

# Two-stage Fine-Tuning for Automatic Identification of Sections in Clinical Documents

Jónathan Heras<sup>1,\*</sup>

<sup>1</sup>*Department of Mathematics and Computer Science, University of La Rioja (Spain)*

## Abstract

Electronic Clinical Narratives (ECN) are the standard for storing relevant information to describe and evaluate a patient's clinical episode or evolution. The ClinAIS task aims to tackle the problem of automatic identification of sections in unstructured Spanish ECNs. In this work, we tackle this challenge by first fine-tuning a language model with the ClinAIS dataset, for later applying a second fine-tuning stage for section identification. The performance of several models was studied using this approach, and a Longformer based model obtained the best results in the validation set. Using this model, we achieved the third position in the ClinAIS challenge with a weighted B2 score of 0.7036 in the test set.

## Keywords

Fine-tuning, ClinAIS, Clinical Documents, HuggingFace

## 1. Introduction

The amount of digitised data available from healthcare systems is increasing exponentially [1]. Among that data, Electronic Clinical Narratives (ECN) have become the standard for storing all the information a practitioner finds relevant to describe and evaluate a patient's clinical episode or evolution. Within these documents, practitioners can find information such as past medical conditions, medical procedures undergone, disease progression, or prescribed treatments. As ECNs have become the standard for data storage, their secondary use has gained prominence in addressing various tasks such as the identification of rare medical events, prediction of hospital readmissions, and public health surveillance.

A fundamental task for the advancement of higher-level applications in healthcare is the accurate identification of medical sections within patient narratives documented in ECNs. This task involves the division of the text into semantic segments and assigning them specific predefined labels. Through section identification, valuable insights can be gleaned regarding different entities, which may vary significantly depending on the section in which they are found. For example, a pathology mentioned in the patient's medical history section could be utilized to predict future conditions and assess the risk of illness. Similarly, the presence of specific symptoms in the Evolution section of the narrative might indicate adverse reactions to a particular treatment.

---

*IberLEF 2023, September 2023, Jaén, Spain*


\*Corresponding author.

✉ [jonathan.heras@unirioja.es](mailto:jonathan.heras@unirioja.es) (J. Heras)

🌐 <https://www.unirioja.es/cu/joheras> (J. Heras)

🆔 0000-0003-4775-1306 (J. Heras)

© 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)

The ClinAIS task presented at IberLEF 2023 [2] aims to tackle the problem of automatic identification of sections in unstructured Spanish clinical documents [3]. The task is focused on identifying seven predefined medical sections: Present Illness, Derived from/to, Past Medical History, Family history, Exploration, Treatment and Evolution. In this work, we tackle this challenge by first fine-tuning a language model with the ClinAIS dataset, for later applying a second fine-tuning stage for section identification. The code of this project is available at <https://github.com/joheras/ClinAIS>.

## 2. Dataset Description

The ClinAIS dataset’s corpus was obtained from the CodiEsp dataset, which was presented in the eHealth CLEF 2020 task [4]. The CodiEsp dataset is a corpus of unstructured clinical case reports from different medical specialties, and it contains 1000 annotated documents for Named Entity Recognition and 2751 unannotated documents as a background set. From the CodiEsp dataset, 1038 distinct notes were randomly selected to form the ClinAIS dataset [5].

For annotation, a set of guidelines were initially created to identify patterns and categorize each section in unstructured clinical notes into 7 categories: Present Illness, Derived from/to, Past Medical History, Family history, Exploration, Treatment and Evolution. From those guidelines, a group of experts went through several rounds of annotating a small set of notes and updating the guidelines accordingly. When the annotation process became more mature, two doctors, trained in clinical report annotation for different tasks, performed a double annotation on the notes. The annotation task was iterative and the evaluation metric was employed to measure the inter-tagger agreement, reaching 75%.

Once the dataset was annotated, the 1038 notes were split into three groups: training (75%, 781 notes), validation (12.5%, 127 notes), and test (12.5%, 130 notes) sets. The training and validation sets were publicly released with their annotations, whereas the annotation of the test set was kept private. The evaluation on the test set was conducted by the judges of the competition.

## 3. Methods

The approach followed here for section identification is based on the transfer learning method proposed in the ULMFIT paper [6]. Given a language model pre-trained on a large dataset, the ULMFIT method consists of two stages. In the first stage, the language model is specialized to a particular context by fine-tuning it with a small dataset of text; after that, the specialised language model is trained to tackle a particular task. In this work, we used the training set of the ClinAIS dataset as small text dataset for the first stage — obtaining in this manner, language models specialized into clinical case reports. Subsequently, for the second stage, we trained those specialized models for the task of clinical section identification using the ClinAIS dataset.

For our work, we have studied several pre-trained language models available at HuggingFace. In particular, we considered two versions of the multi-lingual XLM-RoBERTa model (base and large versions) [7]; a DistilBert Spanish model [8] trained on the Large Spanish Corpus [9]; two RoBERTa-based models one trained on a biomedical-clinical corpus in Spanish collected from

**Table 1**

Models used as basis in the study.

Model	Link
XLM-RoBERTa base	<a href="https://huggingface.co/xlm-roberta-base">https://huggingface.co/xlm-roberta-base</a>
XLM-RoBERTa large	<a href="https://huggingface.co/xlm-roberta-large">https://huggingface.co/xlm-roberta-large</a>
DistilBert Spanish	<a href="https://huggingface.co/dccuchile/distilbert-base-spanish-uncased">https://huggingface.co/dccuchile/distilbert-base-spanish-uncased</a>
RoBERTa biomedical-clinical	<a href="https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es">https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es</a>
RoBERTa clinical	<a href="https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es">https://huggingface.co/PlanTL-GOB-ES/roberta-base-biomedical-clinical-es</a>
Longformer	<a href="https://huggingface.co/PlanTL-GOB-ES/longformer-base-4096-bne-es">https://huggingface.co/PlanTL-GOB-ES/longformer-base-4096-bne-es</a>

several sources and the other trained on a clinical dataset [10]; and a Longformer version of the roberta-base-bne masked language model for the Spanish language [11]. The links to the base models are provided in Table 1

All the models were trained using the functionality provided by the HuggingFace libraries [12] and using a GPU Nvidia GeForce 3090. The hyperparameters used for training the models can be checked in the code available on the project webpage. The validation set was used to evaluate the performance of the models and to select the model that was finally used to perform the predictions in the test set. The metric employed to evaluate the models is an adaption of the boundary distance  $B$  developed by C. Fournier [13] called weighted B2 metric – details about this metric can be found in [3, 5].

Given a sentence, the final models classify each token into B-Class, E-Class or Class – where Class corresponds with one of the 7 predefined medical sections of the ClinAIS dataset, B-Class indicates the beginning of a section, E-Class indicates the end of a section, and Class means that the token is inside a section. Since such an output might produce inconsistencies (for instance a token classified as Class2 inside a sentence that started with B-Class1), the output of the model is post-process to deal with those inconsistencies.

## 4. Results

We start by evaluating the trained models, but we skip the first stage of fine-tuning with the ClinAIS dataset. Instead, we solely train the models for the task of section identification. Please refer to Table 2. The best model using this approach, with a weighted B2 score of 0.7341, was a Longformer-based model combined with the post-processing step. If we analyze the rest of the models, the 3 best models were initially pre-trained with a biomedical Spanish dataset; whereas, the other models, which were trained with generic Spanish or multi-lingual datasets, obtained considerably worse results. It is also worth mentioning the importance of the post-processing stage, since the plain output of all models was inferior to their post-processed counterparts.

In our second set of experiments, we evaluated the models trained using the two-stage approach presented previously, see Table 3. In general, all the models, except for the Longformer-based model, obtained better results than their counterparts trained only for section identification. We can draw similar conclusions to those noticed in the first set of experiments: the models initially pre-trained with biomedical Spanish datasets obtained better results, and the post-processing stage improved the performance of the models. However, the best performing

**Table 2**

One stage fine-tuning. In bold, the best result.

	plain output	post-process
distilbert-base-spanish-uncased	0.4008	0.5908
xlm-roberta-base	0.4078	0.5952
xlm-roberta-large	0.5695	0.6168
roberta-base-biomedical-clinical-es	0.3859	0.6968
bsc-bio-ehr-es	0.0074	0.7175
longformer-base-4096-bne-es	0.6338	<b>0.7341</b>

**Table 3**

Two stage fine-tuning. In bold, the best result.

	plain output	post-process
distilbert-base-spanish-uncased	0.3690	0.6095
xlm-roberta-base	0.47913	0.6178
xlm-roberta-large	0.5919	0.6276
roberta-base-biomedical-clinical-es	0.2873	0.7172
bsc-bio-ehr-es	0.2188	<b>0.7311</b>
longformer-base-4096-bne-es	0.6438	0.7174

**Table 4**

Two stage fine-tuning with augmented dataset

	plain output	post-process
distilbert-base-spanish-uncased	0.0980	0.5944
xlm-roberta-base	0.5712	0.6048
xlm-roberta-large	0.5997	0.6618
roberta-base-biomedical-clinical-es	0.3299	0.7168
bsc-bio-ehr-es	0.0074	<b>0.727</b>
longformer-base-4096-bne-es	0.6315	0.6985

model, a RoBERTa-based model with a weighted B2 score of 0.7311, achieved worse results than those obtained by the Longformer-based model trained only on section identification.

Finally, we applied the two-stage procedure to an artificially augmented version of the ClinAIS dataset. In particular, we hide some of the words of the notes, and used a RoBERTa-based language model to predict those missing words. This allowed us to obtain a training dataset of 2850 notes. The results obtained for this augmented dataset are presented in Table 4. Unfortunately, this approach did not lead to any improvement.

Following this study, we utilized the one-stage fine-tuned Longformer-based model, determined as the best model based on our evaluation, to generate predictions for the test set of the ClinAIS challenge. This model achieved a weighted B2 score of 0.7036.

## 5. Conclusions

In this paper, we have used the two-stage training procedure presented in the ULMFIT work for training several models for medical section identification in unstructured clinical notes. In general, the two stage procedure (first building a specialized language model, and then fine-tuning the model for section identification) provides better results than only training the models for section identification. However, the best model was a Longformer-based model that was only trained for section identification and achieved a final score of 0.7036 in the test set.

## Acknowledgments

This work was partially supported by Ministerio de Ciencia e Innovación [PID2020-115225RB-I00 / AEI / 10.13039/501100011033].

## References

- [1] H. Dalianis, *Clinical text mining: Secondary use of electronic patient records*, Springer Nature, 2018.
- [2] 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.
- [3] I. de la Iglesia, M. Vivó, P. Chocrón, G. de Maeztu, K. Gojenola, A. Atutxa, Overview of ClinAIS at IberLEF 2023: Automatic Identification of Sections in Clinical Documents in Spanish, *Procesamiento del Lenguaje Natural 71* (2023).
- [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: L. Cappellato, C. Eickhoff, N. Ferro, A. Névóel (Eds.), *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: [https://ceur-ws.org/Vol-2696/paper\\_263.pdf](https://ceur-ws.org/Vol-2696/paper_263.pdf).
- [5] I. de la Iglesia, M. Vivó, P. Chocrón, G. de Maeztu, K. Gojenola, A. Atutxa, An Open Source Corpus and Automatic Tool for Section Identification in Spanish Health Records, *Journal of Biomedical Informatics* (2023).
- [6] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, *arXiv preprint arXiv:1801.06146* (2018).
- [7] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, *CoRR abs/1911.02116* (2019). URL: <http://arxiv.org/abs/1911.02116>. *arXiv:1911.02116*.
- [8] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, *ArXiv abs/1910.01108* (2019).

- [9] L. Tunstall, The large spanish corpus, 2022. [https://huggingface.co/datasets/large\\_spanish\\_corpus](https://huggingface.co/datasets/large_spanish_corpus).
- [10] 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](https://arxiv.org/abs/2109.03570).
- [11] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, [arXiv:2004.05150](https://arxiv.org/abs/2004.05150) (2020).
- [12] L. Tunstall, L. Von Werra, T. Wolf, Natural language processing with transformers, "O'Reilly Media, Inc.", 2022.
- [13] C. Fournier, Evaluating text segmentation using boundary edit distance, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013, pp. 1702–1712.