Automatic Segmentation of Clinical Narratives in Sections with Pre-Trained Clinical Transformer Models

Mariia Chizhikova[1,*,†], Manuel Carlos Díaz-Galiano[1], Luis Alfonso Ureña-López[1] and María Teresa Martín Valdivia[1]

[1]Department of Computer Science, University of Jaén, Campus Las Lagunillas, s/n, Jaén, 23071, Spain

Abstract

This paper presents the participation of the SINAI team in the ClinAIS shared task at IberLEF 2023, focusing on the task of section identification in clinical reports. The approach involves a multiclass token classification framework for section boundary detection, utilizing two system variants tuned for detecting one-token and three-token long boundaries. The system is built upon a RoBERTa architecture model pre-trained on biomedical and clinical corpora, and fine-tuned for the token classification task through hyperparameter optimization trials.The results show that fine-tuning for longer boundaries improved performance (0.6766 vs 0.6986 weighted B2 score). Error analysis revealed challenges in detecting "DERIVED_FROM/TO" and "EVOLUTION" sections due to class imbalance and semantic confusion.

Keywords

Clnical Natural Language Processing, Section Identification, RoBERTa language model,

1. Introduction

The widespread implementation of health information systems resulted in the generation of an extensive number of Electronic Health Records (EHR) on a global scale, amounting to billions of records. At large scale, EHRs constitute an invaluable source of medical and clinical information that can be leveraged to impulse both medical research and the quality of clinical assessment.

Regrettably, the secondary utilization of Electronic Health Record (EHR) content presents inherent challenges due to its hybrid structure, which encompasses a diverse range of data types, including coded or structured information, multimedia content and free-text narratives. Structuring clinical information has become increasingly critical as healthcare data needs have evolved. Different strategies from the field of Clinical Information Extraction were proposed to address the need of structured clinical information extraction from free-text reports [1]. These approaches apply Natural Language Processing (NLP) techniques that implement rule-based, machine learning (ML), or deep learning (DL) techniques in order to extract clinically relevant

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

mchizhik@ujaen.es (M. Chizhikova); mcdiaz@ujaen.es (M. C. Díaz-Galiano); laurena@ujaen.es (L. A. Ureña-López); maite@ujaen.es (M. T. M. Valdivia)

0000-0002-0302-912X (M. Chizhikova); 0000-0001-9298-1376 (M. C. Díaz-Galiano); 0000-0001-7540-4059 (L. A. Ureña-López); 0000-0002-2874-0401 (M. T. M. Valdivia)

entities and map them into codes from a controlled vocabulary like SNOMED-CT [2] or classify texts into categories [3] among others methods.

The main challenges of clinical NLP involve dealing with the overall linguistic complexity of an EHR: negated expressions, co-references, misspellings, abbreviations, anaphoric relations and so on [4]. The task of Section Identification (SI) consists in detecting boundaries of text sections and adding semantic annotations to the divided text. This task can be considered as a step towards addressing the aforementioned problems by identification and disambiguation of the narrative structure underlying to each clinical report.

The realization of SI was proven to be beneficial for the performance in clinical information extraction tasks such as entity recognition in Chinese clinical reports [5] and temporal relation extraction [6]. Many works on this specific task focused on clinical narratives written in English: Li et al. [7] implemented a Hidden Markov Model based algorithm trained on a corpus of 9,679 clinical notes from New York-Presbyterian Hospital. More recently, Sadoughi et al. [8] proposed a neural network based approach to SI that trains a model with a Long Short-Term Memory (LSTM) layer in its core. This work also proves that SI is beneficial for the quality of the post-processing of automatically transcribed clinical reports.

However, the capability to analyze clinical text in languages other than English holds significant potential for accessing crucial medical data pertaining to patient cohorts treated in countries where English is not the official language[9]. When it comes to processing clinical narratives written in Spanish, SI appears to be a not so widely explored task, despite the fact of Spanish being the fourth most spoken language in the world. Goenaga et al. [10] evaluate three different approaches to automatically standardizing Spanish electronic discharge summaries from two different hospitals following the HL7 Clinical Document Architecture and state that transfer-learning approaches show the best performance compared to other ML and rule-based techniques.

ClinAIS shared task at the Iberian Languages Evaluation Forum (IberLEF) 2023 aims to impulse the research in the automatic SI applied to the Spanish language by providing a dataset of 1,038 clinical reports annotated with seven predefined medical sections: Present Illness, Derived from/to, Past Medical History, Family history, Exploration, Treatment and Evolution [11, 12].

The main objective of this paper is to describe the system presented by the SINAI team at the ClinAIS shared task. Our approach follows the line of the best performed system described in Goenaga et al. [10]. We tackle the Section Identification task as a token classification problem that intends to detect the borders of each of the seven sections.

The remainder of the paper is organized as follows: Section 2 provides a detailed description of the datasets that were made available by the organizers of the competition, Section 3 is dedicated to the system we developed to tackle the task of automatic SI, Section 3.1 outlines the specific procedures and configurations used during the experiments, Section 4 discloses the results obtained by the presented system during the official evaluation that are subsequently analyzed in Section 5 which critically examines the overall system performance and investigates potential sources of errors. Finally, Section 6 summarizes the main aspects of our contribution and offers insights to enhance future iterations of the automatic section identification system.

2. Data

ClinAIS corpus [11] is a collection of 1,038 clinical case reports from different medical specialties that were obtained by randomly sampling the CodiEsp corpus [13].

The data was divided into three subsets in a stratified manner so that the section distribution is similar in all three sets: training, development and test. The proportion of notes in each is 0.75, 0.125 and 0.125 respectively. It is worth mentioning that the evaluation set was released alongside 2,751 unannotated documents in order to prevent the participants from performing manual corrections and/or annotations.

Each corpus entry was annotated with borders of sections from the following list: Present Illness, Derived from/to, Past Medical History, Family history, Exploration, Treatment and Evolution. Figure 1 shows the class label distribution across training and development datasets. The sections Derived from/to and Family are scarcely represented compared to other labels.

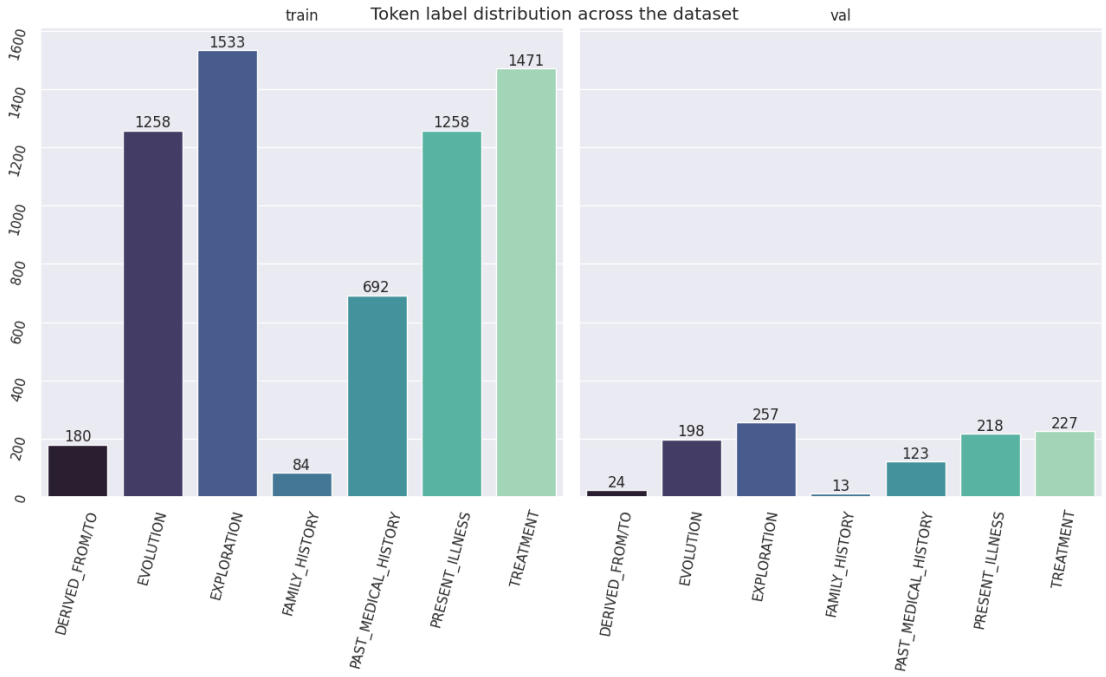

Figure 1: Label distribution across the ClinAIS training and validation subsets

As it can be inferred from the Table 1 which summarizes some relevant statistics of the data subsets, the average report length (measured in tokens produced by RoBERTa tokenizer) doesn't exceed 512, the maximum input length of the majority of pre-trained transformer models available for Spanish, in all of the subsets. We can also point out that the data is quite homogeneous across subsets in terms of the average of the annotated sections.

Notably, section borders in many cases do not match with sentence starts which makes it unviable to rely on a sentence classification approach.

	Train	Dev	Test + Background
Number of reports	781	127	130 + 2,751
Avg. report length (STD)	449.68 (311.12)	436.75 (277.71)	500.82 (392.93)
Avg. sections (STD)	8.29 (4.24)	8.35 (3.16)	–
Avg. unique sections (STD)	4.72 (0.97)	4.88 (0.83)	–

Table 1
Corpus statistics

3. System Description

Fine-tuning large pre-trained transformer models was proven to be an effective approximation to many NLP tasks like text classification [14] or NER [15]. Our approach follows this methodology because we tackle the task of automatic SI as a multiclass token classification problem by fine-tuning a large language model to detect section boundaries.

In order to convert the corpus into a format that would be coherent with our problem formulation, we performed pre-tokenization by splitting the corpus by white space and treating the punctuation marks as separate units. As for the token labeling, we compared the approach of marking only the boundary word as an entity to the method of selecting the first three words of each section and labeling these according to the BIO (Beginning, Inside and Outside) scheme.

We opted for building our system with a RoBERTa architecture model [16] pre-trained on a combination of biomedical and clinical corpora [17]. This model was fine-tuned for the token classification task by adding a dropout and a linear layer on top of the original architecture.

3.1. Experimental Setup

We conducted two experiments in order to compare the aforementioned labeling strategies. In order to maximize the resulting performance of the systems, we carried out optimization of hyperparameters for model fine-tuning. This process relied on the Optuna framework which provides efficient trial pruning and parameter sampling strategies [18].

For each system, we performed 5 trials on a single NVIDIA Ampere A100 GPU. The hyperparameter search space was defined as follows:

- Learning rate: a float value between $3e-5$ and $5e-5$
- Per device train batch size: either 8 or 16
- Weight decay: a float value between $1e-12$ and $1e-1$
- Adam epsilon: a float value between $1e-10$ and $1e-6$
- Warmup steps: an integer value between 0 and 1000

The number of training epochs was selected by making use of the Early Stopping strategy that interrupts fine-tuning when the reference metric doesn't improve during 3 epochs. Table 2 presents the hyperparameters selected for each of the experiments.

	Single word boundary	Three word boundary
Learning rate	4e-5	4.1e-5
Training epochs	18	18
Batch size	16	16
Weight decay	5.9e-7	1.4e-6
AdamW epsilon	8.9e-10	3.8e-8
Warmup steps	485	596

Table 2
Hyperparameters selected for each experiment.

System	Weighted B2
One-token boundary detection	0.6766
Three-token boundary detection	**0.6986**

Table 3
Official evaluation results

4. Results

Our team submitted a total of two runs: one per each of the described pre-processing methodologies. Identification of sections within unstructured clinical notes poses numerous challenges, thereby impeding the comprehensive assessment of this task. Prominent among these challenges is the inherent interconnection between the end of one section and the start of another. Another issue is related to the fact that sections are not delimited by paragraphs, lines or phrases. To address this the organizers designed the B2 evaluation metric, which is an adaptation of the boundary distance [19, 20]. Table 3 displays the results obtained during the official evaluation of the two presented system variants.

5. Performance Analysis

This section attempts to provide insights regarding the overall system performance and investigate the possible reasons for its errors. As shown in Table 3, the variant that involved recognition of the first three tokens of each section was proven to be the best option scoring 0.6986 weighed B2 during the official evaluation on the test set. This performance improvement can be explained by the fact that the majority of the words at the beginning of each section are adpositions (e.g. *con (with)*, *en (in)*), determiners (e.g. *el/la (the)*) and pronouns (e.g. *su (his/her)*). This type of words lack of semantic information almost completely which has a negative impact on the amount of information that can be captured by an embedding of a contextual model such as RoBERTa clinical. Therefore, the classifier layer cannot learn and generalize correctly distinctive features of such tokens which leads to classification errors. Figure 2 displays the frequency of apparition of different parts-of-speech at the beginning of the sections.

Moreover, we carried out an error analysis using the predictions made with the best performing variant of our system on the development subset and evaluated with the official evaluation

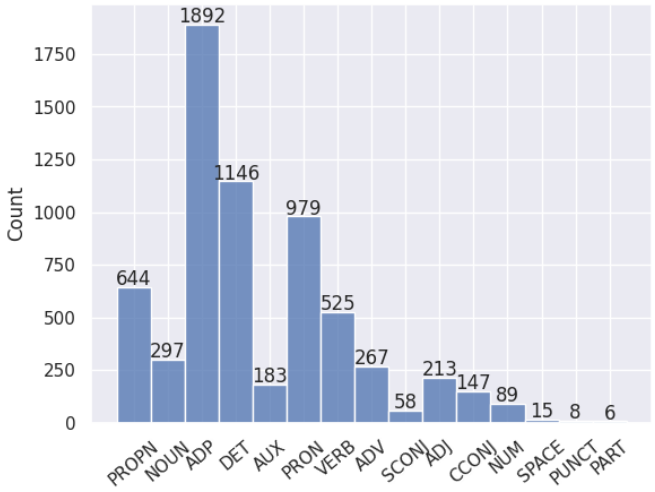

Figure 2: Part-of-speech of tokens at the beginning of sections in the training subset

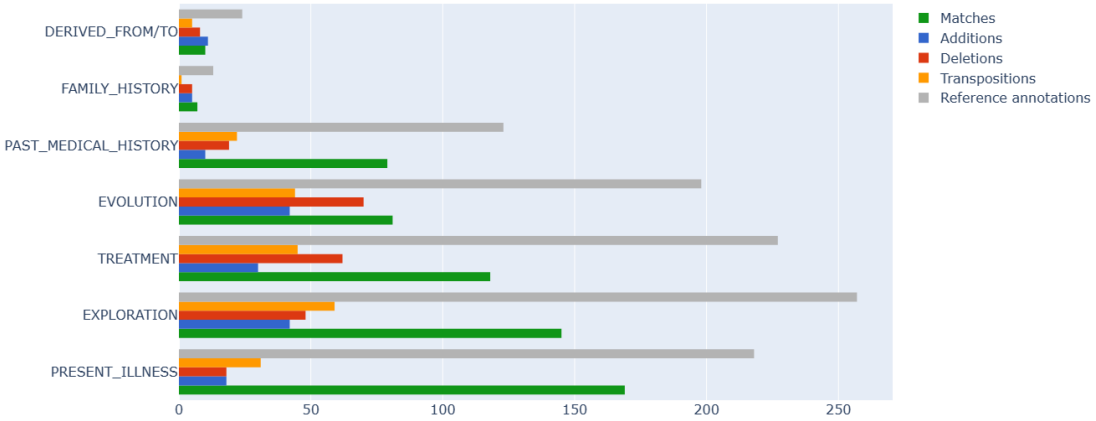

Figure 3: Error statistics per each label. Substitutions are not displayed because technically a substitution is a deletion of one label and addition of the other and there is no straightforward way to include it in the graph

script scoring 0.7477 weighted B2. We distinguish between four types of errors: additions (when the system identifies a section that is not present in the report), deletions (when the system is not capable to detect a section), transpositions (the section was identified, but the extracted span is transposed) and substitutions (the identified section is assigned a wrong label). Figure 3 summarizes the error type and rate per each section type.

The highest number of matches with respect to a total number of reference annotations (169 out of 218) is shown for the label "PRESENT_ILLNESS" that is the most commonly seen across the training dataset. Nevertheless, there is no clear direct relation between the most represented labels in the dataset and the quality of predictions: the categories with the worst matches/reference annotations ratio are "DERIVED_FROM/TO" - 10 matches out of 24 reference annotations

and "EVOLUTION" - 81 matches out of 198 references. Notably, the "DERIVED_FROM/TO" class is the only one where the number of additions (11) is higher than the matches. As for the identification of the "EVOLUTION" sections, the frequency of deletions (70) is worth mentioning.

As for the substitutions, 14 errors of this type were detected when evaluating our best performing system on the development dataset. The most confounded label resulted to be "EVOLUTION" which was substituted with "EXPLORATION" 4 times and once with "TREATMENT". Moreover, it substituted the label "TREATMENT" twice and "PRESENT_ILLNESS" once. The "EVOLUTION" was defined in the annotation guidelines as the "*Evolution of the patient's health status. It may include differential diagnoses*", which means it can include references to the explorations performed and its results, as well as the response to treatments and updates on the present illness, so semantically this section might be close to many other sections depending on the report and section position within it.

6. Conclusions and Future Work

This paper covers the participation of the SINAI team at the ClinAIS shared task held on IberLEF 2023. We describe our approach to the task of section identification in clinical reports. Our problem formulation follows a multiclass token classification approach for section boundary detection and we compare two variants of systems: one tuned to detect one-token boundary and the other tuned to recognize three-token-long boundaries. We based our system on a RoBERTa architecture model pre-trained on a combination of biomedical and clinical corpora that was fine-tuned for the token classification task with hyperparameters selected during a 5 trial optimization.

The approach of fine-tuning a model to detect longer section boundaries performed better scoring 0.6986 weighted B2 score. This improvement is probably due to the fact that the majority of sections start with a function word like a preposition or a determiner.

In order to shed light on the errors made by our system we conducted an error analysis of the predictions of our best performing system on the development set. The two sections that resulted to be the hardest to detect have proven to be "DERIVED_FROM/TO", one of the less represented classes in the training set, and "EVOLUTION" which might be semantically close to other sections like "TREATMENT", "PRESENT_ILLNESS" or "EXPLORATION" and is thus frequently confounded or omitted.

With the purpose of dealing with these issues, an implementation of a weighted loss function can be proposed as a way of mitigating the impact of class imbalance. With the same aim data augmentation techniques like back translation or generative language models can be used. A more in-depth dataset analysis can provide more detailed information regarding the linguistic characteristics of each section which can be employed as an extension to the contextual embedding during the classification process.

Acknowledgments

This work has been partially supported by WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, and projects CONSENSO (PID2021-122263OB-

C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government, and project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government.

References

[1] 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.

[2] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings, 2022.

[3] M. Chizhikova, P. López-Úbeda, J. Collado-Montañez, T. Martín-Noguerol, M. C. Díaz-Galiano, A. Luna, L. A. Ureña-López, M. T. Martín-Valdivia, Cares: A corpus for classification of spanish radiological reports, Computers in Biology and Medicine 154 (2023) 106581.

[4] 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.

[5] 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.

[6] S. Kropf, P. Krücken, W. Mueller, K. Denecke, Structuring legacy pathology reports by openehr archetypes to enable semantic querying, Methods of information in medicine 56 (2017) 230–237.

[7] Y. Li, S. Lipsky Gorman, N. Elhadad, Section classification in clinical notes using supervised hidden markov model, in: Proceedings of the 1st ACM International Health Informatics Symposium, IHI '10, Association for Computing Machinery, New York, NY, USA, 2010, p. 744–750. URL: https://doi.org/10.1145/1882992.1883105. doi:10.1145/1882992.1883105.

[8] 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: Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20, Springer, 2018, pp. 563–573.

[9] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical natural language processing in languages other than english: opportunities and challenges, Journal of biomedical semantics 9 (2018) 1–13.

[10] I. Goenaga, X. Lahuerta, A. Atutxa, K. Gojenola, A section identification tool: towards hl7 cda/ccr standardization in spanish discharge summaries, Journal of Biomedical Informatics 121 (2021) 103875.

[11] 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).

[12] 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.

[13] 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.

[14] P. López-Ubeda, M. C. Díaz-Galiano, L. A. U. López, M. T. Martín-Valdivia, T. Martín-Noguerol, A. Luna, Transfer learning applied to text classification in spanish radiological reports, in: Proceedings of the LREC 2020 Workshop on Multilingual Biomedical Text Processing (MultilingualBIO 2020), 2020, pp. 29–32.

[15] A. Miranda-Escalada, E. Farré, M. Krallinger, Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results., IberLEF@ SEPLN (2020) 303–323.

[16] 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, arXiv preprint arXiv:1907.11692 (2019).

[17] 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, arXiv preprint arXiv:2109.03570 (2021).

[18] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.

[19] 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), Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 1702–1712. URL: https://aclanthology.org/P13-1167.

[20] 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).