LSI_UNED at ClinAIS 2023: Transformer Models for Section Identification in Spanish Medical Reports

Andres Duque[1,2,*], Lourdes Araujo[1,2], Juan Martinez-Romo[1,2] and Hermenegildo Fabregat[1,3]

[1]*NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal 16, Madrid 28040, Spain*

[2]*IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad, Monforte de Lemos 5, Madrid 28019*

[3]*Avature Machine Learning, Spain*

Abstract

This paper presents our participation in the ClinAIS task of the IberLEF 2023 shared evaluation campaign, devoted to the automatic section identification in medical reports written in the Spanish language. Our proposal is based on the use of Transformer-based models to perform a token classification task. In this task, we use a Named Entity Recognition-like annotation of the provided training dataset for jointly detecting the boundaries of each section in the report, and classifying the detected section. Two different pre-trained models are tested and their hyperparameters are explored, and two annotation schemes are tested in our experiments. Our approach achieves the second best results in the task, thus supporting the use of this type of techniques for performing clinical section identification.

Keywords

Clinical section identification, Transformer models, annotation schemes, Spanish language

1. Introduction

Clinical narratives can be found in different types of unstructured medical documents such as Electronic Health Records (EHR) or Electronic Discharge Summaries (EDS). In these documents, doctors and practitioners gather many different aspects of patient information related to a clinical case, such as past and present medical history, diagnoses, treatments or laboratory results [1]. Although healthcare systems usually provide guidelines for writing clinical documents, most of the information contained in these documents can be considered unstructured or poorly structured. However, medical entities and expressions usually offer different information depending on the section they are found within the medical document. Therefore, and given the vast amount of this kind of data within the medical domain, the development of automatic systems performing section identification may help improving a variety of downstream tasks

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

aduque@lsi.uned.es (A. Duque); lurdes@lsi.uned.es (L. Araujo); juaner@lsi.uned.es (J. Martinez-Romo); gildo.fabregat@gmail.com (H. Fabregat)

0000-0002-0619-8615 (A. Duque); 0000-0002-7657-4794 (L. Araujo); 0000-0002-6905-7051 (J. Martinez-Romo); 0000-0001-9820-2150 (H. Fabregat)

in the field such as Named Entity Recognition (NER) [2], temporal relation extraction [3] or ICD-10 coding [4].

In this paper we present our participation in the ClinAIS task of the IberLEF 2023 shared evaluation campaign [5], to be held as part of the XXXIX conference of the Spanish Society for Natural Language Processing (SEPLN 2023). Systems participating in the ClinAIS task must automatically identify different medical sections in unstructured clinical documents written in the Spanish language. The proposed system employs Transformer-based architectures using different context windows for identifying those sections, through a Named Entity Recognition (NER)-like methodology with different annotation schemes. For this purpose, we consider tokenized texts and define the different labels or NER tags indicating the starting token of each particular section. Then, we train the different selected models for recognizing those labels and hence the starting point of each identified section.

The rest of the paper is structured as follows: Section 2 is devoted to exploring existing systems performing automatic medical section identification. The particular task addressed in this work, together with the available dataset and evaluation metrics, are described in Section 3. The developed system is described in Section 4 and the obtained results in Section 5. Finally, Section 6 offers some conclusions regarding this research and possible lines of work to be followed in the future.

2. Related Work

The automatic identification of sections in clinical documents has been normally addressed through three main different techniques: rule-based systems, machine learning systems and a combination of both [6]. Among the rule-based methods, some previous works make use of exact matching techniques for determining the boundaries (beginning and end of a section) [7, 8], although regular expressions allow the systems to generalize the detection of indicators of these boundaries [4, 9]. Also, probabilistic methods can be found in the literature complementing these previous approaches by taking advantage of clinical corpora from which specific probabilistic knowledge can be modeled and then applied to new medical reports for detecting the desired sections [10, 11].

Regarding machine learning methods, classical methods such as logistic regression can be found performing section identification at sentence level [12], while Conditional Random Fields (CRFs) and Support Vector Machines (SVMs) usually show good performance in this kind of tasks [13, 14]. However, in the last few years, the rise of deep learning models has produced important advances in the field. Section detection in medical dictations is addressed in [15] through the use of LSTMs [16] as a binary classification task at token level, this is, classifying each token as positive if it indicates the beginning of a section and negative otherwise. Recurrent Neural Networks (RNNs) and BERT [17] are used in [18] for classifying sections in Electronic Health Records (EHRs), including transfer learning techniques in order to overcome the lack of training data.

3. The ClinAIS Task

The ClinAIS task is part of the shared evaluation campaign IberLEF 2023 and aims to develop systems for the automatic identification of sections within medical reports written in Spanish. The following seven possible sections are defined within the reports, which may or may not occur and may appear more than once in the same clinical report: “Present Illness”, “Derived from/to”, “Past Medical History”, “Family History”, “Exploration”, “Treatment” and “Evolution”.

Therefore, the main objective of a system participating in the task is to detect the boundaries of the sections (i.e., the start and end tokens) as well as to classify the correct section. Additional information about the task can be found in [19].

3.1. Dataset

The dataset provided by the organizers is a subset of the CodiEsp dataset, originally used in the CLEF eHealth 2020 task [20]. Specifically, the corpus used for the ClinAIS task is a revised version of the dataset presented in [21] and contains 1,038 clinical reports and is divided into training (781 reports), development (127 reports), and test (130 reports) splits. Each of these datasets consists of a JSON file that stores a set of clinical reports, both in raw text format and tokenized text format, so participants do not need to perform text tokenization. Therefore, each of the tokens within a clinical report may or may not be the beginning of one of the sections to be identified. The average length of the reports is 374.04 tokens, while each report has an average of 6.94 sections (4.38 unique sections).

3.2. Evaluation Metrics

The metric proposed in the task to evaluate participating systems is called B2 and takes into account operations such as additions, deletions, substitutions and n-wise transpositions for evaluating the differences between the detected sections and the Gold Standard sections. Additions and deletions can be considered complete errors, as they represent either detected sections that do not exist in the Gold Standard or undetected sections that are present in the Gold Standard. Substitutions occur when the boundaries of a section are correctly detected, but the section type is classified incorrectly. Finally, n-wise transpositions refer to situations in which a section is correctly detected, but the predicted boundaries are displaced n words with respect to the Gold Standard.

With all this information, different weights are assigned to each operation for designing the final formula. This formula assigns a score to the system on each of the considered medical reports. The final evaluation of the system is computed by performing a weighted average of all the reports in the test dataset, using the number of sections in the Gold Standard of each report as a means of assigning weights. More information about the specifics of the evaluation metric can be found in [19].

4. Proposed System

The system developed in this research mainly focuses on the use of pre-trained models based on the Transformer architecture [22]. The addressed task is modeled as a Named Entity Recognition task, hence the main objective of the developed models will be to correctly identify the label of each token within a medical report. This way, we will be able to subsequently determine the start and end token of each section, this is, the section's boundaries. For this purpose, a token classification layer is added on top of the pre-trained models in order to fine-tune it with the available training dataset.

4.1. Transformer Models

Two different Transformer-based pre-trained models are used in this research: the first one is a RoBERTa [23] model pre-trained with clinical and biomedical information from different sources, written in the Spanish language [24]. However, the main limitation of this model is that the maximum size of the documents that can be processed is 512 tokens, while the task dataset presents longer documents. In order to get around this limitation, we divide the documents into smaller chunks by performing a sentence-based splitting while setting a minimum context size of 128 or 256 tokens. This way we ensure that these new instances to be processed by the model do not exceed the 512-token limitation, while maintaining a minimum size that avoids losing important information for performing token classification.

The second Transformer-based model used in this work is a Longformer architecture [25] also pre-trained with biomedical and clinical information written in the Spanish language [26]. The main advantage of this model is the extension of the maximum document size to 4096 tokens, which in this case allows us to process all the documents in the task dataset without the need of splitting them into smaller instances.

Models have been trained during 10 and 100 epochs in order to determine whether overfitting issues can affect the final performance of the trained models, as well as the impact of this epoch size on the final results. A batch size of 4 has been used for all the experiments, with a learning rate of 5×10^{-5}. The AdamW optimizer with a weight decay of 10^{-5} has been also used in all the proposed experiments. Finally, categorical cross-entropy is used as the loss function for fine-tuning the models.

4.2. Annotation Schemes

We propose two different annotation schemes for addressing this task as a Named Entity Recognition task, this is, a token classification task. The first scheme, denoted "Simple-NER", defines a particular label for a token indicating the beginning of a new section within the document, and a label "O" for the rest of the tokens in the document. Hence a total of 8 different labels can be assigned to each token: "O", "PRESENT_ILLNESS", "DERIVED_FROM/TO", "PAST_MEDICAL_HISTORY", "FAMILY_HISTORY", "EXPLORATION", "TREATMENT" and"EVOLUTION". Then, each test instance will be post-processed in order to determine the boundaries of each section. A section will start each time a token classified with a label different from "O" is found. The following "O" tokens will be considered to be within the found section, until a new label different from "O" is found, except if this new label indicates the same section that was being considered.

The second annotation scheme, denoted "Full-NER", performs a slightly more sophisticated labeling of the tokens within the training instances. In this case, we differentiate the starting, inner and ending tokens of each section. Hence, for each section "S" we define labels "B-S" (starting token of section "S"), "I-S" (inner token of section "S") and "L-S" (last token of section "S"). A total of 21 different labels are then defined for this scheme. Some additional post-processing rules are then added in this case for performing the final annotation of each test instance: we consider that a new section begins whenever a label belonging to a different section is found, regardless of whether it is a starting, inner or last label. Also, we consider that a new section must contain at least three tokens, hence if a section with only one or two tokens is found, it is not classified, and the tokens are considered to belong to the previous section. These post-processing rules have been designed after manual observation of the development dataset.

5. Results

In this section we will discuss the main results obtained by the proposed system on the test dataset of the ClinAIS task, as well as a comparative among the different systems participating in the task.

Up to 5 runs could be submitted for evaluation in the task. Table 1 shows the characteristics of the different runs submitted by our team: Transformer-based model employed in the experiment, context size, number of epochs and annotation scheme. The context size depends on the selected model: since the RoBERTa model admits a maximum size of 512 tokens, smaller documents were created from the original reports by performing sentence-based splitting, yet assuring a minimum size of 128 or 256 tokens. On the other hand, since Longformer models admit up to 4096 tokens per instance, no document splitting was performed in those cases.

Table 1
Configurations of the different runs submitted to the competition. Model used, context size allowed by the model, total number of epochs and type of annotation scheme are shown.

Run	Model	Context size	No. epochs	Annotation Scheme
Run 1	RoBERTa	128 (min)	100	Simple-NER
Run 2	RoBERTa	256 (min)	100	Simple-NER
Run 3	Longformer	4096 (max)	10	Simple-NER
Run 4	Longformer	4096 (max)	100	Simple-NER
Run 5	Longformer	4096 (max)	100	Full-NER

Table 2 shows results obtained by the different configurations of the proposed model for the development and test datasets, evaluated using the weighted B2 metric.

As we can observe in the table, all results fall within a similar range of weighted B2 values, which indicates that the differences between the considered configurations of the model are not especially relevant. In particular, although the best performing model for the development

Table 2
Results obtained by the proposed system for the development and test datasets. Bold indicates the best result in each case (Metric: weighted B2).

Run	Weighted B2 (Dev)	Weighted B2 (Test)
Run 1	0.7677	0.7660
Run 2	**0.7820**	0.7587
Run 3	0.7723	0.7754
Run 4	0.7799	**0.7873**
Run 5	0.7763	0.7793

dataset is the RoBERTa model with a context size of 256 tokens, this result is not maintained when it comes to the test dataset. The remaining configurations show consistent results for both the development and the test datasets, being run 4 (Transformer model trained during 100 epochs, "Simple-NER" annotation scheme) the best performing configuration for the test dataset. Finally, results for run 5 indicate that despite being more complex and hence more difficult for the model to learn, the "Full-NER" annotation scheme also allows the system to obtain competitive results.

Finally, table 3 shows a comparison of the best performing runs of every team participating in the ClinAIS competition, evaluated with the weighted B2 metric on the test dataset.

Table 3
Comparative of results obtained for the test dataset by the different teams participating in the competition.

Team	Weighted B2 (Test)
ELiRF	**0.8022 (1)**
LSI_UNED (Ours)	0.7873 (2)
joheras	0.7036 (3)
SINAI	0.6986 (4)
plncmm	0.6958 (5)

The table shows how our system is able to obtain the second best result in the competition, only around 1.5% below the best performing system, and significantly above the third best system. This indicates that the use of Transformer-based models is an interesting starting point for performing automatic section identification in medical reports, due to the current ease of use and adaptation to almost any problem of this type of systems.

6. Conclusions and Future Work

This paper describes our participation in the ClinAIS task of the IberLEF 2023 shared evaluation campaign, devoted to the automatic identification of sections within medical reports. The main contribution of this research is the use of two different Transformer-based models for performing a NER-like token classification, in which the main objective is to detect those tokens indicating the beginning of a new section within the report. Two different annotation schemes similar to those employed in NER tasks are also studied. The main results indicate that these models are useful for accurate detection of the different sections and their respective boundaries. The use of Longformer models, which allow for documents up to 4096 tokens in length, also avoids the need of particular pre-processing techniques such as sentence splitting. The best performing configuration of our system is able to achieve the second best result in the competition.

As future lines of work, we consider a further analysis and tuning of the hyperparameters of the deep learning models employed in this work, such as learning rates or loss functions, as well as the study of additional pre-trained models in order to determine their impact in the final results. We also envisage the inclusion of additional information such as keywords and keyphrases in the training loop in order to direct the model's attention to those expressions able to more accurately define the beginning or end of a section. Finally, the annotation schemes should be also studied and refined as much as possible in order to make it easier for the model to learn the labels associated to the boundaries of the sections.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the DOTT-HEALTH Project (MCI/AEI/FEDER, UE) under Grant PID2019-106942RB-C32 and OBSER-MENH Project (MCIN/AEI/10.13039/501100011033 and NextGenerationEU/PRTR) under Grant TED2021-130398B-C21 as well as project RAICES (IMIENS 2022).

References

[1] I. Goenaga, X. Lahuerta, A. Atutxa, K. Gojenola, A section identification tool: Towards HL7 CDA/CCR standardization in spanish discharge summaries, J. Biomed. Informatics 121 (2021) 103875. URL: https://doi.org/10.1016/j.jbi.2021.103875. doi:10.1016/j.jbi.2021.103875.

[2] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, H. Xu, A comprehensive study of named entity recognition in chinese clinical text, Journal of the American Medical Informatics Association 21 (2014) 808–814.

[3] H. Lee, Y. Zhang, M. Jiang, J. Xu, C. Tao, H. Xu, Identifying direct temporal relations between time and events from clinical notes, BMC Medical Informatics Decis. Mak. 18 (2018) 23–34. URL: https://doi.org/10.1186/s12911-018-0627-5. doi:10.1186/s12911-018-0627-5.

[4] A. Duque, H. Fabregat, L. Araujo, J. Martínez-Romo, A keyphrase-based approach for interpretable ICD-10 code classification of spanish medical reports, Artif. Intell. Medicine 121 (2021) 102177. URL: https://doi.org/10.1016/j.artmed.2021.102177. doi:10.1016/j.artmed.2021.102177.

[5] 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.

[6] A. Pomares-Quimbaya, M. Kreuzthaler, S. Schulz, Current approaches to identify sections within clinical narratives from electronic health records: a systematic review, BMC medical research methodology 19 (2019) 1–20.

[7] X. Wang, H. Chase, M. Markatou, G. Hripcsak, C. Friedman, Selecting information in electronic health records for knowledge acquisition, Journal of biomedical informatics 43 (2010) 595–601.

[8] N. D. Phuong, V. T. N. Chau, Automatic de-identification of medical records with a multilevel hybrid semi-supervised learning approach, in: T. Cao, Y. Ho (Eds.), 2016 IEEE RIVF International Conference on Computing & Communication Technologies, Research, Innovation, and Vision for the Future, RIVF 2016, Hanoi, Vietnam, November 7-9, 2016, IEEE, 2016, pp. 43–48. URL: https://doi.org/10.1109/RIVF.2016.7800267. doi:10.1109/RIVF.2016.7800267.

[9] M. J. Schuemie, D. Trieschnigg, E. Meij, Dutchhattrick: Semantic query modeling, context, section detection, and match score maximization, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, USA, November 15-18, 2011, volume 500-296 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2011. URL: http://trec.nist.gov/pubs/trec20/papers/DutchHatTrick.med.update.pdf.

[10] J. C. Denny, R. A. Miller, K. B. Johnson, A. Spickard III, Development and evaluation of a clinical note section header terminology, in: AMIA annual symposium proceedings, volume 2008, American Medical Informatics Association, 2008, p. 156.

[11] C. P. Shivade, P. Malewadkar, E. Fosler-Lussier, A. M. Lai, Comparison of UMLS terminologies to identify risk of heart disease using clinical notes, J. Biomed. Informatics 58 (2015) S103–S110. URL: https://doi.org/10.1016/j.jbi.2015.08.025. doi:10.1016/j.jbi.2015.08.025.

[12] M. Tepper, D. Capurro, F. Xia, L. Vanderwende, M. Yetisgen-Yildiz, Statistical section segmentation in free-text clinical records, in: N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012, European Language Resources Association (ELRA), 2012, pp. 2001–2008. URL: http://www.lrec-conf.org/proceedings/lrec2012/summaries/1016.html.

[13] J. Jancsary, J. Matiasek, H. Trost, Revealing the structure of medical dictations with conditional random fields, in: 2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2008, pp. 1–10. URL: https://aclanthology.org/D08-1001/.

[14] D. L. Mowery, J. Wiebe, S. Visweswaran, H. Harkema, W. W. Chapman, Building an automated SOAP classifier for emergency department reports, J. Biomed. Informatics 45 (2012) 71–81. URL: https://doi.org/10.1016/j.jbi.2011.08.020. doi:10.1016/j.jbi.2011.08.020.

[15] N. Sadoughi, G. P. Finley, E. Edwards, A. Robinson, M. Korenevsky, M. Brenndoerfer, N. Axtmann, M. Miller, D. Suendermann-Oeft, Detecting section boundaries in medical

dictations: Toward real-time conversion of medical dictations to clinical reports, in: A. Karpov, O. Jokisch, R. Potapova (Eds.), Speech and Computer - 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18-22, 2018, Proceedings, volume 11096 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 563–573. URL: https://doi.org/10.1007/978-3-319-99579-3_58. doi:10.1007/978-3-319-99579-3_58.

[16] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735. doi:10.1162/neco.1997.9.8.1735.

[17] 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.

[18] S. Rosenthal, K. Barker, Z. Liang, Leveraging medical literature for section prediction in electronic health records, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 4864–4873.

[19] 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).

[20] 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. Névéol (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.

[21] 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).

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[24] 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, CoRR abs/2109.03570 (2021). URL: https://arxiv.org/abs/2109.03570. arXiv:2109.03570.

[25] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, CoRR abs/2004.05150 (2020). URL: https://arxiv.org/abs/2004.05150. arXiv:2004.05150.

[26] 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.