BIT.UA at MultiCardioNER: Adapting a Multi-head CRF for Cardiology

Notebook for the BioASQ Lab at CLEF 2024

Richard A. A. Jonker[1],*, Tiago Almeida[1] and Sérgio Matos[1]

[1] IEETA/DETI, LASI, University of Aveiro, Aveiro, Portugal

Abstract

This paper presents the participation of the University of Aveiro Biomedical Informatics and Technologies (BIT.UA) group in the MultiCardioNER task at BioASQ 12, specifically in the CardioDis subtrack, which focuses on adapting Named Entity Recognition (NER) systems to Spanish cardiology case reports. We aimed to address two primary research questions: 1) the generalizability of a NER model trained on general medical concepts to the specialized sub-domain of cardiology, and 2) the robustness of our Multi-Head CRF model. Our team achieved the top result in the competition with an F1 score of 81.99, using the Multi-Head CRF model. Our findings indicate that task-specific data is beneficial to the overall performance of the model, although a model without this data can still be competitive. Additionally, our Multi-Head CRF model demonstrated consistent reliability and robustness, performing well on single-class NER tasks.

Keywords

Named Entity Recognition, Spanish Clinical Procedures, Transformers, Data Augmentation, Multi-head CRF, Robust ML

1. Introduction

Named Entity Recognition (NER) is a fundamental task in the field of natural language processing, especially crucial in the medical domain where it aids significantly in structuring unstructured text for enhanced patient care and medical research. While general NER technologies have seen considerable advancement, their application to medical texts presents unique challenges due to the complexity and specificity of the medical language. To address these challenges, numerous competitions have been organized to foster the development of NER systems specifically tailored to the biomedical domain.

Our team has continually engaged in these competitions, gaining experience through participation in BioCreative events such as the NLM-Chem Track [1, 2] and the BioRED Track [3, 4], both of which were focused on English biomedical articles. These challenges, allowed us to build a solid foundation in NER methodologies, initially leveraging state-of-the-art BERT-based models and masked Conditional Random Fields (CRF) [2, 5] over BIO-tagged sequences [6, 7]. We subsequently expanded our efforts to include Spanish medical texts, participating in challenges like MedProcNER [8] and SYMPTEMIST [9], where we secured first and second places respectively in the NER evaluations. This extensive experience has led to the creation of our versatile Multi-Head CRF model [10], a highly competitive NER solution that encapsulates our accumulated expertise.

This year, as part of the BioASQ challenge [11], the Text Mining Unit (TEMU) at Barcelona Supercomputing Center (BSC), introduced the MultiCardioNER [12, 13] challenge. This challenge addresses the need for better recognition of clinical variables in cardiology, given the high mortality rate from cardiovascular diseases (CVDs), which cause approximately 17.9 million deaths annually [14]. It includes two subtracks: CardioDis, which adapts NER systems to Spanish cardiology case reports, and MultiDrug, which tests these systems on medication mentions in English, Spanish, and Italian. The dataset includes a training set of 1,000 general clinical case reports, a development set of 258 cardiology cases, and a test

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

richard.jonker@ua.pt (R. A. A. Jonker); tiagomeloalmeida@ua.pt (T. Almeida); aleixomatos@ua.pt (S. Matos)
0000-0002-3806-6940 (R. A. A. Jonker); 0000-0002-4258-3350 (T. Almeida); 0000-0003-1941-3983 (S. Matos)

set of 250 cardiology reports. Systems are evaluated on micro-averaged Precision, Recall, and F-measure to determine their adaptability and accuracy in diverse medical settings.

This paper details the participation of the Biomedical Informatics and Technologies of University of Aveiro (BIT.UA) in the MultiCardioNER challenge, where we utilize our Multi-Head CRF model [10]. Due to time constraints, we focused our efforts solely on the CardioDis subtrack. Unlike other challenges, the CardioDis subtrack is designed to adapt general concept recognition systems, trained on the DISTEMIST dataset, to cardiology case reports. This leads to our first research question:

1. How effectively can a model trained on general concepts adapt to the specialized domain of cardiology cases?

Additionally, we are utilizing this challenge to test the robustness of our Multi-Head CRF model, prompting our second research question:

2 Is our Multi-Head CRF model capable of delivering competitive out-of-the-box performance in a specialized clinical setting?

The remainder of the paper is organized as follows: Section 2 provides a review of related work, focusing on the latest advancements in biomedical Named Entity Recognition. Section 3 describes our methodology, detailing the application of our Multi-Head CRF model to address the specific challenges of the MultiCardioNER competition. Section 4 presents our validation results and the official challenge evaluations, showcasing the performance of our model. In Section 5, we discuss the outcomes in relation to our initial research questions, providing insights into the adaptability and robustness of our approach. Section 6 concludes the paper, summarizing our key findings.

2. Related Work

Named Entity Recognition (NER) in the biomedical domain presents unique challenges due to the limited availability of annotated data. The annotation process is both time-consuming and requires a high level of expertise, making it expensive [15, 16]. Most research in biomedical NER has been focused on the English language [17], but there is a growing need to extend this work to other languages.

Several competitions have aimed to address clinical NER in the Spanish language, focusing on various entity types such as compounds and drugs (PharmaCoNER [18]), diseases (DisTEMIST [19]), tumor morphology (CANTEMIST [20]), medical procedures (MedProcNER [21]), and symptoms (SympTEMIST [22]). All of these competitions utilize the Spanish Clinical Case Corpus (SPACCC), which comprises 1,000 clinical case reports from Spanish medical publications (SciELO).

Recent advancements in NER predominantly utilize transformer-based models for sequence labelling, which have proven effective in managing complex entity recognition tasks [23, 8, 24, 25]. These models often leverage pretrained language-specific versions of BERT [26] and RoBERTa [27], such as BETO, a BERT model trained on a Spanish corpus [28], and *bsc-bio-es*, a RoBERTa model tailored to Spanish biomedical vocabulary [29].

Further enhancements have been observed by integrating masked Conditional Random Fields (CRFs)[2, 5] atop the transformer backbone, a technique that our research group has profoundly explored [1, 2, 3, 4, 8, 9]. In these works, we also demonstrate strong transfer-learning capabilities, effectively adapting a Spanish NER model trained on clinical notes [30] to the diverse target domains.

Another significant development is the introduction of the SpanMarker model, which utilizes the novel Packed Levitated Markers (PL-Marker) approach [31]. This model enhances NER performance by employing a neighborhood-oriented packing strategy to accurately model entity boundaries and a subject-oriented strategy for complex span pair classification tasks. Concurrently, innovative methods like those in the AIONER system, which prepends special tokens to input texts, enable the adaptation to annotated corpora lacking comprehensive coverage of all entity classes [32]. Similarly, HunFlair2 utilizes a multi-class BIO tagging scheme, enhancing its ability to distinguish between various entity types such as genes and diseases, thereby showcasing the adaptability and versatility of these advanced NER systems [33].

3. Methodology

In this section, we describe the dataset, the evaluation metrics used, and provide a brief overview of the methodology used.

3.1. Dataset

The MultiCardioNER challenge utilizes two datasets: the previously released DisTEMIST dataset and the newly released CardioCCC dataset. The DisTEMIST dataset comprises 1,000 documents from the SPACCC corpus, annotated with disease mentions. For the validation and evaluation of the system, the CardioCCC dataset was created. This collection consists of 508 cardiology clinical case reports, split into 258 documents for development and 250 for testing. The goal of the task is to train a generic system capable of classifying diseases and to evaluate it within the more specific cardiology domain. Whilst the goal of the competition is to evaluate the adaption of the model, we utilize the validation set in order to train some models[1].

3.2. Evaluation Metrics

The official metrics used in this work are the standard micro-averaged precision, recall and F1-scores.

- **Precision (P)**: The ratio of true positive (TP) predictions to the total number of positive predictions (TP + FP). It is defined as:
$$P = \frac{TP}{TP + FP}.$$
- **Recall (R)**: The ratio of true positive (TP) predictions to the total number of actual positives (TP + FN). It is defined as:
$$R = \frac{TP}{TP + FN}.$$
- **F1-Score (F1)**: The harmonic mean of precision and recall. It is defined as:
$$F1 = 2 \cdot \frac{P \cdot R}{P + R}.$$

3.3. System

The system utilizes the Multi-Head CRF model as a basis, in order to test the secondary objective of this work, the robustness of the multi-head-CRF model [10]. All the work presented in this work is done utilising the same code and methods from that work. Whilst the multi-head CRF architecture is designed, and tested for performing multi-class NER, in this work we configure the architecture to use only one head for single class classification as illustrated in Figure 1. Whilst the general architecture is the same, in order to keep this work self-contained, we provide a brief overview of the Multi-Head-CRF architecture.

The architecture was inspired by several existing works [1, 2, 8, 9], achieving competitive results in various challenges. The main idea behind the architecture is to utilize several CRF heads, one per entity class, with a shared transformer as a base, in this case using a Spanish RoBERTa model [30]. By having several classification heads, we can solve the problem of overlapping entity classes, since each entity is trained separately. However, since each head shares the same transformer, significant overhead is reduced compared to training several individual classifiers. Going more in-depth, the work utilizes the well-known BIO tagging schema, where each entity has its own tagging schema assigned to it. The CRF classification heads utilize several dense layers, a classification layer, and a CRF layer. Each of these heads then produces a series of labels corresponding to the BIO tagging for the specific entity of the head. The model is trained using a joint loss function, aggregated from each classification head.

[1] The event organizers explicitly mentioned: "Participants are encouraged to experiment with the documents and annotations as they see fit."

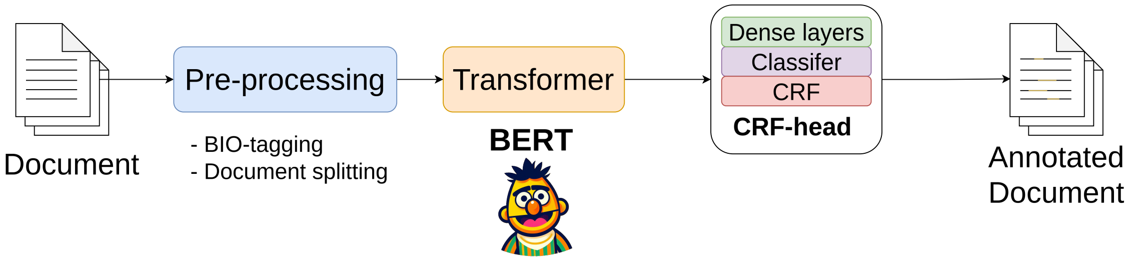

Figure 1: Multi-head CRF architecture, adapted for MultiCardioNER.

The model also employs a document splitting system to overcome the maximum context length of the transformer. Each document is split in a sliding window fashion, with each piece of the document being encapsulated with a fixed length context. The work also utilizes some data augmentation techniques, namely random token replacement and a variation, random token replacement with unknown. In the first technique, a random input token is replaced with a random token from the vocabulary, while in the latter, the token is replaced with a special token '[UNK]', however in this work we follow the conclusions of the original work, and utilize only random token replacement, as it performed better. To better control the augmentation, two hyperparameters were put in place: one determines the chance of selecting a sample for augmentation (the augmentation probability), and the other determines how many tokens within the sample get augmented (percentage tags).

Finally, following the work of our previous NER submission [9, 8], we employed an entity-level ensemble to merge the outputs from various models, which proved to improve overall results. The entity-level ensemble is a majority voting approach over the exact entities predicted by the models, where each entity is added to the final submission if enough support is present for the given entity.

4. Results

In this section we will present the results obtained with the proposed system. Initially, we evaluate the performance of the model on a validation set, before discussing the results on the final test set of the competition.

4.1. Validation results

In order to find the optimal hyperparameters for the models to be submitted, we performed basic hyperparameter tuning, investigating varying amounts of training epochs, different augmentation configurations, and adjusting the context size and number of hidden layers. The validation set used for this work was the 258 document development set, containing cardiac data. Figures describing this basic hyperparameter search can be seen in Figures 2 - 4. The best-performing model configurations on validation can be seen in Table 1.

Looking first at Figure 2, we can see that the performance difference between random augmentation and no augmentation is significant, and the use of augmentation improves the overall performance of the system. This is inline with the conclusions drawn from the original multi-head CRF model, however we did not investigate the use of the 'unk' augmentation technique. We also note that training for more epochs does not necessarily result in significant performance gains, especially for models with random augmentation.

Examining the optimal augmentation configuration in Figure 3, we observe that lower values of percentage tags perform better, especially with increasing augmentation probabilities. This corresponds to selecting a large number of documents and augmenting a small number of tokens. While this is not the exact same configurations used in the original Multi-Head CRF work, the configuration is relatively similar, with similar conclusions being drawn.

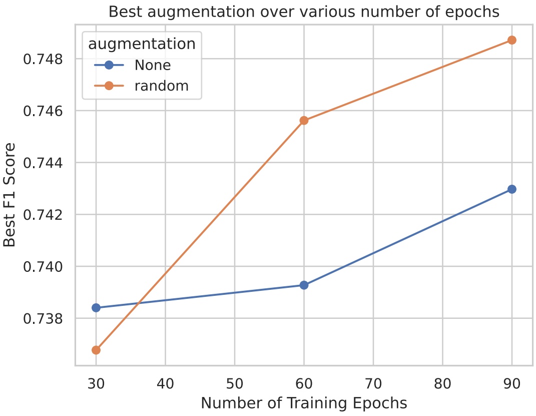

Figure 2: Line graph describing the best augmentation technique against no augmentation with varying number of epochs.

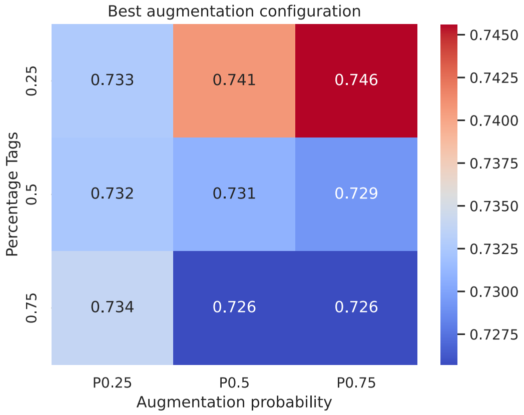

Figure 3: Heatmap showcasing the best augmentation configuration.

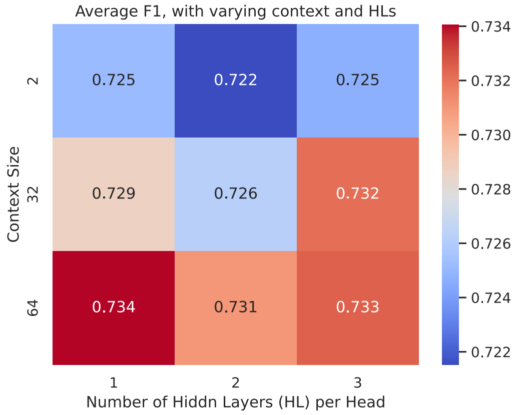

Figure 4
Heatmap showing the average performance of models with varying context size and Hidden Layers in the CRF head.

Table 1
Top 5 model configurations on validation data. Ctx. represents the context size of the model. HL represents the number of hidden layers used in the CRF head, aug. is either random augmentation or None, PT is the percentage tags and AP is the augmentation probability.

Ctx.	HL	Epoch	Aug.	PT	AP	F1
64	1	90	random	0.25	0.75	74.87
64	1	60	random	0.25	0.75	74.56
32	3	90	random	0.25	0.75	74.40
64	3	90	None	0.2	0.5	74.30
64	3	60	random	0.25	0.75	74.21

Finally, discussing the context size and number of hidden layers as described in Figure 4, we note that higher context size performs better on average, with the optimal number of hidden layers being either 1 or 3, which is the same as the optimal model in the original paper. We also note that the performance for our 123 model search ranged from 69.47 to 74.87, with an average of 72.89 and a median of 72.96.

4.2. Competition results

Below we present the official results of our systems in the competition. The competition uses the F1-score as the official metric, with the test set containing 250 documents.

For the competition, we submitted five different systems. We followed two separate approaches. The first approach was to keep the validation set separate in order to test the adaptability of a system trained exclusively on diseases to directly identify cardiology-related concepts. Our second approach was to utilize the validation set to train a model, in combination with the generic disease data, with the intuition that more data is always beneficial for training a model. A summary of our submitted systems is below, with there performance of the models being displayed in Table 2.

- **run0**: This submission used an ensemble of our top 5 validation models (ranging from 74.20-74.87), trained using all data including the validation set.

- **run1**: This submission used an ensemble of top 17 runs, trained using all data including the validation set. (all above 74)
- **run2**: This submission used our best performing model on validation, trained without validation data.
- **run3**: This submission used an ensemble of our top 24 runs, all trained without validation data.
- **run4**: This represents an ensemble of all submissions containing 41 runs.

Table 2
Official competition results. The relative rank corresponds to the ranking of the systems against our best system, as the competition organizer did not provide an official ranking, with only the best and median systems availble as comparison.

System	Precision	Recall	F1 Score	Relative Rank
run0-top5-full	81.10	81.81	81.45	2
run1-all-full	81.55	**82.43**	**81.99**	1
run2-best-val	74.80	75.42	75.11	5
run3-all-val	75.44	75.88	75.66	4
run4-all	79.81	78.27	79.03	3
Best	89.19	82.43	81.99	-
Median	-	-	72.29	-

Looking at the results, we obtained the best submission in the competition. This was achieved by using a large ensemble with multiple runs that included validation data for training. It is not surprising that run1 outperformed run0. Generally, from previous work, we have observed that larger ensembles tend to perform better. We also have no assurance that the top 5 models on validation would achieve the best results on the test set. Next, we note that the performance of the models trained using the validation data greatly outperformed those that were not. While the performance was 6 percentage points higher, the base system still performed relatively well, considering it was not directly trained for the cardiac domain. Similarly, to the comparison of run0 and run1, we see a slight performance increase with run3 over run2 with the increase in the number of models. However, the performance gains were not as significant. Finally, given these results, our final submission, run4, performed as expected—somewhere between our best and worst models—given the large discrepancy in data performances. Overall, all our submissions achieved F1 scores above the median, with our best submission obtaining the top F1 score and recall in the competition.

5. Discussion

In this work, we proposed two research questions: 1. investigating the impact of training with validation data for domain-adaptation and 2. investigating the robustness of the multi-head CRF model.

With regards to the first research question, we can look to our submission results. We utilized two different approaches to see the performance difference when using the validation data in training. The conclusions we drew indicate that the models have a significant performance gain when using the validation data. This was an expected outcome; however, the models that did not have access to this data still obtained competitive performance, being well above the average submission.

Discussing the second research question, we can see the robustness of the model by looking at both the overall performance of the model in the competition and our validation results. Our model performed well overall, obtaining the top performance within the competition. Looking at the validation we performed, we obtained comparative results to the original work, indicating the overall robustness and reliability of the model, showcasing its ability to perform well on not only multi-class NER but also single-class NER.

6. Conclusion

This study aimed to address two research questions using the MultiCardioNER challenge as a basis. The first question, related to the task, was whether an NER model trained on a specific domain—in this case, Diseases—could generalize to a sub-domain, in this case, cardiology data. The conclusions we drew from our work indicate that while the system does perform well in generalization, better performance is always achieved when using task-specific data.

Our second research question was aligned with our previous work and focused on testing the robustness of our Multi-Head CRF architecture [10]. In this work, we demonstrated that the architecture is indeed robust, achieving top performance in the competition, which utilizes only a single entity, as opposed to previous work which utilizes multi-class NER. We further observed the robustness of the model within our validation tests, where we drew many similar conclusions to those of the original work. Overall, we believe that this Multi-Head CRF architecture stands as a solid basis for future work in NER.

Acknowledgments

This work was funded by the Foundation for Science and Technology (FCT) in the context of the project doi.org/10.54499/UIDB/00127/2020. Tiago Almeida is funded by the grant doi.org/10.54499/2020.05784.BD. Richard A. A. Jonker is funded by the grant PRT/BD/154792/2023. This work was funded by FCT I.P. under the project Advanced Computing Project 2023.10766.CPCA.A0, platform Vision at University of Évora.

References

[1] T. Almeida, R. Antunes, J. F. Silva, J. R. Almeida, S. Matos, Chemical detection and indexing in pubmed full text articles using deep learning and rule-based methods, CDR 1500 (2021) 15943.

[2] T. Almeida, R. Antunes, J. F. Silva, J. R. Almeida, S. Matos, Chemical identification and indexing in PubMed full-text articles using deep learning and heuristics, Database 2022 (2022) baac047. URL: https://doi.org/10.1093/database/baac047. doi:10.1093/database/baac047.

[3] T. Almeida, R. A. A. Jonker, D. da Silva, J. Almeida, S. Matos, BIT.UA at Biocreative VIII track 1: A joint model for relation classification and novelty detection, 2023. URL: https://doi.org/10.5281/zenodo.10117952. doi:10.5281/zenodo.10117952.

[4] T. Almeida, R. A. A. Jonker, R. Antunes, J. R. Almeida, S. Matos, Towards Discovery: An End-to-End System for Uncovering Novel Biomedical Relations, Database (to appear) 2024 (2024).

[5] T. Wei, J. Qi, S. He, S. Sun, Masked conditional random fields for sequence labeling, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2024–2035. URL: https://aclanthology.org/2021.naacl-main.163. doi:10.18653/v1/2021.naacl-main.163.

[6] L. Ratinov, D. Roth, Design challenges and misconceptions in named entity recognition, in: S. Stevenson, X. Carreras (Eds.), Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), Association for Computational Linguistics, Boulder, Colorado, 2009, pp. 147–155. URL: https://aclanthology.org/W09-1119.

[7] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: K. Knight, A. Nenkova, O. Rambow (Eds.), Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270. URL: https://aclanthology.org/N16-1030. doi:10.18653/v1/N16-1030.

[8] T. Almeida, R. A. Jonker, R. Poudel, J. M. Silva, S. Matos, Bit. ua at medprocner: discovering medical procedures in spanish using transformer models with mcrf and augmentation, Working Notes of CLEF (2023).

[9] R. A. A. Jonker, T. Almeida, S. Matos, J. Almeida, Team BIT.UA @ BC8 SympTEMIST Track: A Two-Step Pipeline for Discovering and Normalizing Clinical Symptoms in Spanish., 2023. URL: https://doi.org/10.5281/zenodo.10103360. doi:10.5281/zenodo.10103360.

[10] R. A. A. Jonker, T. Almeida, R. Antunes, J. R. Almeida, S. Matos, Multi-head CRF classifier for biomedical multi-class Named Entity Recognition on Spanish clinical notes, Database (to appear) 2024 (2024).

[11] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, M. Krallinger, N. Loukachevitch, V. Davydova, E. Tutubalina, G. Paliouras, Overview of BioASQ 2024: The twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. Maria Di Nunzio, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), 2024.

[12] S. Lima-López, E. Farré-Maduell, J. Rodríguez-Miret, M. Rodríguez-Ortega, L. Lilli, J. Lenkowicz, G. Ceroni, J. Kossoff, A. Shah, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of MultiCardioNER task at BioASQ 2024 on Medical Speciality and Language Adaptation of Clinical NER Systems for Spanish, English and Italian, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), CLEF Working Notes, 2024.

[13] S. Lima-López, E. Farré-Maduell, J. Rodríguez-Miret, M. Krallinger, MultiCardioNER Corpus: Multilingual Adaptation of Clinical NER Systems to the Cardiology Domain, 2024. URL: https://doi.org/10.5281/zenodo.11368861. doi:10.5281/zenodo.11368861.

[14] World Health Organization, Cardiovascular diseases (cvds), 2021. URL: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1, accessed: 08-06-2024.

[15] Z. Li, S. Zhang, Y. Song, J. Park, Extrinsic factors affecting the accuracy of biomedical ner, 2023. arXiv:2305.18152.

[16] D. Demner-Fushman, W. W. Chapman, C. J. McDonald, What can natural language processing do for clinical decision support?, J. Biomed. Inform. 42 (2009) 760–772.

[17] E. French, B. T. McInnes, An overview of biomedical entity linking throughout the years, Journal of biomedical informatics 137 (2023) 104252.

[18] A. Gonzalez-Agirre, M. Marimon, A. Intxaurrondo, O. Rabal, M. Villegas, M. Krallinger, Pharmaconer: Pharmacological substances, compounds and proteins named entity recognition track, in: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, 2019, pp. 1–10.

[19] 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: CLEF (Working Notes), 2022, pp. 179–203.

[20] 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.

[21] S. Lima-López, E. Farré-Maduell, L. Gascó, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of medprocner task on medical procedure detection and entity linking at bioasq 2023, Working Notes of CLEF (2023).

[22] S. Lima-López, E. Farré-Maduell, L. Gasco-Sánchez, J. Rodríguez-Miret, M. Krallinger, Overview of symptemist at biocreative viii: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text, in: Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models, 2023.

[23] S. Vassileva, G. Grazhdanski, S. Boytcheva, I. Koychev, Fusion @ bioasq medprocner: Transformer-based approach for procedure recognition and linking in spanish clinical text, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs

of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 190–205. URL: https://ceur-ws.org/Vol-3497/paper-017.pdf.

[24] E. Zotova, A. G. Pablos, M. Cuadros, G. Rigau, VICOMTECH at medprocner 2023: Transformers-based sequence-labelling and cross-encoding for entity detection and normalisation in spanish clinical texts, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 206–218. URL: https://ceur-ws.org/Vol-3497/paper-018.pdf.

[25] G. Grazhdanski, S. Vassileva, I. Koychev, S. Boytcheva, Team Fusion@SU @ BC8 SympTEMIST track: Transformer- based Approach for Symptom Recognition and Linking, 2023. URL: https://doi.org/10.5281/zenodo.10103750. doi:10.5281/zenodo.10103750.

[26] J. Devlin, M.-W. 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, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[27] 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, 2019. arXiv:1907.11692.

[28] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[29] 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: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), 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.

[30] L. Campillos-Llanos, A. Valverde-Mateos, A. Capllonch-Carrión, A. Moreno-Sandoval, A clinical trials corpus annotated with umls© entities to enhance the access to evidence-based medicine, BMC Medical Informatics and Decision Making 21 (2021) 1–19.

[31] D. Ye, Y. Lin, P. Li, M. Sun, Packed levitated marker for entity and relation extraction, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 4904–4917. URL: https://aclanthology.org/2022.acl-long.337. doi:10.18653/v1/2022.acl-long.337.

[32] L. Luo, C.-H. Wei, P.-T. Lai, R. Leaman, Q. Chen, Z. Lu, Aioner: all-in-one scheme-based biomedical named entity recognition using deep learning, Bioinformatics 39 (2023) btad310.

[33] M. Sänger, S. Garda, X. D. Wang, L. Weber-Genzel, P. Droop, B. Fuchs, A. Akbik, U. Leser, Hunflair2 in a cross-corpus evaluation of named entity recognition and normalization tools, arXiv preprint arXiv:2402.12372 (2024).