–S36155.
- [12] S. Lima-López, E. Farré-Maduell, V. Brivá-Escalada, L. Gascó, M. Krallinger, MEDDOPLACE Shared Task overview: recognition, normalization and classification of locations and patient movement in clinical texts, *Procesamiento del Lenguaje Natural* 71 (2023).
- [13] P. Zhang, W. Liang, Medical name entity recognition based on lexical enhancement and global pointer, *International Journal of Advanced Computer Science and Applications* 14 (2023). URL: <https://vpnssl.urjc.es/dana/home/index.cgi/scholarly-journals/medical-name-entity-recognition-based-on-lexical/docview/2807223042/se-2>, copyright - © 2023. This work is licensed under <http://creativecommons.org/licenses/by/4.0/> (the “License”). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Última actualización - 2023-05-02.