Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 289-299 recibido 26-06-2023 revisado 09-07-2023 aceptado 17-07-2023

Overview of ClinAIS at IberLEF 2023: Automatic Identification of Sections in Clinical Documents in Spanish

Resumen de la tarea de ClinAIS en IberLEF 2023: Identificación Automática de Secciones en Documentos Clínicos en Castellano

Iker De la Iglesia,[1] **María Vivó,**[2] **Paula Chocrón,**[2]
Gabriel de Maeztu,[2] **Koldo Gojenola,**[1] **Aitziber Atutxa**[1]
[1]HiTZ Basque Center for Language Technology (UPV/EHU), Spain
[2]IOMED Medical Solutions SL, Spain
{iker.delaiglesia, koldo.gojenola, aitziber.atutxa}@ehu.eus
{maria.vivo, paula.chocron, gabriel.maeztu}@iomed.es

Abstract: The ClinAIS shared task organized by IOMED and the HiTZ center aims to tackle the identification of seven section types within unstructured clinical records in the Spanish language. These records, known as Electronic Clinical Narratives (ECNs), store crucial individual health information. However, their lack of standardized formats poses challenges in the development and evaluation of automated systems for clinical document analysis. Twenty-seven participants registered for the task, with five submitting results. This paper presents the outcomes and methodologies used in ClinAIS, contributing to the advancement of clinical text analysis and its application in improving healthcare decision-making and patient care.
Keywords: Section Identification, Unstructured Clinical Documents, Language Models, Deep Learning.

Resumen: La tarea ClinAIS organizada por IOMED y el centro HiTZ tiene como objetivo abordar la identificación de siete tipos de secciones dentro de registros clínicos no-estructurados en español. Estos registros, conocidos como Narrativas Clínicas Electrónicas (ECNs), almacenan información crucial acerca de la salud personal. Sin embargo, la falta de estandarización en los formatos plantea desafíos en el desarrollo y evaluación de sistemas automatizados para el análisis de documentos clínicos. Veintisiete participantes se registraron para la tarea, de los cuales cinco presentaron resultados. Este artículo presenta los resultados y metodologías utilizadas en la tarea ClinAIS, contribuyendo al avance del análisis de notas clínicas y su aplicación en la mejora de la toma de decisiones en la atención médica y el cuidado al paciente.
Palabras clave: Identificación de Secciones, Documentos Clínicos No-Estructurados, Modelos de Lenguaje, Aprendizaje Profundo.

1 Introduction

Electronic Clinical Narratives (ECNs) have become the standard for storing all the information a practitioner finds relevant to describe and evaluate a patient's clinical episode or evolution. These documents contain descriptions of previous pathologies, undergone procedures, the evolution of a given disease, or prescribed treatments. Secondary use of ECNs tackles diverse tasks, including identifying rare medical events, predicting hospital re-admissions, or in Public Health Surveillance among others.

Identifying medical sections in ECNs is a crucial task for higher-level applications. Section identification consists in dividing the text into semantically coherent segments categorized with a set of predefined labels.

ISSN 1135-5948 DOI 10.26342/2023-71-22

Section identification provides new insights about entities, which might be completely different depending on the section in which they occur. For example, a pathology referenced in the *patient's medical history* could be used to predict future conditions and risks of illness. Similarly, symptomatology in the *evolution* section could indicate adverse reactions to a given treatment.

The successful resolution of this task will enable the improvement of higher-level applications that can extract valuable, actionable information from clinical documents, such as medical entity recognition, patient cohort retrieval, and temporal relation extraction. This will ultimately improve patient care and clinical decision-making.

The Iberian Languages Evaluation Forum (IberLEF) features several shared tasks providing benchmarks to enable fair comparison across participants' systems. In 2023, the ClinAIS task is devoted to learning the identification of sections in unstructured Spanish clinical documents. 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 ECNs, mainly progress notes.

In this paper, we define the task and evaluation methodology, describe data preparation, analyze the main results, and provide a brief categorization of the approaches used by the participating systems.

In the rest of the paper, after examining related work in Section 2, we describe the datasets and methods used in Section 3. Section 4 is devoted to presenting the main results followed by Section 5 where we discuss the relevant findings, ending with a final section containing the main conclusions.

2 Background

Clinical case reports are unstructured documents narrating the patient's clinical history chronologically, grouping together in sections clauses, sentences, or phrases that describe the different dimensions of a patient. Although in recent years we have seen exciting developments in the processing of clinical texts, with robust and precise tools for tasks like medical Named Entity Recognition and Relation Extraction (Lee et al., 2019), or automatic ICD coding (Atutxa et al., 2019), higher levels of processing to reach the requirements of medical personnel are still an issue. Structuring this information is a crucial step for developing more advanced functionalities. For example, knowing if some symptom belongs to the current patient's health status or to a previous episode (Personal or Family History) is very relevant to the clinician. Similarly, diagnostics, treatments, and procedures might happen within different parts of the document, for example in the *present illness* (description of the reasons that made patient seek medical care), during the physical *exploration*, in the description of the *evolution* and the implications might be different for the care professional. Overall, structuring a document is fundamental for developing any system that can model clinical diagnosis, clinical text understanding, and reasoning (Goenaga et al., 2021; Terroba, 2018; Gao et al., 2022b; Gao et al., 2022a). (Pomares-Quimbaya, Kreuzthaler, and Schulz, 2019) present an exhaustive review of different approaches to identify sections in clinical narratives. Regarding the different languages that have been approached, 78% of the reviewed articles were intended for English. (Goenaga et al., 2021) can be considered an antecedent of this work as it deals with section identification in clinical notes written in Spanish. However, the types of documents differ significantly, because the documents treated in (Goenaga et al., 2021) are semi-structured, in the sense that the texts contain in many cases explicit headings or linguistic clues (paragraphs or line endings) that allow identifying section starts in an easier way, while in the documents presented in this task (see Figure 1) the notes appear without any kind of section header, with several sections contained in a single paragraph, and also with sentences containing more than one section.

As defined in (Gao et al., 2022a), section identification is critical "to train and evaluate future NLP models for clinical text understanding, clinical knowledge representation, inference, and summarization." The task aims, given a clinical document, to detect and classify the parts of text corresponding to each section. To our knowledge this is the first time that this task has been proposed for Spanish, while a similar task has recently been proposed for English (Gao et al., 2022b; Gao et al., 2022a). The currently advanced state of the art in clinical NLP for

Spanish presents an opportunity for the development of this type of tool that can give a competitive advantage over other medical NLP systems.

3 Task Description

The task will consist in identifying seven sections that might appear in a given document. The number and types of sections are derived from the ideas presented in (Terroba, 2018; Goenaga et al., 2021) and adapted to the specific needs of the current task:

- **Present Illness (PI)**. This section contains a brief statement that answers to who, what, where, why, and when, and a detailed description of the symptoms and other relevant issues, as the reason for consultation, including treatments, diagnoses, and explorations performed prior to admission. Anamnesis is also included in this section when collected in the clinical case.
- **Past Medical History / Medical History (MH)** covers past symptoms, medications, diseases or procedures. A mention of the absence of previous medical history is also considered part of this section.
- There is also a specific section for **Family History (FH)**, giving a description of family members' pathologies. If its absence is indicated, it will also be noted as belonging to this section.
- **Exploration (E)** describes observations, including physical examination, vital signs, and muscle power examination of different organs, especially ones that might be related to the symptoms. Additionally, specialized tests, like ECG, laboratory tests, or radiography results. This section includes autopsies and their results as well.
- **Evolution (EV)** describing the evolution of the patient's health status during the current episode. It may include differential diagnoses.
- **Treatment (T)** which contains a plan of the proposed treatment or procedures performed on the patient to treat his condition, including "dieting".
- And, finally, **derived from/to (D)**. This section contains information about another hospital or service where part of the current illness was treated. It can contain a referral to any transfer from/to any department, center, or primary care physician who has made the transfer request and its justification if any.

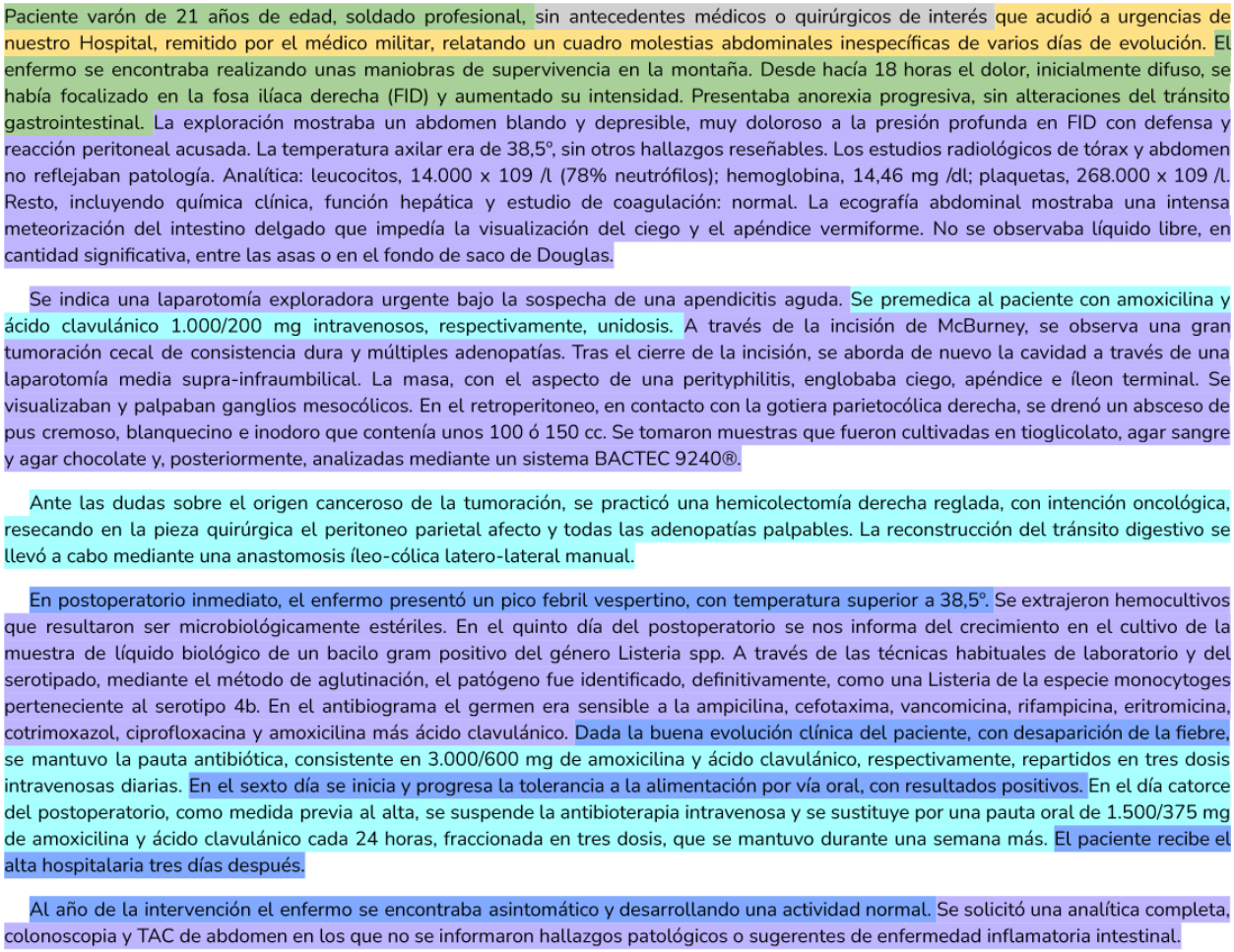
Figure 1 presents an example of an annotated document. Different colors represent the section types: green for Present Illness, gray for Medical History, yellow for "derived from/to", violet for Exploration, turquoise for Treatment, and dark blue for Evolution. The difficulty of the task comes from the fine-grained level of annotation, because each section can extend over several sentences, and a sentence can also contain instances of different sections. Additionally, a document can present several instances of each section.

Number of notes	1,038
Number of tokens	360,224
Average note length (in tokens)	347.04 ±235.52
Average number of sections in note	6.94 ±3.36
Average number of unique sections in note	4.38 ±0.99

Table 1: Dataset statistics.

3.1 Dataset

For the task, a subset of the CodiEsp corpus (Miranda-Escalada et al., 2020) was selected by the organizers. CodiEsp is a collection of Spanish unstructured clinical case reports from different medical specialties. This corpus was originally used in a Named Entity Recognition (NER) task (eHealth CLEF 2020), with the aim of identifying procedures and diagnoses labeled with the Spanish version of ICD-10 in a subset of 1,000 documents. An additional collection of 2,751 unannotated documents was also provided as a background set. The present corpus is a randomly-selected subset of the background CodiEsp corpus, consisting of 1038 distinct notes. Table 1 presents some of its relevant


Paciente varón de 21 años de edad, soldado profesional, sin antecedentes médicos o quirúrgicos de interés que acudió a urgencias de nuestro Hospital, remitido por el médico militar, relatando un cuadro molestias abdominales inespecíficas de varios días de evolución. El enfermo se encontraba realizando unas maniobras de supervivencia en la montaña. Desde hacía 18 horas el dolor, inicialmente difuso, se había focalizado en la fosa ilíaca derecha (FID) y aumentado su intensidad. Presentaba anorexia progresiva, sin alteraciones del tránsito gastrointestinal. La exploración mostraba un abdomen blando y depresible, muy doloroso a la presión profunda en FID con defensa y reacción peritoneal acusada. La temperatura axilar era de 38,5º, sin otros hallazgos reseñables. Los estudios radiológicos de tórax y abdomen no reflejaban patología. Analítica: leucocitos, 14.000 x 109 /l (78% neutrófilos); hemoglobina, 14,46 mg /dl; plaquetas, 268.000 x 109 /l. Resto, incluyendo química clínica, función hepática y estudio de coagulación: normal. La ecografía abdominal mostraba una intensa meteorización del intestino delgado que impedía la visualización del ciego y el apéndice vermiforme. No se observaba líquido libre, en cantidad significativa, entre las asas o en el fondo de saco de Douglas.

Se indica una laparotomía exploradora urgente bajo la sospecha de una apendicitis aguda. Se premedica al paciente con amoxicilina y ácido clavulánico 1.000/200 mg intravenosos, respectivamente, unidosis. A través de la incisión de McBurney, se observa una gran tumoración cecal de consistencia dura y múltiples adenopatías. Tras el cierre de la incisión, se aborda de nuevo la cavidad a través de una laparotomía media supra-infraumbilical. La masa, con el aspecto de una perityphilitis, englobaba ciego, apéndice e íleon terminal. Se visualizaban y palpaban ganglios mesocólicos. En el retroperitoneo, en contacto con la gotiera parietocólica derecha, se drenó un absceso de pus cremoso, blanquecino e inodoro que contenía unos 100 ó 150 cc. Se tomaron muestras que fueron cultivadas en tioglicolato, agar sangre y agar chocolate y, posteriormente, analizadas mediante un sistema BACTEC 9240®.

Ante las dudas sobre el origen canceroso de la tumoración, se practicó una hemicolectomía derecha reglada, con intención oncológica, resecando en la pieza quirúrgica el peritoneo parietal afecto y todas las adenopatías palpables. La reconstrucción del tránsito digestivo se llevó a cabo mediante una anastomosis íleo-cólica latero-lateral manual.

En postoperatorio inmediato, el enfermo presentó un pico febril vespertino, con temperatura superior a 38,5º. Se extrajeron hemocultivos que resultaron ser microbiológicamente estériles. En el quinto día del postoperatorio se nos informa del crecimiento en el cultivo de la muestra de líquido biológico de un bacilo gram positivo del género Listeria spp. A través de las técnicas habituales de laboratorio y del serotipado, mediante el método de aglutinación, el patógeno fue identificado, definitivamente, como una Listeria de la especie monocytoges perteneciente al serotipo 4b. En el antibiograma el germen era sensible a la ampicilina, cefotaxima, vancomicina, rifampicina, eritromicina, cotrimoxazol, ciprofloxacina y amoxicilina más ácido clavulánico. Dada la buena evolución clínica del paciente, con desaparición de la fiebre, se mantuvo la pauta antibiótica, consistente en 3.000/600 mg de amoxicilina y ácido clavulánico, respectivamente, repartidos en tres dosis intravenosas diarias. En el sexto día se inicia y progresa la tolerancia a la alimentación por vía oral, con resultados positivos. En el día catorce del postoperatorio, como medida previa al alta, se suspende la antibioterapia intravenosa y se sustituye por una pauta oral de 1.500/375 mg de amoxicilina y ácido clavulánico cada 24 horas, fraccionada en tres dosis, que se mantuvo durante una semana más. El paciente recibe el alta hospitalaria tres días después.

Al año de la intervención el enfermo se encontraba asintomático y desarrollando una actividad normal. Se solicitó una analítica completa, colonoscopia y TAC de abdomen en los que no se informaron hallazgos patológicos o sugerentes de enfermedad inflamatoria intestinal.

Figure 1: Example of annotated document.

statistics.

The training set consists of 500 documents, while the development and test sets consist of 250 documents each. All documents were exhaustively manually annotated by professional clinical annotators with section information. The division of the data set into train, development, and test sets has followed several principles. The proportion of notes in each set is 0.75, 0.125, 0.125 respectively, and the allocation of notes is randomly stratified by category and annotator to ensure a similar proportion of categories in all sets and to account for different annotator expertise levels (see Table 2).

Split	%	Number of notes
Train	75%	781
Dev	12.5%	127
Test	12.5%	130

Table 2: Dataset split.

The documents were annotated by a group of computational linguist experts and doctors from IOMED and the HiTZ Center following an iterative methodology. First, documentation describing the general patterns and labeling strategies for every category was conceived. Subsequently, the computational linguists conducted a few cycles of annotation and discussion of a small set of clinical cases. At the end of every round, the documentation was updated and extended. When the annotation became more fluent, and there was a lower level of ambiguity in deciding among the categories, the first version of the annotation guidelines was released. This guide was used by a group of doctors, trained in clinical report annotation for different tasks, to start a second iterative annotation phase. In this case, a double annotation was performed on the notes. The section evaluation metric was used to decide which notes are ambiguous and sent for revision. Moreover, these complex notes also served to refine the anno-

tation guidelines. The size of the set of annotated documents generated at every iteration increased every time the inter-annotator agreement improved.

3.2 Evaluation Metric

The evaluation of systems dedicated to the detection of text sections is influenced by the type of task, due to the specific requirements for each task. For this task, the usual metrics used for the evaluation of entity classification tasks that check the match of the prediction and the actual annotation at the segment level (exact start and end of the entity) are not useful because the boundaries between sections are usually diffuse. Since the end of one section is always linked to the beginning of another (except for the start and end of the document), this type of annotation would count two sections as wrong if one of these boundaries were not correct, even if the error was a single word.

The metric used by (Goenaga et al., 2021) is not useful in this task either, because, in the documents they dealt with, each section occupied a range of full lines, while in the clinical documents used in the current work, we can find more than one section in the same line, and in some cases, there are documents made up of a single line with more than one or two sections.

In the context of document segmentation, existing works such as (Fournier and Inkpen, 2012; Fournier, 2013) may not be entirely suitable for addressing the specific requirements of this task. To overcome these limitations, we employed a new evaluation metric, B2, which is based on the B metric (Fournier, 2013).[1] This metric utilizes a modified version of the edit distance algorithm (Levenshtein, 1966) and introduces three distinct operations: *additions/deletions*, *substitutions*, and *transpositions*. The first two operations are inherent to the edit distance calculation, while transpositions are a novel operation that enables the movement of section boundaries by a limited number of tokens to align with the reference segmentation. (De la Iglesia et al., 2023) presents a comprehensive understanding of the motivation and precise details of our evaluation metric.

In order to assess the overall performance across the dataset, we compute a weighted average of the metrics obtained for each note. The weights assigned to each note are determined by the number of sections in the corresponding ground truth note, as we identified the number of sections as a meaningful indicator of note complexity.

To facilitate the evaluation process, we provided all participants with an evaluation script that incorporates the tailored parameters of the metric for this task.

4 Submitted Systems and Results

In the following section we provide detailed summaries of the approaches, methodologies, and results presented by each team, shedding light on the innovations and insights gained during the task.

4.1 System Overview

The approaches employed by the participating teams encompassed various strategies, including token classification, text chunking, sentence-based splitting, and the utilization of different annotation schemes. Additionally, some participants incorporated augmentation techniques and hyperparameter optimization. Table 3 showcases the main characteristics of the various submitted systems by each team. All the teams participating in the ClinAIS shared task utilized the pre-trained bsc-bio-ehr-es model, which was introduced by (Carrino et al., 2022). This model, based on the RoBERTa architecture (Liu et al., 2019), has been specifically trained on Spanish biomedical and clinical corpora. By leveraging the knowledge encoded in this pre-trained model, the teams were able to capture the domain-specific nuances and linguistic patterns present in the clinical text, enhancing the performance and accuracy of their systems for section identification.

1. **LSI UNED** (Duque et al., 2023). They develop a system that leverages pre-trained Transformer-based models including a RoBERTa model that has been pre-trained using clinical and biomedical information in the Spanish language, as well as a Longformer architecture trained on biomedical and clinical data. To overcome the token limitation of the RoBERTa model, a sentence-based splitting technique is applied to divide documents into smaller chunks, ensuring a minimum context size of either 128 or 256 tokens. Two distinct annotation

[1] https://ixa2.si.ehu.eus/clinais/evaluation

System		Pre-Trained Model	Section Classification Approach	Data Augmentation	Hyperparameter Optimization	(Long) Document Chunking
Team	Submission #					
ELiRF	1	RoBERTa (ESP/Cli)	Token Classification	None	Grid Search	Chunk documents at the model's max token limit without overlapping.
	2			Back-Translation	Grid Search	
	3			None	Optuna	
	4			Back-Translation	Optuna	
	5			Back-Translation	Optuna	
LSI UNED	1	RoBERTa (ESP/Cli)	Token Classification (IO-style)	None	Training Epochs	Chunk documents that surpass models' limit performing a sentence-based splitting ensuring a minimum of 128 or 256 tokens.
	2	RoBERTa (ESP/Cli)				
	3	Longformer (ESP/Cli)				
	4	Longformer (ESP/Cli)				
	5	Longformer (ESP/Cli)	Token Classification (BIOE-style)			
Grupo Informática UR	1	Longformer (ESP)	Token Classification	None	None	Truncation.
	2	RoBERTa (ESP/Cli + ClinAIS)				
SINAI	1	RoBERTa (ESP/Cli)	Sections' First Token Classification	None	Optuna	Truncation.
	2		Section's First Three Token Classification			
PLNCMM	1	RoBERTa (ESP/Cli)	Sentence Segmentation + Sentence Classification	None	None	Text chunking into sentences.

Table 3: Main characteristics of the systems submitted by each team, providing an overview of the different approaches, methodologies, and key features employed. Models marked with ESP or Cli are pre-trained on Spanish and clinical corpora respectively, and the one marked with ClinAIS has been further pre-trained using the task corpus.

schemes are implemented: Simple-NER, which utilizes a predefined label set to identify section boundaries while assigning an "O" label to the remaining tokens, and Full-NER, which expands the label set to encompass three labels per section, denoting the beginning, inner, and end tokens. The models are trained for varying epochs (10 and 100) to evaluate the impact of overfitting and epoch size on performance. The results demonstrate comparable weighted B2 values across different configurations, with the Longformer model trained on the Simple-NER annotation achieving the highest performance on the test dataset. As avenues for future research, the study recommends conducting in-depth analyses and fine-tuning of hyperparameters, exploring additional pre-trained models, incorporating keywords and keyphrases for improved attention focus, and refining the annotation schemes to facilitate the model's learning process and enhancing overall performance.

2. **PLNCMM** (Carvallo et al., 2023). Their approach consists of a two-step pipeline. In the first step, they employ a text chunking module to partition the input Electronic Clinical Narrative (ECN) into sections using a machine learning-based method. This method utilizes a BIO tag system to define the sections and their boundaries. Each chunk obtained from the text chunking process is then passed through a transformer-based language model. Specifically, they use a RoBERTa architecture that has been fine-tuned using a dedicated methodology for Spanish language and Clinical NLP tasks. This language model, augmented with a feed-forward neural network, performs sentence classification to determine the section to which each chunk belongs. Their analysis highlights the text chunking module as the primary limitation of their system, and they propose two potential approaches to address this limitation in future work.

3. **Grupo Informática de la Universidad de La Rioja** (Heras, 2023). The methodology proposed for addressing the section identification task is grounded in the ULMFIT method, which follows a two-phase approach. In the initial phase, a large pre-trained language model (Carrino et al., 2022) is selected and fine-tuned using domain-specific data, specifically the CodiEsp records annotated for the ClinAIS task. The resultant model is then trained for section identification. Several pre-trained language models were assessed,

and the biomedical Spanish pre-trained models demonstrated superior performance, highlighting the benefits of employing a domain-specific pre-trained language model to enhance the model's effectiveness for the task. Additionally, an augmentation technique involving word masking and prediction using a RoBERTa model was implemented, but it did not yield any noticeable improvements in model performance. Experimental investigations employing various language models and different phases of the ULMFIT two-phase methodology indicate that while the two-phase approach does enhance model performance, the improvement is relatively modest compared to directly training the model solely for section identification. Consequently, the most effective model achieved optimal results by solely focusing on the section identification task and employing a Longformer-based architecture.

4. **SINAI** (Chizhikova et al., 2023). The approach adopted by the SINAI team centers around a token classification framework designed to identify section boundaries. They utilize two system variants specialized in detecting different boundary lengths: one variant concentrates on identifying the first word of a section, while the other variant focuses on the first three words. Their system is constructed using a RoBERTa architecture model, which undergoes pretraining on biomedical and clinical corpora (Carrino et al., 2022) and subsequent fine-tuning through hyperparameter optimization. The results underscore the notable performance enhancement achieved through fine-tuning for longer boundaries.

5. **ELiRF** (Marco et al., 2023). In their participation in the ClinAIS task at IberLEF 2023, this team approached the task as a word sequence classification problem and leveraged the biomedical Spanish language model presented by (Carrino et al., 2022). Their system, during the fine-tuning phase, assigns labels to words to determine their corresponding sections. To enhance the diversity of phrasing and word choice, they employ two data augmentation techniques based on back-translation. The team conducted a hyperparameter search employing both grid search and Optuna strategies, resulting in the presentation of five systems that encompass various combinations of hyperparameter search strategies and the utilization of data augmentation. Their models achieved exceptional results, positioning them at the forefront of this task.

4.2 Team Submission Results

Table 4 provides a summary of the results obtained by the participating teams in the ClinAIS shared task. The table showcases the weighted B2 scores achieved by each team in accurately identifying section boundaries within Spanish ECNs. These results offer interesting information about the effectiveness and efficacy of the different approaches and methodologies employed by the teams.

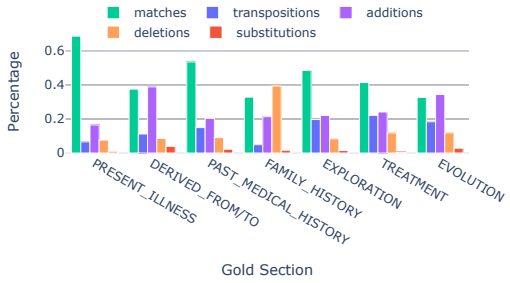
Team	Submission #	Weighted B2
ELiRF	5	80.22
	2	80.08
	1	78.11
	3	77.75
	1	77.26
LSI UNED	4	78.73
	5	77.93
	3	77.54
	1	76.60
	2	75.87
Grupo Informática UR	1	70.36
	2	70.01
SINAI	2	69.86
	1	67.66
PLNCMM	1	69.58

Table 4: Results obtained in the final evaluation on the test set by the submissions of each team.

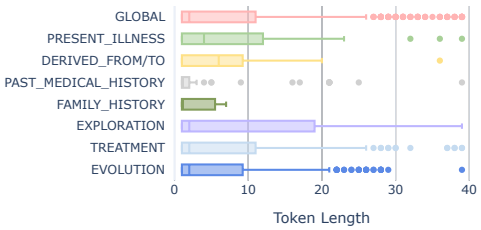
4.3 Analysis

Further analysis reveals valuable insights into the performance of the systems in detecting and classifying different sections within the clinical records.

Figure 2 provides a comprehensive overview of the teams' submissions in terms of matching sections, transpositions, additions, deletions, and substitutions of section



(a) Overview of the operations made per section. Percentages are calculated based on the total number of operations for each section type.



(b) Representation of token movements during transpositions. The *Global* section represents the overall transposition length.

Figure 2: Aggregate error analysis of the best system submissions from each team for the test set, providing a comprehensive overview of combined statistics.

boundaries to match the gold standard; it also presents boxplots showing the token movements during transpositions for each section type.

Upon analyzing Figure 2a, it becomes evident that the *Derived From/To* and *Evolution* sections exhibit a higher proportion of additions compared to other sections. This indicates that the systems struggle the most in identifying these specific sections, resulting in a greater number of missed instances. It highlights the complexity and challenges associated with accurately detecting and classifying these sections in unstructured clinical records as they tend to be directly related to other types of sections. Furthermore, the plot reveals that the *Family History* section has the highest number of deletions. This suggests that it is the section with the highest number of false positives, meaning that systems often mistakenly include unrelated information in this section. The higher error frequency for both the *Family History* and *Derived From/To* sections might be attributed to their limited representation in the dataset.

Upon examining Figure 2b, we observe that, on average, the transpositions do not exceed 10 tokens. This suggests that the majority of section transpositions involve a relatively small number of token movements, indicating a limited need for extensive rearrangements within the text. Interestingly, the *Exploration* section stands out as having the largest transpositions in terms of token movement. This finding aligns with the observation that the *Exploration* section has the longest overall length among the different sections.

Figure 3 provides insights into the substitution errors between predicted and gold-standard sections. Notably, the section pairs that demonstrate the highest degree of mismatch or confusion are *Exploration* and *Evolution*. This can be attributed to the frequent association of patient analysis with monitoring their progress, such as blood analyses or other examinations. Additionally, the *Evolution* section exhibits confusion with the *Treatment*, *Derived From/To*, and *Present Illness* sections. This confusion arises from the similarities and semantic connections between these sections. Furthermore, the *Exploration* section is occasionally replaced with the *Present Illness* section, as preliminary exploratory analyses are sometimes conducted before the commencement of the clinical case.

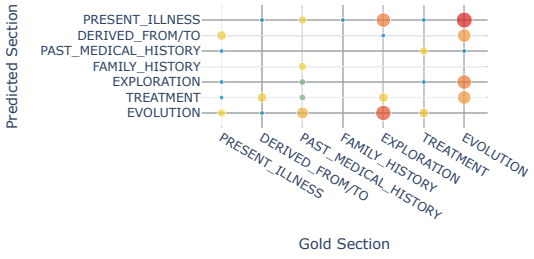


Figure 3: Confusion matrix plot displaying the mismatched section pairs in substitutions. The results of the best system submissions from each team appear combined.

Table 5 showcases the performance of the best submissions from each team, providing insights into the operations performed and

Team	Additions	Deletions	Substitutions	Transpositions	Transposition Length
ELiRF	263	76	21	302	5.7±8.5
LSI UNED	143	128	24	76	16.8±9.9
Grupo Informática UR	327	158	24	216	8.46±10.0
SINAI	188	231	16	164	9.0±11.1
PLNCMM	652	63	18	322	5.0±7.7

Table 5: Comprehensive error operation statistics generated by the best system of each team.

the transposition lengths. The analysis reveals that the most prevalent types of operations across all teams are additions and transpositions. On the other hand, substitutions are observed to be less frequent, suggesting that teams were successful in accurately labeling correctly detected sections in most cases.

5 Conclusions

The ClinAIS shared task at IberLEF 2023 serves as a crucial step toward addressing the challenges posed by the lack of standardized structure in clinical documents. By focusing on the identification of sections within unstructured clinical records in the Spanish language, ClinAIS promotes research and exploration of methodologies that can contribute to the standardization of clinical documents. Furthermore, it paves the way for the development and evaluation of automated systems that can accurately extract and contextualize clinical data. The task made public the first Spanish dataset for section annotation, showcasing various models, characteristics, and techniques employed in the proposed approaches.

In line with the current significance of transformers in NLP tasks, the majority of approaches in the ClinAIS shared task relied on transformer-based models. Interestingly, all the models employed were encoders, and there was a notable absence of generative-based approaches. Despite this commonality, the presented methods exhibited diversity in their classification and annotation techniques.

The analysis of submissions highlights the considerable challenges involved in accurately detecting certain sections, particularly the *Family History* and *Derived From/To* sections, which have a lower occurrence rate compared to other sections. Additionally, distinguishing the *Exploration* and *Evolution* sections from other semantically similar sections has been identified as another significant challenge, indicating potential areas for future research and improvement. Despite these challenges, the submitted results demonstrate strong performance, indicating the effectiveness of the proposed approaches. Moving forward, integrating the section classification task into other downstream tasks presents an exciting avenue for further exploration and advancement in the field of clinical document analysis.

Acknowledgements

This work was partially funded by the Spanish Ministry of Science and Innovation (MCI/AEI/FEDER, UE, DOTT-HEALTH/PAT-MED PID2019-106942RB-C31), the Basque Government (IXA IT1570-22), MCIN/AEI/ 10.13039/501100011033, European Union NextGeneration EU/PRTR (DeepR3 TED2021-130295B-C31, ANTIDOTE PCI2020-120717-2 EU ERA-Net CHIST-ERA), and the Government of the United States IARPA BETTER program (INT NOCORE 19/08 project, via Contract No. 2019- 19051600006).

References

Atutxa, A., A. D. de Ilarraza, K. Gojenola, M. Oronoz, and O. P. de Viñaspre. 2019. Interpretable deep learning to map diagnostic texts to ICD-10 codes. *International Journal of Medical Informatics*, 129:49 – 59.

Carrino, C. P., J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, and M. Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland, May. Association for Computational Linguistics.

Carvallo, A., M. Rojas, C. Muñoz-Castro, C. Aracena, R. Guerra, B. Pizarro, and J. Dunstan. 2023. Automatic Section Classification in Spanish Clinical Narratives Using Chunked Named Entity Recognition. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*, CEUR Workshop Proceedings. CEUR-WS.org.

Chizhikova, M., M. C. Díaz-Galiano, L. A. Ureña-López, and M. T. M. Valdivia. 2023. Automatic Segmentation of Clinical Narratives in Sections with Pre-Trained Clinical Transformer Models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*, CEUR Workshop Proceedings. CEUR-WS.org.

De la Iglesia, I., M. Vivó, P. Chocrón, G. de Maeztu, K. Gojenola, and A. Atutxa. 2023. An Open Source Corpus and Automatic Tool for Section Identification in Spanish Health Records. *Journal of Biomedical Informatics*.

Duque, A., L. Araujo, J. Martinez-Romo, and H. Fabregat. 2023. LSI_UNED at ClinAIS 2023: Transformer Models for Section Identification in Spanish Medical Reports. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*, CEUR Workshop Proceedings. CEUR-WS.org.

Fournier, C. 2013. Evaluating text segmentation using boundary edit distance. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1702–1712, Sofia, Bulgaria, August. Association for Computational Linguistics.

Fournier, C. and D. Inkpen. 2012. Segmentation similarity and agreement. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 3-8, 2012, Montréal, Canada*, pages 152–161. The Association for Computational Linguistics.

Gao, Y., J. Caskey, T. Miller, B. Sharma, M. M. Churpek, D. Dligach, and M. Afshar. 2022a. Tasks 1 and 3 from progress note understanding suite of tasks: Soap note tagging and problem list summarization (version 1.0.0). In *PhysioNet*.

Gao, Y., D. Dligach, T. Miller, S. Tesch, R. Laffin, M. M. Churpek, and M. Afshar. 2022b. Hierarchical annotation for building a suite of clinical natural language processing tasks: Progress note understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5484–5493, Marseille, France, June. European Language Resources Association.

Goenaga, I., X. Lahuerta, A. Atutxa, and K. Gojenola. 2021. A section identification tool: Towards HL7 CDA/CCR standardization in Spanish discharge summaries. *Journal of Biomedical Informatics*, 121:103875.

Heras, J. 2023. Two-stage Fine-Tuning for Automatic Identification of Sections in Clinical Documents. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*, CEUR Workshop Proceedings. CEUR-WS.org.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09.

Levenshtein, V. I. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8).

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Marco, P., M. Jose Castro-Bleda, E. Segarra, and L. F. Hurtado. 2023. ELiRF at ClinAIS Task: Automatic Identification of Sections in Clinical Documents. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*, CEUR Workshop Proceedings. CEUR-WS.org.

Miranda-Escalada, A., A. Gonzalez-Agirre, J. Armengol-Estapé, and M. Krallinger. 2020. 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.

Pomares-Quimbaya, A., M. Kreuzthaler, and S. Schulz. 2019. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. *BMC Medical Research Methodology*, 19.

Terroba, A. R. 2018. Mejora de la calidad del informe clínico de alta hospitalaria desde el punto de vista lingüístico. *PhD Thesis, University of La Rioja.*