BIT.UA at MedProcNer: Discovering Medical Procedures in Spanish Using Transformer Models with MCRF and Augmentation

Tiago Almeida[1], Richard A. A. Jonker[1], Roshan Poudel[1], Jorge M. Silva[1] and Sérgio Matos[1,*,†]

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

Abstract

This paper presents the participation of the University of Aveiro Biomedical Informatics and Techologies (BIT) group in the Medical Procedure Named Entity Recognition (MedProcNER) task at BioASQ 11, which includes Clinical Procedure Recognition, Clinical Procedure Normalization, and Clinical Procedure-based Document Indexing. We employ transformer models with Masked Conditional Random Fields for procedure recognition, utilizing a contextualized sliding window strategy to handle large documents, and data augmentation. For normalization, we utilize sentence transformers to generate embeddings for the predicted clinical procedure and the entire SNOMED CT corpus, and utilize cosine similarity for mapping. For the indexing task, we index all the SNOMED CT codes identified in the normalization phase. Our ensemble method, which combines models at the token, span, and entity level, has shown considerable success in addressing the MedProcNER task. Our system achieved the best results in terms of NER (subtask-1) with a score of 79.85%. On the other hand, in the second and third subtasks we achieved a score of 31.68% and 35.85%, respectively. Code to reproduce our submissions is available at https://github.com/ieeta-pt/MedProcNER.

Keywords

Named Entity Recognition, Spanish Clinical Procedures, Transformers, Entity Linking, Data Augmentation, SNOMED CT, Normalization, Document Indexing.

1. Introduction

Medical Named Entity Recognition (NER), a fundamental task in natural language processing, plays a pivotal role in extracting and understanding critical biomedical information from vast and varied text sources. Specifically, recognising clinical procedures in documents, a subset of biomedical NER, is of prime significance given the vast amount of valuable information encapsulated in these terminologies. Moreover, this task becomes even more challenging when dealing with languages other than English, such as Spanish, given the scarcity of annotated resources in such languages. In this context, the MEDical PROCedure Named Entity Recognition

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

✉ tiagomeloalmeida@ua.pt (T. Almeida); richard.jonker@ua.pt (R. A. A. Jonker); proshan@ua.pt (R. Poudel); jorge.miguel.ferreira.silva@ua.pt (J. M. Silva); aleixomatos@ua.pt (S. Matos)
🌐 https://t-almeida.github.io/online-cv/ (T. Almeida); https://jorgeMFS.com/ (J. M. Silva)
iD 0000-0002-4258-3350 (T. Almeida); 0000-0002-3806-6940 (R. A. A. Jonker); 0000-0001-6450-023X (R. Poudel); 0000-0002-6331-6091 (J. M. Silva); 0000-0003-1941-3983 (S. Matos)

CEUR Workshop Proceedings (CEUR-WS.org)

(MedProcNER [1]) task by BioASQ [2] is a timely and ambitious challenge.

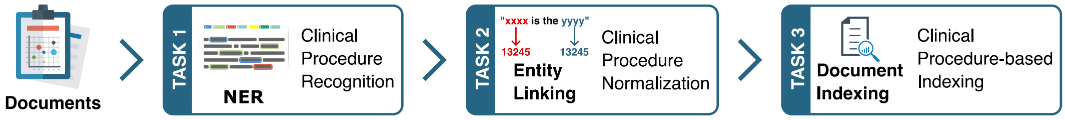
Documents
TASK 1
NER
Clinical Procedure Recognition
TASK 2
"xxxx is the yyyy"
13245
13245
Entity Linking
Clinical Procedure Normalization
TASK 3
Document Indexing
Clinical Procedure-based Indexing

Figure 1: Pipeline of the subtasks performed in MedProcNER.

The MedProcNER task presents three distinct subtasks: Clinical Procedure Recognition, Clinical Procedure Normalization, and Clinical Procedure-based Document Indexing, as depicted in Figure 1. The first task focuses on detecting clinical procedures in published reports, the second subtask aims to normalise these mentions by assigning SNOMED CT codes, and the third subtask requires indexing whole clinical case report texts using these codes. The complexity of these tasks stems from the varied representation of clinical procedures in text, the inherent ambiguity in medical language, and the vastness of the SNOMED CT ontology.

Our work focused on the methodologies proposed in Almeida et al. [3] to build a strong solution for the MedProcNER task, which focuses on clinical procedure recognition, normalization, and indexing in Spanish clinical documents. More concretely, our methodology pivots on transformer models, a contemporary class of models that have proven efficient in various NLP tasks. To surmount the limitation of transformer models related to maximum input length, we split larger documents, adding context to each side of the split to preserve crucial contextual information. Further enhancing the performance, we employ data augmentation techniques, namely Random Token Replacement, to boost the model's generalisation abilities.

This paper elucidates our approach, its merits, and its performance in the MedProcNER task. The succeeding sections detail the related work, our methodology, data handling, model training, and performance evaluation, then a comparison of our approach with existing methods, a discussion and conclusion, and finally, potential future directions.

2. Related Work

Biomedical Named Entity Recognition (NER) is a crucial task in Natural Language Processing (NLP), which aims to identify structured information from clinical texts [4, 5]. Such texts often contain complex, technical terms and show a high degree of variability. However, NER in the biomedical domain is particularly challenging due to the limited availability of annotated data. In addition, the annotation process requires a high level of expertise, and it's both time-consuming and expensive [4, 6].

Several state-of-the-art methods have been proposed to tackle these challenges. For example, the Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) incorporates part-of-speech tagging and named entity recognition for information extraction from electronic medical record clinical free-text [5]. This system effectively captures structured information from unstructured texts, improving the overall understanding of clinical narratives.

In addition to these systems, recent research has emphasized the importance of identifying biomedical entities such as genes, proteins, chemical compounds, and clinical entities. This process is often paired with a normalization step to reduce ambiguity and ensure that different terms referring to the same concept are treated as identical [3, 7, 8, 9, 10, 11, 12, 13]. This normalization process is particularly crucial for accurate entity recognition and retrieval in the biomedical field, where terminology can be highly variable and context-specific.

Transformer models, which excel at capturing complex dependencies in input data, have been widely used in NER tasks. However, these models are limited by a maximum input length, which can be problematic when dealing with lengthy clinical narratives. Researchers have suggested segmenting larger documents and attaching contextual information to each segment to mitigate this. This approach ensures the preservation of essential contextual information despite the limitation in input length [3, 14, 15, 16].

In a similar vein, research has shown the efficacy of machine learning in normalizing disease mentions [17, 18], ensemble models in identifying and normalizing chemicals [19, 20], and neural models for the joint extraction of entities and relations [21, 22, 23, 24]. These innovative techniques have significantly improved the accuracy and comprehensiveness of biomedical entity normalization.

Pappas et al. [25] has studied the effect of seven data augmentation methods in factoid question answering, focusing on the biomedical domain. They demonstrated that data augmentation could substantially enhance the performance of NER tasks, even when large pre-trained Transformers are employed [25]. This finding suggests the importance of augmenting training data in domains with limited annotated resources, such as the biomedical field.

3. Methodology

The methodology section commences with a comprehensive overview of the dataset. Subsequently, it details the methods employed for each subtask we participated in.

3.1. Dataset

The dataset consists of a collection of documents containing biomedical information, namely clinical procedures. In total, the dataset comprises 750 annotated Spanish documents, containing a total of 11065 entities. Each annotation consists of three components: the label, which is always the same, the start span, and the end span. Further, each entity is mapped to a SNOMED CT code from a list of 242228 codes.

We define the entity recognition problem as a sequence labelling problem, where (sub)-tokens need to be classified as part of the entity or not. For that purpose, we adopted the BIO tagging schema, which allows us to label each token in the documents as either the beginning of an entity (B), inside an entity (I), or outside an entity (O).

In order to be coherent with the BIO tagging schema, we had to address overlapping annotations, since the BIO schema only allows us to perform single label classification. Therefore, in cases where multiple annotations overlapped, we resolved the conflict by merging the overlapping annotations. We believe that this assumption if reasonable, since if two overlapping annotation are correct, then the concatenation of both should also be a valid annotation.

3.2. Subtask-1: Clinical Procedure Recognition

Our approach for subtask-1 follows the works of Almeida et al. [3], which consider the NER problem as a sequence classification problem. We utilize a transformer-based model with a Masked Conditional Random Field (MCRF) [26] as a classification layer and incorporate data augmentation during training. Specifically, our model comprises three key components: a transformer-based model trained in the Spanish language, an encoder layer, and a classification head. To select the most suitable pretrained transformer-based models for our problem, we rely on the Huggingface hub. The following are the candidate transformer-based models we consider:

- lcampillos/roberta-es-clinical-trials-ner: RoBERTa [27] based model fine tuned for medical NER on spanish text [28].
- PlanTL-GOB-ES/roberta-base-biomedical-clinical-es: RoBERTa based model trained on a biomedical-clinical corpus in Spanish for Masked Language Modelling (MLM) [29].
- ccarvajal/beto-prescripciones-medicas: BETO based model fine tuned for medical entity detection. BETO is a BERT model trained on a Spanish corpus [30].
- plncmm/bert-clinical-scratch-wl-es: BETO based model fine tuned for medical entity detection.

Then, as an encoder layer we experimented with a simple dense layer and a BiLSTM layer. Finally, as the classification head we used the Masked CRF layer, which corresponds to a normal Conditional Random Field layer with a mask over its transitional weights. The main idea, is that the BIO tagging schema by itself encodes information about its structure, for instance, with BIO schema a model should never predict a I (inside) tag after an O (outside) tag. Therefore, by using a mask, we can directly encode this property into the model by setting a large negative weight to the transition O to I, and hence, making it impossible for the model to take that decoding path.

Additionally, as the transformer models have limitations regarding the maximum input length (usually 512 tokens), we needed to split larger documents to fit within the input size constraints. To enhance the performance on the split documents, we keep a right and left context region that is not decoded, but it will help the model to gain context about the previous and next sequences, as depicted in Figure 2. For instance, if we set the context size to 64, we applied a left context of 64 tokens, the actual content of 384 tokens, and a right context of 64 tokens. This approach proves beneficial for capturing contextual dependencies and improving the accuracy of the model. This allows us to overcome some limitations of the transformer models and ensure comprehensive coverage of the document content for clinical procedure recognition.

Similar to our previous work Almeida et al. [3], data augmentation techniques were employed to enhance the model's ability to generalize and improve overall performance. By incorporating augmentation techniques, we aim to increase the robustness and generalization capabilities of our model. The two augmentation techniques utilized in our methodology are as follows:

- **Random Token Replacement:** This technique involves randomly replacing tokens in the input sequence by any other valid token in the vocabulary [3, 31].

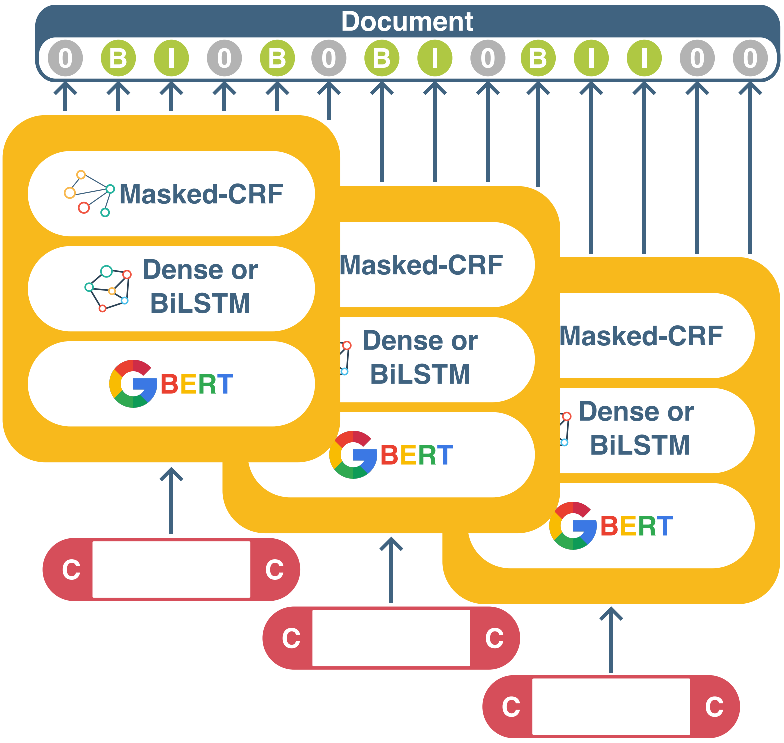

Figure 2: Diagram depicting our MedProcNER subtask-1 architecture. We fed different contexts to the model generating BIO labels of the documents.

- **Random Token Replacement with Unknown:** This technique can be seen as a special case of the previous one, where we replace randomly selected tokens in the input sequence by the [UNK] special token.

The intuition behind data augmentation techniques, such as unknown replacement, is to encourage the model to rely on contextual information when making predictions. By replacing certain tokens with an unknown (UNK) token, we create situations where the model must consider the surrounding tokens to determine whether the replaced token should be part of an entity or not. For example, let's consider the entity "tomografía computarizada abdominal" If we perform data augmentation with unknown replacement, we might obtain "tomografía UNK abdominal." In this case, the model can only predict that the UNK token should be part of the entity if it properly utilizes the context provided by the surrounding tokens. This technique effectively forces the model to learn to look beyond the representation of an individual token and consider its neighboring tokens to make accurate predictions. Lastly, it is important that we also balance this method by also replacing some non-entity tokens with UNK token, so that the model does not learn that the UNK token would always correspond to an entity.

In order to extract the entities from the predicted BIO tags, we build a simpler decoder component responsible for that. The decoder considers both the current and previous tokens to determine the appropriate boundaries for each annotation. The decoding process follows the following rules:

- If the current token is labeled as “B” (indicating the beginning of an annotation), a new annotation is started.
- If the previous token indicates no annotation (“O”), but the current token is labeled as “I” (indicating inside an annotation), it is assumed to be the beginning of the annotation.
- If the previous token is labeled as “B” or “I” (indicating the beginning or inside an annotation, respectively), and the current token is also labeled as “I”, the span of the annotation is extended.
- If the current token is labeled as “O” (indicating no annotation), and the previous token is labeled as “B” or “I”, it signifies the end of the annotation.

Lastly, in order to grasp the knowledge of several models, we proposed to use ensemble techniques at different levels to enhance the performance and reliability of our models. Here is an overview of the ensemble methods used:

- **Span Level Ensemble:** After decoding the BIO tags into spans, we perform a span-level ensemble. This ensemble combines all the spans predicted by the individual models. By merging the spans, we aim to capture all relevant annotations by the models, however, this may also increase the number of incorrect annotations.
- **Entity Level Ensemble:** At the entity level, we focus on each annotated entity. We perform an exact matching process to compare the entities predicted by different models. A majority voting scheme is used to select the which entities are used to represent the document.

3.3. Subtask-2: Clinical Procedure Normalization

In the next phase of the challenge, we need to normalize the predicted clinical procedures to the SNOMED CT terminology. To achieve this, we employed Sentence Transformers [32, 33] and tested various pre-trained models to create embeddings for the annotations labeled in the first phase, as well as the entire SNOMED CT corpus. It is important to note that no fine-tuning was performed on these models. We tested both Spanish models and English models by translating the corpus to English. An automatic translation of the codes and predicted procedures was performed using the following open-source model [34].

Using the embeddings generated by the pre-trained models, we employed cosine similarity to find the best matching term from the SNOMED CT terminology for each clinical procedure. The cosine similarity metric allowed us to measure the similarity between the vectorized embeddings of the annotations and the terms in the SNOMED CT corpus. The cosine similarity between two vectors is calculated using the dot product of the vectors and their respective magnitudes. The equation for cosine similarity is as follows:

$$\text{Cosine Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \times \|\mathbf{B}\|}.$$

3.4. Subtask-3: Clinical Procedure-based Document Indexing

In this subtask, we are tasked with automatically assigning clinical procedure codes to full clinical case report texts in order to enable effective indexing of the documents. Due to time constraints, we utilize all matched SNOMED CT codes labeled from the previous stage to index the document.

4. Results and Discussion

4.1. Validation Results

Table 1 shows the preliminary tests of using data augmentation for entity recognition. As observable, the UNK augmentation provided the best results for both models. Based on these observations we assume that data augmentation is beneficial, or in the worse case, will not negatively impact the performance, whilst still increasing the models generalization capabilities. In the remainder of the paper these augmentation techniques will be used.

Table 1
Comparison of microF1 scores for different data augmentation techniques on PlanTL and lcampilos models.

Augmentation	microF1	
	PlanTL	lcampilos
None	0.7638	0.7723
UNK	**0.7666**	**0.7731**
Random	0.7639	0.7710

Following the previous experiments, work was focused in finding optimal model and architecture combinations. A large number of models were tested during the first phase due to the large number of possible model combinations available. In the initial tests of combinations, the main focus was selecting the best model architecture and the best context size to use (Figure 3). From the initial results, we conclude that the 2 best models used were PlanTL and lcampillos. It was noted that a context size of 128 performed significantly worse. It is theorized that by reducing the context size of the models, more documents were able to fit into a single context window which would have better result than documents that were split. The context size of 8 and 16 performed the best.

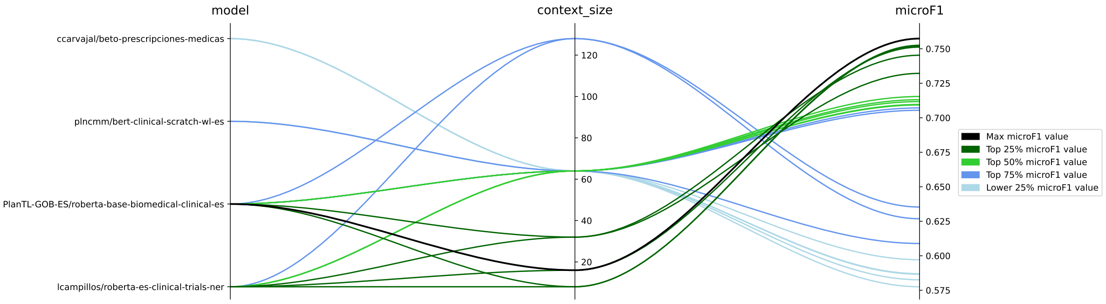

Figure 3: Parallel plot showing impact of different models and context size in microF1 score.

Looking at some further preliminary tests (Figure 4), it was noted that the BiLSTM architecture type performed better than the dense architecture. We hypothesise that this difference is due to the BiLSTM capability of considering long-term dependencies, while a simple dense layer encodes the transformer representation independently. It also seems that using UNK augmentation performed better than just using simple random token replacement. It is possible that the model's performance is negatively impacted by random replacement, as it may inadvertently alter token embeddings in an incorrect manner. In contrast, utilizing the UNK token, which is never used during inference, ensures that any changes made to its representation will not hinder the model's performance. This is also evidenced by the fact that with random replacement, the models with lower replacement percentage achieved higher results, suggesting that the random replacement is hurting the performance. On the other hand, with the UNK replacement it seems that using higher percentages of replacement is beneficial, further evidenced the benefices of using that data augmentation technique.

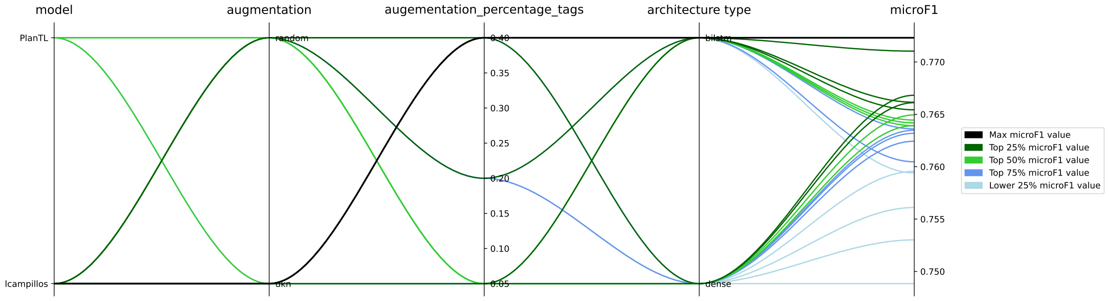

Figure 4: Impact of parameter variation on final metric. Each line represents an individual run of the model, depicting the influence of different parameters on the final metric.

Regarding our ensemble methods for entity recognition, we also performed some validation experiments that are summarized in Table 2. We show the best, worst and average microF1, precision and recall values that were obtained from training 21 models. These values are compared against both ensemble strategies, containing all models. Overall it is observable that the entity level ensemble was consistently better, achieving higher scores than the best models, similar to Almeida et al. [3]. This shows that this ensemble method can consistently combine the knowledge from different models improving the final score.

Table 2
Performance comparison of ensemble techniques at different levels in comparison with individual models in terms of microF1, precision and recall. 21 models were used in the ensemble.

Model	microF1	Precision	Recall
Best	0.7778	0.7794	**0.7768**
Worst	0.7537	0.7573	0.7503
Average	0.7673 ± 0.0057	0.7683 ± 0.0068	0.7665 ± 0.0078
Ensemble - Entity level	**0.7808**	**0.7978**	0.7645
Ensemble - Span level	0.7711	0.7716	0.7706

Focusing now in subtask-2 (Clinical Procedure Normalization), Table 3 summarizes the performance of various models, both in Spanish and English. It is important to note that the accuracies reported in this table are relatively low, with the best model only achieving a maximum of 22% accuracy. These low accuracies can largely be attributed to the disparities in the embedding space and the lack of context in the SNOMED-CT corpus. Models were not pre-trained for this task, which would have otherwise aligned the embedding spaces more closely and potentially led to more accurate results.

A notable point is that the translation of documents from Spanish to English did not result in a significant improvement in the model performance. It was originally hypothesized that translating the documents to English could potentially improve the performance, considering the availability of more advanced and powerful models trained on English data. However, as the results indicate, this was not the case, which led to the decision to keep the documents in their original Spanish language to avoid any loss of information that might occur during the translation process.

Interestingly, the performance of English models displayed a consistent trend, while the Spanish models showed a much wider range of accuracies. This might indicate that the quality and the performance of models trained on Spanish data can vary more significantly, suggesting room for improvement in the development and training of Spanish models.

The model selected for the submission of the work is the best-performing Spanish model, "LaBSE-sentence-embeddings", achieving an accuracy of 22.0%. Despite its modest performance, the selection of this model aligns with our goal of maintaining the documents in their original language while attempting to leverage the benefits of pre-trained models for the task at hand.

Future work could involve fine-tuning the models for the specific task of normalizing clinical procedures to SNOMED CT terminology, potentially leading to significantly improved results. In addition, further research into developing and refining models for the Spanish language could also provide more advanced tools for tackling this and similar tasks.

Table 3
Model Performance on Validation Set: Comparison of Spanish and English Models for subtask-2. (*) represents models from sentence transformer package.

Model (Spanish)	**Accuracy**
paraphrase-xlm-r-multilingual-v1*	19.0
sentence_similarity_spanish_es*	18.0
LaBSE-sentence-embeddings	**22.0**
sn-xlm-roberta-base-snli-mnli-anli-xnli	16.8
distiluse-base-multilingual-cased-v2-finetuned-stsb_multi_mt-es	17.6
S-Biomed-Roberta-snli-multinli-stsb	18.4
roberta-base-biomedical-clinical-es	15.0
Model (English)	
BioSimCSE-BioLinkBERT-BASE*	19.6
S-Biomed-Roberta-snli-multinli-stsb*	**22.1**
S-BioBert-snli-multinli-stsb*	21.8
LaBSE-sentence-embeddings	19.8

4.2. Official Results

Table 4 outlines the official results of our five systems (system-0 to system-4) in comparison with the best results achieved in the competition. In subtask-1 (NER), our systems performed commendably. Notably, system-4, an ensemble of all plantl and lcampillos models, achieved the highest F1 score of 79.85, which was also the best score across the competition. In subtasks-2 and 3, however, our systems fell short. Our best score for subtask-2 (Entity Linking) was 31.68 by system-4, considerably below the competition's best score of 57.07. For subtask-3 (Indexing), our top performer was system-3 with an F1 score of 35.98, significantly lower than the competition's best of 62.42.

Table 4
Comparison of System Performances on Subtasks with Best Competition Results.

System	NER System			F1		
	Models used	Aug.	Data	Subtask-1	Subtask-2	Subtask-3
system-0	5 x lcampillos-dense	UNK	no val	79.46	31.53	35.67
system-1	5 x lcampillos-dense	UNK	full	79.24	31.32	35.42
system-2	30 x lcampillos-bilstm	mixed	mixed	78.81	31.10	35.37
system-3	20 x plantl-mixed	mixed	no val	79.23	31.66	35.98
system-4	All plantl and lcampilos (91)	mixed	mixed	**79.85**	31.68	35.85
Best				**79.85**	**57.07**	**62.42**

5. Conclusion

This study sought to analyze and compare the effectiveness of different models for Named Entity Recognition (NER), Entity Linking, and Indexing in Spanish in the context of the MedProcNER shared task. For the first subtask (NER), there was little variation in performance among the various models tested, with our system achieving the best result for the task. It was noted that smaller context size improves the performance of the system, and that ensembling a variety of models is beneficial.

In the second and third subtasks, our score was almost half of the best performing team. Given the lack of fine tuning for the task, this result is expected. It is also noted that the performance from the first subtask is directly carried over to the other tasks, so an improvement in subtask-1 will directly carry over.

Interestingly, translating text to English before testing did not significantly improve the performance of the models. This outcome prompted us to maintain the documents in their original Spanish language to avoid any potential loss of information during the translation process. Although English models are generally more powerful than Spanish ones, in this case, we found no substantial disparity in performance between the two language models.

Acknowledgments

This work was partially supported by national funds through the Foundation for Science and Technology (FCT) in the context of the projects DSAIPA/AI/0088/2020 and UIDB/00127/2020. Tiago Almeida is funded by FCT under the grant 2020.05784.BD. Jorge Miguel Silva has received funding from the EC under grant agreement 101081813, Genomic Data Infrastructure.

References

[1] 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, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.

[2] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, L. Gasco, M. Krallinger, G. Paliouras, Overview of BioASQ 2023: The eleventh BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), 2023.

[3] 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).

[4] Z. Li, S. Zhang, Y. Song, J. Park, Extrinsic factors affecting the accuracy of biomedical NER, arXiv preprint arXiv:2305.18152 (2023).

[5] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, C. G. Chute, Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, Journal of the American Medical Informatics Association 17 (2010) 507–513.

[6] D. Demner-Fushman, W. W. Chapman, C. J. McDonald, What can natural language processing do for clinical decision support?, Journal of biomedical informatics 42 (2009) 760–772.

[7] A. M. Cohen, W. R. Hersh, A survey of current work in biomedical text mining, Briefings in bioinformatics 6 (2005) 57–71.

[8] S. Sarawagi, et al., Information extraction, Foundations and Trends® in Databases 1 (2008) 261–377.

[9] A. S. Yeh, L. Hirschman, A. A. Morgan, Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup, Bioinformatics 19 (2003) i331–i339.

[10] M.-S. Huang, P.-T. Lai, P.-Y. Lin, Y.-T. You, R. T.-H. Tsai, W.-L. Hsu, Biomedical named entity recognition and linking datasets: survey and our recent development, Briefings in Bioinformatics 21 (2020) 2219–2238.

[11] Ö. Uzuner, B. R. South, S. Shen, S. L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, Journal of the American Medical Informatics Association 18 (2011) 552–556.

[12] S. Henry, K. Buchan, M. Filannino, A. Stubbs, O. Uzuner, 2018 n2c2 shared task on adverse

drug events and medication extraction in electronic health records, Journal of the American Medical Informatics Association 27 (2020) 3–12.

[13] A. J. Jimeno-Yepes, B. T. McInnes, A. R. Aronson, Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation, BMC bioinformatics 12 (2011) 1–14.

[14] X. Dai, Recognising biomedical names: Challenges and solutions, arXiv:2106.12230 (2021).

[15] K. Chowdhary, K. Chowdhary, Natural language processing, Fundamentals of artificial intelligence (2020) 603–649.

[16] 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.

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

[18] R. I. Doğan, R. Leaman, Z. Lu, NCBI disease corpus: a resource for disease name recognition and concept normalization, Journal of biomedical informatics 47 (2014) 1–10.

[19] R. Leaman, C.-H. Wei, Z. Lu, tmChem: a high performance approach for chemical named entity recognition and normalization, Journal of cheminformatics 7 (2015) 1–10.

[20] R. Leaman, R. Khare, Z. Lu, Challenges in clinical natural language processing for automated disorder normalization, Journal of biomedical informatics 57 (2015) 28–37.

[21] M. Miwa, M. Bansal, End-to-end relation extraction using LSTMs on sequences and tree structures, arXiv preprint arXiv:1601.00770 (2016).

[22] G. Bekoulis, J. Deleu, T. Demeester, C. Develder, Joint entity recognition and relation extraction as a multi-head selection problem, Expert Systems with Applications 114 (2018) 34–45.

[23] S. Zhao, T. Liu, S. Zhao, F. Wang, A neural multi-task learning framework to jointly model medical named entity recognition and normalization, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 817–824.

[24] 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.

[25] D. Pappas, P. Malakasiotis, I. Androutsopoulos, Data augmentation for biomedical factoid question answering, arXiv:2204.04711 (2022).

[26] T. Wei, J. Qi, S. He, S. Sun, Masked conditional random fields for sequence labeling, in: 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.

[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, arXiv preprint arXiv:1907.11692 (2019).

[28] 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.

[29] 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, 2021.

[30] 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.

[31] A. Erdengasileng, K. Li, Q. Han, S. Tian, J. Wang, T. Hu, J. Zhang, A BERT-based hybrid system for chemical identification and indexing in full-text articles, bioRxiv (2021). URL: https://www.biorxiv.org/content/early/2021/10/28/2021.10.27.466183. doi:10.1101/2021.10.27.466183.

[32] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019.

[33] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2020.

[34] J. Tiedemann, The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT, in: Proceedings of the Fifth Conference on Machine Translation, Association for Computational Linguistics, Online, 2020, pp. 1174–1182. URL: https://www.aclweb.org/anthology/2020.wmt-1.139.