LinkMed: Entity Recognition and Relation Extraction from Clinical Notes in Spanish

Carlos Muñoz-Castro[1,3,6], Andrés Carvallo[1], Matías Rojas[4], Claudio Aracena[2,6], Rodrigo Guerra[2,5], Benjamín Pizarro[3] and Jocelyn Dunstan[3,4,5,6]

[1]*National Center for Artificial Intelligence, Chile*
[2]*Faculty of Physical and Mathematical Sciences, Universidad de Chile*
[3]*Department of Computer Science, Pontificia Universidad Católica de Chile, Chile*
[4]*Institute for Mathematical and Computational Engineering, Pontificia Universidad Católica de Chile, Chile*
[5]*Center for Mathematical Modeling, Universidad de Chile, Chile*
[6]*Millenium Institute Foundational Research on Data, Chile*

Abstract
Relation extraction is an essential component of Natural Language Processing (NLP) and significantly influences information retrieval and structured information extraction. Within clinical notes, the task is needed to establish connections among illnesses, therapies, indications, and other medical concepts. Motivated by the above, in this work, we propose a two-step model approach for entity linking; in the first step, we solve entity recognition, and in the second, a relation classification approach. We evaluated our approach in a Spanish corpus of the *TESTLINK* challenge in *IberLEF2023* (Iberian Languages Evaluation Forum), comprising 81 clinical notes to train and 80 clinical notes to test. Our results show competitive performance with a *precision* of 0.47, *recall* of 0.43, and *F1-score* of 0.45, presenting an effective strategy for relation extraction from clinical notes in Spanish.

Keywords
Natural Language Processing, Link prediction, Named Entity Recognition, Clinical Text

1. Introduction

The increasing complexity of healthcare services has accentuated the importance of clinical notes as indispensable sources of insights into patients' health conditions. These documents contain data from clinical visits, physical examinations, diagnoses, and follow-up treatments and often encompass critical outcomes of laboratory tests and measurements - integral elements in disease and disorder diagnosis. However, the extraction of these pertinent data, particularly from Spanish-language documents, has yet to be explored, and research on this field, especially in the Spanish language, still needs to be explored.

This paper introduces *LinkMed*, our proposed method, to the *TESTLINK* task at *IberLEF2023* [1]. This task, grounded in clinical cases from the E3C corpus, poses a challenge of relation extraction from clinical narratives. It demands identifying test results and measurements within the text, establishing links between these results, and the textual mentions of the corresponding laboratory tests and measurements.

IberLEF 2023, September 2023, Jaén, Spain

Differing from a conventional Named Entity Recognition (NER) task, this challenge requires the interpretation of numeric values and ranges, establishing its identity as a Relation Extraction (RE) task that considers elements involved in the relation and its directionality.

Our approach, *LinkMed*, introduces a two-step solution to this challenge. Initially, a Named Entity Recognition (NER) model is used to identify potential entities of interest in the clinical notes that could be linked. The NER approach is followed by a relation classification system on all the combinations of found pairs to ascertain the existence of a valid relation between those identified entities. Our method mainly addresses the entity-linking problem in Spanish clinical notes.

This paper presents an in-depth description of our proposed solution to the *TESTLINK* task. While the challenge encompasses both Spanish and Basque languages, our work focuses explicitly on the former, thus enriching resources available for Spanish-language clinical documents and promoting thorough patient care and clinical decision-making processes within this linguistic context.

2. Related work

Entity linking for medical text analysis presents a unique challenge in non-English corpora, exacerbated by the scarcity of resources available for other languages [2]. The methodologies applied to address this task predominantly fall into two distinct categories: **rule-based** systems and **machine learning-based** approaches.

The source of **rule-based** systems, initially designed to facilitate medical evidence searches in databases like MEDLINE by identifying specific medical terms in texts [3], marked a significant breakthrough. Systems such as CLARIT [4], SAPHIRE [5], and MetaMap [6] capitalized on linguistic rules and dictionaries to map concept mentions to Medical Subject Headings (MeSH) terms, significantly enhancing the interpretability and accessibility of medical text data. Subsequent systems such as CHARTLINE [7] and MedLEE [8] expanded upon these ideas, employing dictionary-matching techniques to extract and link entities within clinical reports to the Unified Medical Language System (UMLS). Innovations like REX [9] pushed boundaries by linking clinical note mentions to ICD-9-CM codes, thereby aiding medical record coding. However, the main limitation of rule-based systems is their struggle with semantic understanding and the diverse terminology present in clinical narratives [10, 11, 12].

On the other hand, **machine learning-based** methods transitioned entity linking from a mere matching problem to a complex mapping task, leveraging numerical representations of mentions and concepts [13]. The emergence of deep learning techniques and contextual embeddings, such as ELMo [14] and BERT [15], caused a paradigm shift in entity-linking research. Currently, the majority of state-of-the-art systems employ deep contextualized embeddings, combining these with a variety of methods, including binary [16], multi-class [17], and clustering approaches [18]. However, a persistent challenge in this domain is the scarcity of resources for effective training of entity-linking models.

Despite significant advances in Spanish language models such as BETO [19] and DistillBETO [20], along with the development of their evaluation frameworks for both general-domain [21, 22] and biomedical [23, 24, 25] contexts, the specific problem of entity linking in medical texts remains relatively unexplored.

This work proposes a methodology that integrates the strengths of **machine-learning** focused on natural language processing and a pairwise text classification approach. In detail, we first utilize a transformer-based Named Entity Recognition (NER) model to identify potential entities, followed by a relation classification method to determine potential links. This novel approach addresses the present limitations in entity-linking solutions and contributes to the underexplored area of Spanish entity-linking in medical texts.

3. Dataset

We assessed the method used for clinical cases extracted from the *E3C* corpus [26], featured in the *TESTLINK* challenge in *IberLEF2023* [1, 27, 28], with a primary emphasis on documents in the Spanish language. This clinical corpus represents a medical history of different anonymized patients, where we can find; a general patient description, the reason for a visit, the medical history associated with the consultation, the diagnosis, the results of treatments, and more.

The composition of the dataset is 81 documents for training and 80 for testing; the first has 597, and the second has 668 annotated relations for humans. Both are under *PubTator format* [1], indicating an ordered pair of entity mentions (i.e., *RML*, *EVENT*). *RML (Resource Mapping Language)* is a tag created to mark test results, and the tag *EVENT* corresponds to activities, conditions, and situations significant to an individual's medical history.

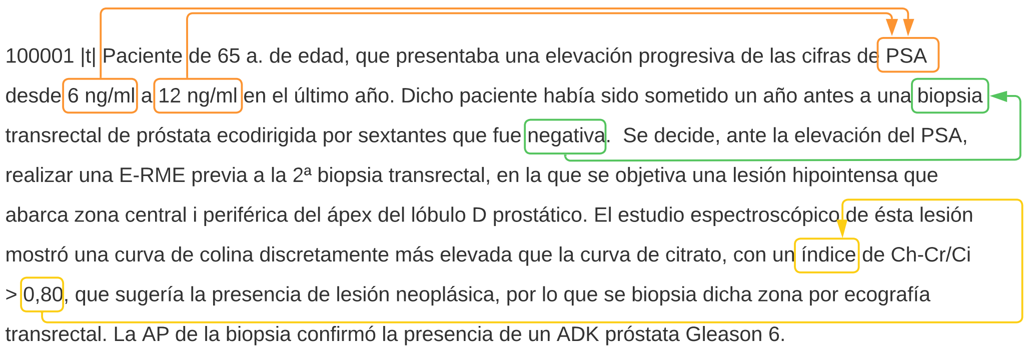

Figure 1: Training example number 10001, composed by the following relationships: *6 ng/ml* $\rightarrow$ *PSA*, *12 ng/ml* $\rightarrow$ *PSA*, *negativa* $\rightarrow$ *biopsia*, and *0,80* $\rightarrow$ *índice*

.

An example of the task is shown in Figure 1. It can be seen the task of entity linking within clinical narratives in Spanish presents intricate complexity. One element contributing to this complexity is the variable context size surrounding an entity pair; it may be either broad or minimal. Moreover, the entities may comprise multiple tokens, adding further difficulty. Additionally, the directionality of relationships between these entities is subject to change, further complicating the task. The multifaceted structure of this task underscores the fact that the *TESTLINK* challenge [1] is far from a simple problem to solve, instead manifesting as a compelling challenge within the realm of Clinical Natural Language Processing (NLP). Thus, it is essential to explore and decipher these complexities to advance in the field and improve the efficiency and accuracy of clinical data interpretation.

4. Proposed Model

To address the relation extraction task, we propose *LinkMed*, a deep learning model based on two sequential steps; entity recognition and the relation classification modules. Specifically, two models were created: a NER and a relation classification model. The NER model obtains mentions of tag events and their corresponding results. Then, the classification model takes those pair of mentions and predicts whether there is a relation between them. Figure 2 depicts the relation between the two used modules in *LinkMed*.

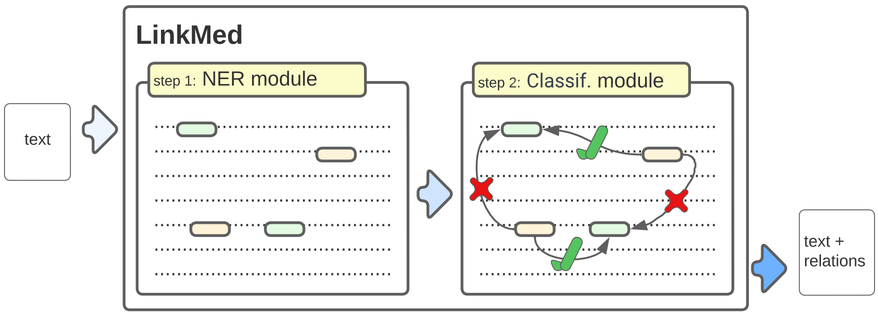

Figure 2: The *LinkMed* diagram involves two distinct steps. Initially, it identifies entities in Clinical Notes, such as test results and measurement categories. Then, the second component discerns relationships between all the combinations of these found entities.

.

Our approach for solving the *Teslink* task is shown in Figure 2, which indicates that the task consists of two consecutive components: a NER module and a relationship classification module between the found entities. Regarding the NER model, we solved the task using the FLERT model [29], which has shown outstanding results in Spanish NER tasks [30]. This approach consists of fine-tuning a transformer-based model, not at the sentence level but at the document level. In other words, the input of the model considers not only the sequence of tokens of the current sentence but also a window of tokens of the previous and following sentences, thus incorporating more context. For our experiments, we tested three language models; biomedical and clinical versions of RoBERTa [31], and the Spanish version of BERT [19].

Then the model used for the relation extraction module followed the same architecture. We fine-tuned a domain-specific language model to create contextual representations of the spans found in the NER module. Then, these representations are fed into a linear layer to determine whether there is a relation between both entities. We are once again contemplating a focus on the level of documents. In experiments using the validation partition, our NER model obtained a mean of 0.84 according to the *F1-score* using the biomedical version of RoBERTta, and the combination of both modules obtained a 0.51 *F1-score* using the official evaluation script. In both modules, the language model that obtained the best results was the biomedical version of RoBERTa.

5. Results

After obtaining the results of our proposed solution, we employed a range of metrics to measure our model's performance. These include False Positives (*FP*), False Negatives (*FN*), *precision*, *recall*, and the *F1-score* as shown in Table 1. The model's efficacy was assessed utilizing a test dataset containing distinct entities and their corresponding relationships.

Concerning the general performance of the model shown in the last row (*all*) of Table 1, the system registered 326 *False Positive* instances, denoting scenarios where it incorrectly recognized a relationship between two entities within the test dataset. In addition, it failed to identify 379 existing relationships, classifying them as *False Negatives*. The *precision* of the model, a ratio representing accurately identified relationships over the sum of identified relationships, approximates 0.47. This suggests that nearly 47% of the relationships the system identified align with those defined within the test set. The model's *recall*, calculated as the proportion of accurately identified relationships and overall actual relationships in the test set, approximates 0.43. This inference reveals that the system correctly identified around 43% of all actual relationships embedded within the test set context. Finally, the *F1-score*, representing the harmonic mean of *precision* and *recall*, was determined to be approximately 0.45. This score embodies a trade-off between *precision* and *recall*, indicating a need for further refinement in the model's performance. Collectively, these findings elucidate both the promising prospects and inherent challenges of entity-relationship recognition within defined contexts. The occurrence of both *false positives* and *false negatives* pinpoints areas for potential enhancement in the model's performance.

token division	**count**				**evaluation metrics**		
	# relations	**TP**	**FP**	**FN**	**precision**	**recall**	**F1-score**
single	326	170	177	156	0.49	0.52	0.51
two	250	76	92	174	0.45	0.30	0.36
multiple	92	43	57	49	0.43	0.47	0.45
all	668	289	326	379	0.47	0.43	0.45

Table 1
Evaluation results for different entity categories based on the token quantity of *RML*. The categories classification are; *single* as one token, *two* as two tokens, *multiple* as more than two tokens, and *all* tokens.

In order to clarify the overall results shown in Table 1, we divided the data into three classifications (first three rows of Table 1) based on the number of tokens present in the *RML* entity: *single* token, *two* tokens, and *multiple* tokens. As the event entity (*EVENT*) invariably contained only one token, it did not facilitate the creation of these categories.

The classification of the *single* token obtained a *precision* of 0.49, a *recall* of 0.52, and an *F1-score* of 0.51. Interestingly, the *single-token* category outperformed the general results and other classifications across all the metrics evaluated, thereby indicating a robust structure for resolving the composite entity comprising one token.

In contrast, comparing the *two-token* and *multiple-token* results reveals an interesting pattern, deviating from the established trend. The performance in the *multiple-token* category surpassed that of the *two-token* category, yielding a recall of 0.47 compared to 0.30 and an *F1-score* of

0.45 compared to 0.36. This suggests that the model employed in this study demonstrated a heightened performance under more complex circumstances. Specifically, there was a relative increase in True Positives (*TP*) and a relative decrease in False Negatives (*FN*) concerning the evaluated category. This unique performance underlines the capability of the model to adapt and perform proficiently, even under more challenging conditions.

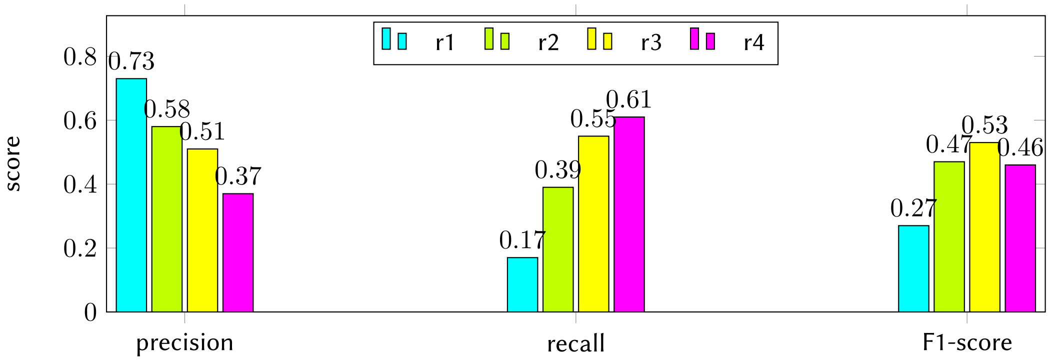

Figure 3: Comparison of metrics for different character distance categories between entity pair. The total of relationships was separated into four groups with a similar number of relationships, motivated by the amount of context needed to generate the relationship between entities. Where *r1* indicates one, *r2* is between 2 and 8, *r3* is between 9 and 25, and *r4* is greater than 26 characters between entities.

Figure 3 presents a comparative analysis of various context levels, approximated by the number of characters separating entity pairs. The first level portrays a close relationship between entities, typically a single space. This category yields the highest *precision* of 0.73, contrasted with the lowest *recall* of 0.17 and an *F1-score* of 0.27, compared with other categories. The observed results could be attributed to the considerable number of False Negative (*FN*) cases (135) and the system's reduced capability to detect closely related True Positives (*TP*) (27 out of 162). This finding suggests the potential for incorporating a rule-based approach in future models. The second level, characterized by a character difference ranging from 2 to 8, signifies scenarios where the entity pair has some words intervening. Here, the model demonstrates commendable performance with a *precision* of 0.58, *recall* of 0.39, and an *F1-score* of 0.47, reflecting its ability to comprehend the immediate context and establish relationships. A similar pattern is observed in the third level, extending from 9 to 25 spaces, which indicates an expanded context within the clinical text. The weighted *F1-score* in this category approximates 0.53, presenting a balanced option for identifying relationships, considering both quality and quantity. Lastly, the fourth level, characterized by more than 26 characters, displays a substantial capability to identify 108 *TP* instances while showing a reduced capacity to detect *FP* cases, with 181 instances identified. This result could be attributed to the complexity of comprehending larger contextual expanses within the text.

6. Limitations

Our proposed model, while achieving commendable performance in various scenarios for section identification in Electronic Clinical Narratives (ECNs), has its limitations. One of the most significant restrictions stems from the text-chunking module. When this component incorrectly identifies a section, the error tends to propagate to subsequent sections due to the sequential nature of the task, where all parts of the text have a designated section. This has a cascading effect on the precision of the classification for the entire ECN.

Another limitation relates to the interdependency of the two modules of our model. As the performance of the section classification module is inherently dependent on the accuracy of the text-chunking module, any classification error can negatively impact the overall matching accuracy. Thus, the combined performance of both modules is a critical factor for the effectiveness of our model.

7. Conclusions and Future Work

In conclusion, our hybrid approach effectively drives the challenges of entity linking in the Spanish clinical domain. Combining the strengths of a transformer-based Named Entity Recognition (NER) model with a pair-wise classification module, we successfully identify relevant entities and their potential relationships within clinical notes. However, there is still space for improvement, particularly within the pair-wise classification component. Its current design has limitations in capturing semantic relatedness between entities and the directional nature of their relationships, which can impede overall performance. Future work should enhance this component by integrating a method capable of discerning semantic relatedness and directionality among identified entities. Ultimately, such advancements will improve the model's performance and further contribute to exploring the entity linking task in Clinical-NLP.

Acknowledgements

This work was funded by ANID Chile: Basal Funds for Center of Excellence FB210017 (CENIA), FB210005 (CMM); Millennium Science Initiative Program ICN17_002 (IMFD) and ICN2021_004 (iHealth), Fondecyt grant 11201250, and National Doctoral Scholarships 21211659 (Claudio Aracena) and 21221155 (Carlos Muñoz-Castro). This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02) and the Patagón supercomputer of Universidad Austral de Chile (FONDEQUIP EQM180042).

References

[1] IberLEF, TESTLINK@IberLEF 2023, 2023. URL: https://e3c.fbk.eu/testlinkiberlef.
[2] E. French, B. T. McInnes, An overview of biomedical entity linking throughout the years, Journal of Biomedical Informatics (2023).
[3] H. J. Lowe, G. O. Barnett, Micromesh: a microcomputer system for searching and exploring the national library of medicine's medical subject headings (mesh) vocabulary, in: Pro-

ceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association, 1987, p. 717.

[4] D. A. Evans, K. Ginther-Webster, M. Hart, R. G. Lefferts, I. A. Monarch, Automatic indexing using selective nlp and first-order thesauri, in: Intelligent Text and Image Handling-Volume 2, 1991, pp. 624–643.

[5] W. Hersh, T. Leone, The saphire server: a new algorithm and implementation., in: Proceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association, 1995, p. 858.

[6] A. R. Aronson, Effective mapping of biomedical text to the umls metathesaurus: the metamap program., in: Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, p. 17.

[7] R. A. Miller, F. M. Gieszczykiewicz, J. K. Vries, G. F. Cooper, Chartline: providing bibliographic references relevant to patient charts using the umls metathesaurus knowledge sources., in: Proceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association, 1992, p. 86.

[8] C. Friedman, G. Hripcsak, W. DuMouchel, S. B. Johnson, P. D. Clayton, Natural language processing in an operational clinical information system, Natural Language Engineering 1 (1995) 83–108.

[9] J. Friedlin, C. J. McDonald, A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports, in: AMIA annual symposium proceedings, volume 2006, American Medical Informatics Association, 2006, p. 269.

[10] J. D'Souza, V. Ng, Sieve-based entity linking for the biomedical domain, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015, pp. 297–302.

[11] H. Liu, S. T. Wu, D. Li, S. Jonnalagadda, S. Sohn, K. Wagholikar, P. J. Haug, S. M. Huff, C. G. Chute, Towards a semantic lexicon for clinical natural language processing, in: AMIA Annual Symposium Proceedings, volume 2012, American Medical Informatics Association, 2012, p. 568.

[12] A. Leal, B. Martins, F. M. Couto, Ulisboa: Recognition and normalization of medical concepts, in: proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015, pp. 406–411.

[13] R. Leaman, R. Islamaj Doğan, Z. Lu, Dnorm: disease name normalization with pairwise learning to rank, Bioinformatics 29 (2013) 2909–2917.

[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations. arxiv preprint, arXiv preprint arXiv:1802.05365 (2018).

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[16] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.

[17] W.-T. Kao, H.-y. Lee, Is bert a cross-disciplinary knowledge learner? a surprising finding

of pre-trained models' transferability, arXiv preprint arXiv:2103.07162 (2021).
[18] A. Subakti, H. Murfi, N. Hariadi, The performance of bert as data representation of text clustering, Journal of big Data 9 (2022) 1–21.
[19] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, Pml4dc at iclr 2020 (2020) 1–10.
[20] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, Albeto and distilbeto: Lightweight spanish language models, arXiv preprint arXiv:2204.09145 (2022).
[21] V. Araujo, A. Carvallo, S. Kundu, J. Cañete, M. Mendoza, R. E. Mercer, F. Bravo-Marquez, M.-F. Moens, A. Soto, Evaluation benchmarks for spanish sentence representations, arXiv preprint arXiv:2204.07571 (2022).
[22] C. Aspillaga, A. Carvallo, V. Araujo, Stress test evaluation of transformer-based models in natural language understanding tasks, arXiv preprint arXiv:2002.06261 (2020).
[23] V. Araujo, A. Carvallo, C. Aspillaga, C. Thorne, D. Parra, Stress test evaluation of biomedical word embeddings, arXiv preprint arXiv:2107.11652 (2021).
[24] V. Araujo, A. Carvallo, C. Aspillaga, D. Parra, On adversarial examples for biomedical nlp tasks, arXiv preprint arXiv:2004.11157 (2020).
[25] A. Carvallo, D. Parra, Comparing word embeddings for document screening based on active learning., in: BIRNDL@ SIGIR, 2019, pp. 100–107.
[26] B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli, The e3c project: Collection and annotation of a multilingual corpus of clinical cases, Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 (2020).
[27] 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.
[28] B. Altuna, R. Agerri, L. Salas-Espejo, J. J. Saiz, R. Zanoli, M. Speranza, B. Magnini, A. Lavelli, G. Karunakaran, Overview of TESTLINK at IberLEF 2023: Linking Results to Clinical Laboratory Tests and Measurements, Procesamiento del Lenguaje Natural 71 (2023).
[29] S. Schweter, A. Akbik, Flert: Document-level features for named entity recognition, ArXiv abs/2011.06993 (2020).
[30] M. Rojas, J. Barros, K. Martin, M. Araneda-Hernandez, J. Dunstan, PLN CMM at SocialDisNER: Improving Detection of Disease Mentions in Tweets by Using Document-Level Features, in: Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task, Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 52–54. URL: https://aclanthology.org/2022.smm4h-1.15.
[31] 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.