The READ-BioMed Team in LivingNER Task 1 (2022): Adaptation of an English Annotation System to Spanish

Antonio Jimeno Yepes[1,2], Karin Verspoor[1,2]

[1]School of Computing Technologies, RMIT University, Melbourne, Australia
[2]School of Computing and Information systems, The University of Melbourne, Melbourne, Australia

Abstract

We describe the work of the READ-BioMed team for the preparation of a submission to the LivingNER Species Named Entity Recognition (NER) Task (Task 1) in 2022. We had previously developed a system for named entity recognition for identifying biomedical concepts in MEDLINE citations written in the English language. We adapted this system to process the challenge data, to process reports written in the Spanish language. We show that minimal adaptation of our methodology was required to perform named entity recognition in the Spanish language, given the availability of pre-trained language models for Spanish, in conjunction with the LivingNER training data.

Keywords

Species Named Entity Recognition, Transformer Language Model, Multilingual Adaptation

1. Introduction

In this paper, we describe the READ-BioMed (**R**eading, **E**xtraction, and **A**nnotation of **D**ocuments in **BioMed**icine) approach to the 2022 LivingNER Task 1. The documents in the LivingNER Task 1 are Spanish-language medical reports and the task involves the annotation of HUMAN and other biological SPECIES entities mentioned in the texts.

The READ-BioMed team has extensive experience in natural language processing for the biomedical domain, specifically for concept/named entity recognition [1] and relation extraction [2, 3]. Most of our previous work has focused on English language texts, including our prior work addressing annotation of biological pathogens [4]. The LivingNER challenge provided an opportunity to explore the adaptation of our methods to clinical texts in the Spanish language.

Our approach was to adapt a system previously developed for annotation of MEDLINE citations in the English language. That system relies on a pre-trained transformer based language model, which was fine-tuned to the context for biomedical concept recognition for the LitCOIN challenge earlier this year (https://ncats.nih.gov/funding/challenges/litcoin), where we ranked in the top 5 submissions (https://ncats.nih.gov/funding/challenges/litcoin/winners). We did not use any terminological resources for our submission, solely relying on machine learning

IberLEF 2022, September 2022, A Coruña, Spain.
antonio.jose.jimeno.yepes@rmit.edu.au (A. Jimeno Yepes); karin.verspoor@rmit.edu.au (K. Verspoor)
0000-0002-6581-094X (A. Jimeno Yepes); 0000-0002-8661-1544 (K. Verspoor)

methods applied to the training data to build a model. Evidence of the utility of contextualised word embeddings for Spanish-language clinical NER was available from previous work [5].

2. Methods

In this section, we describe our approach for the task 1 of the LivingNER challenge. This description follows the steps that we followed to prepare our submission, including the pre-processing and post-processing of the data and training and annotation steps. We mention which modifications were required to adapt our previously existing system for biomedical named entity recognition to the context of the challenge.

2.1. Data

We used the data provided by the challenge organizers [6, 7]. The provided training data consisted of 1000 documents annotated with HUMAN and SPECIES entity labels. An additional 500 documents constituted a validation set. The testing set contains over 13k documents.

2.2. Conversion of Text to BIO format

The training and validation documents were provided as tab separated values (TSV) files. We attempt to use pre-processing tools, mostly provided by the brat system [8], which we previously used for MEDLINE citations, but sentence splitting and alignment with the original text was a problem in this context; sentences in the training and validation sets would start with spaces in some cases. Instead of further investigating this problem and since we had to build a processor for the TSV files, we redeveloped the pre-processing scripts to deal with the LivingNER task 1 documents directly.

To make use of our existing system, we needed to adapt the data to conform to the BIO labeling schema. This was achieved using a tokeniser built with regular expressions. In the BIO labeling schema (*aka* IOB [9]), tokens not belonging to any entity are labeled as O, the B label prefix is used to select where the entity starts and I is used for the following tokens belonging to the same entity.

Token boundaries were identified using the following set of characters: {"' -.[]()‚/%:}. The tokens and the start and end offset were used in the BIO conversion. Spaces were not considered as a token and a separation was added between sentences. Sentence boundaries were identified using the "." character and some rules for end of sentence. Using the entities in the training and validation sets, we could identify cases in which entities might have been incorrectly assigned to two different sentences, e.g. *E. coli* would have been split into *E*, . and *coli* and each token assigned to a different sentence.

2.3. Training

In our previous work in biomedical English named entity recognition as part of the LitCoin challenge (https://bitgrit.net/competition/13#private-leaderboard), we evaluated pre-trained language models such as BioBERT [10], SciBERT [11] and PubMedBERT [12]. We utilised the

NERDA (https://github.com/ebanalyse/NERDA) framework for named entity recognition to support fine-tuning of language models to the specific task. The annotation approach consists on using a BERT based system followed by a fully connected layer with as many outputs as labels need to be predicted. Our motivation was to be able to make changes and adapt the software according to [13], which sets a loss to control capacity of the BERT system that might help with overfitting with data sets with a small number of annotations. Since the challenge data was in Spanish, we identified a pre-trained language model for Spanish biomedical data [14] to use in our system. More specifically, we have used the model *PlanTL-GOB-ES/roberta-base-biomedical-es* available from Huggingface. This pre-trained is based on a RoBERTa [15]. It has been trained trained on a biomedical-clinical corpus in Spanish collected from several sources.

We used the training data set provided by the organisers (1000 documents) for learning the model and the validation set (500 documents) was used to control the training process. The BIO labels that our system was trained for included the O label for out entity tokens and the B-HUMAN and I-HUMAN for HUMAN related tokens and the B-SPECIES and I-SPECIES for the species ones, representing a total of 5 token labels.

We trained the models using the SPARTAN system (https://dashboard.hpc.unimelb.edu.au). The models were fine tuned using a P100 NVIDIA GPU and it took a bit over 2 hours to fine tune for 10 epochs.

2.4. Annotation

To process the test data set, we followed a similar process to the one used for the training and validation sets. We identified that this process was generating many sentence identification errors. To solve this, we used the sentenciser in spacy [16] as confirmation of sentence boundaries.

Results from the training step showed that there are errors made by the system that required post-filtering. After correcting for errors due to incorrect sentence splitting, we identified that terms such as *animal*, *huésped* and *huéspedes* were sometimes incorrectly annotated as HUMAN. We prepared a filtering step to remove those annotations.

Annotation of the complete testing set (over 13k documents) provided by the organisers took a bit over 3 hours to annotate using the model trained as explained in the previous section. The annotation relied as well on the SPARTAN system P100 GPUs as during the training step. We used a version of the testing set tokenised and sentencised similar to the processing of the training data and detailed in the next section.

2.5. Conversion of the Annotated Text to the Submission Format

The submission format follows the same format as the training and validations sets, so from the BIO annotated files, we extracted the annotated entities and generated the TSV submission file. From the output of the annotation process, for each document, we have the tokenised sentences and the BIO annotated tokens. We aligned the tokenised sentences and the BIO annotated tokens and from the B and I annotations for each entity type, the annotation spans are decided and the output annotation file is produced. Conversion of the annotations on the testing set and the generation of the submission format would take less than a few seconds.

3. Results

The hyperparameters used for the fine tuning of the pre-trained language model include a maximum length in tokens of the sentences (266), the learning rate (lr), and the number of epochs used to fine tune the model. We realised that by using the learning rate set to 0.0001, the loss would increase for both the training and validation sets after two epochs and decided to use a lower learning rate of 0.00001, which allowed training the system for longer, which we set to 10 epochs (which we used in the LitCoin challenge).

Results in Table 1 shows the performance in predicting the B and I label prefixes for each configuration, over the validation set. We selected the setup with learning rate of 0.00001 and 10 epochs for our submission, which seemed to have an overall slightly better performance.

Method	Token	Precision	Recall	F1
lr=0.0001 epochs=2	B-HUMAN	0.9504	0.9739	0.9620
	I-HUMAN	0.9141	0.8779	0.8956
	B-SPECIES	0.9625	0.9500	0.9562
	I-SPECIES	0.9364	0.9202	0.9282
lr=0.00001 epochs=10	B-HUMAN	0.9475	0.9806	0.9638
	I-HUMAN	0.8900	0.8812	0.8856
	B-SPECIES	0.9677	0.9569	0.9623
	I-SPECIES	0.9392	0.9347	0.9369

Table 1
Evaluation of the fine tuned models on the validation data set for B and I label prefixes for the HUMAN and SPECIES entity types.

Table 2 shows the results of our submission over the test set, as per the official challenge evaluation. The results are compared to the mean results of all participants. We observe that results for the HUMAN entity type are higher than the results for the SPECIES entity type. As well, our results are above the mean results for all the evaluation measures (precision (P), recall (R) and F1-measure (F1)).

	LivingNER NER			NER only SPECIES			NER only HUMAN		
	P	R	F1	P	R	F1	P	R	F1
READ-BioMed	0.9540	0.9411	0.9475	0.9399	0.9195	0.9296	0.9736	0.9702	0.9719
MEAN	0.8763	0.8077	0.8239	0.8112	0.7579	0.7781	0.9312	0.8750	0.8849
STD	0.1542	0.2465	0.2371	0.2490	0.2565	0.2510	0.1156	0.2388	0.2230

Table 2
Official results. The READ-BioMed submission, run *run1-roberta-NERDA*, is compared to the mean and standard deviation of the participants in task 1.

4. Discussion

The results in tables 1 and 2 indicate that the performance for the HUMAN entity type is higher than for the SPECIES entity type. There are several possible explanations for this. One possible reason is that there is a larger number of different terms that are relevant for SPECIES, and that therefore generalization is more challenging [17]. As shown in Table 5, both the training and validation datasets have more instances of SPECIES annotations, relative to the instances of HUMANS. Table 3 shows statistics about the number of unique tokens and terms and table 4 shows the most frequent tokens per entity type.

Set	Annotation	Tokens	Terms
training	HUMAN	461	494
training	SPECIES	1,936	2,508
validation	HUMAN	302	300
validation	SPECIES	1,147	1,404

Table 3
Unique number of tokens and terms per entity type in the training and validation sets.

Training				Validation			
Human		Species		Human		Species	
Frequency	Token	Frequency	Token	Frequency	Token	Frequency	Token
3,204	paciente	456	-	1,490	paciente	285	-
342	varón	452	virus	169	varón	189	VIH
251	personales	402	VIH	116	personales	156	virus
217	mujer	391	.	104	mujer	130	.
215	pacientes	305	de	99	madre	117	de
161	familiares	223	spp	80	familiares	106	2
150	madre	160	2	77	pacientes	89	spp
133	familia	127	b	67	Paciente	89	SARS
131	de	126	y	64	de	89	CoV
130	familiar	125	CMV	54	niño	78	cannabis

Table 4
Token count per entity type in the training and validations sets.

Another one is that the species terms are more complex. The data in Table 5 also shows that the number of entities with multiple tokens is lower in the HUMAN entity type than in the SPECIES one, indicated by the proportion of B versus I labels for each entity type. In addition, the performance on the I label prefixes is lower compared to the B label prefixes as can be seen in Table 1.

Set	Annotation	Count	Annotation	Count
training	B-HUMAN	6,712	I-HUMAN	641
training	B-SPECIES	8,889	I-SPECIES	6,052
validation	B-HUMAN	3,147	I-HUMAN	303
validation	B-SPECIES	3,758	I-SPECIES	2,495

Table 5
B and I annotation for the HUMAN And SPECIES entity types on the training and validation sets using the model used for our submission.

5. Conclusions and Future Work

For the READ-BioMed challenge submission to LivingNER, we adapted an existing approach for the annotation of biomedical entities in English to the Spanish language. We identify that while the training and prediction steps remain mostly unchanged, with the biggest change the use of a Spanish pre-trained language model, special attention is needed in the pre-processing and post-processing steps. Overall, few changes were required to adapt the code from English to Spanish.

As future work, it might be interesting to evaluate the adaptation of other approaches prepared for the English language might be adapted to Spanish, or other languages, using pre-trained transformer based language models. It might also be interesting to explore the performance of language models pre-trained in one or multiple languages in several scenarios of multi-lingual tasks.

Acknowledgments

We thank the LivingNER organisers for preparing the task and the datasets required for the challenge. We additionally thank Carrino et al. [14] for preparing the pre-trained language model used in our submission.

This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

References

[1] C. Funk, W. Baumgartner, B. Garcia, C. Roeder, M. Bada, K. B. Cohen, L. E. Hunter, K. Verspoor, Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters, BMC Bioinformatics 15 (2014) 1–29.

[2] H. Liu, L. Hunter, V. Kešelj, K. Verspoor, Approximate subgraph matching-based literature mining for biomedical events and relations, PloS One 8 (2013) e60954.

[3] D. Q. Nguyen, K. Verspoor, End-to-end neural relation extraction using deep biaffine attention, in: European conference on information retrieval, Springer, 2019, pp. 729–738.

[4] A. Jimeno Yepes, A. Albahem, K. Verspoor, Using discourse structure to differentiate focus entities from background entities in scientific literature, in: Proceedings of the The

19th Annual Workshop of the Australasian Language Technology Association, 2021, pp. 174–178.

[5] L. Akhtyamova, P. Martínez, K. Verspoor, J. Cardiff, Testing contextualized word embeddings to improve ner in spanish clinical case narratives, IEEE Access 8 (2020) 164717–164726.

[6] A. Miranda-Escalada, E. Farré-Maduell, G. González Gacio, M. Krallinger, LivingNER corpus: Named entity recognition, normalization & classification of species, pathogens and food, 2022. URL: https://doi.org/10.5281/zenodo.6642852. doi:10.5281/zenodo.6642852, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

[7] A. Miranda-Escalada, E. Farré-Maduell, S. Lima-López, D. Estrada, L. Gascó, M. Krallinger, Mention detection, normalization & classification of species, pathogens, humans and food in clinical documents: Overview of livingner shared task and resources, Procesamiento del Lenguaje Natural (2022).

[8] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, J. Tsujii, Brat: a web-based tool for nlp-assisted text annotation, in: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 102–107.

[9] L. Ramshaw, M. Marcus, Text chunking using transformation-based learning, in: Third Workshop on Very Large Corpora, 1995. URL: https://aclanthology.org/W95-0107.

[10] 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.

[11] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, arXiv preprint arXiv:1903.10676 (2019).

[12] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23.

[13] A. Jimeno Yepes, Hyperplane bounds for neural feature mappings, arXiv preprint arXiv:2201.05799 (2022).

[14] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, M. Villegas, Biomedical and clinical language models for Spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, arXiv preprint arXiv:2109.03570 (2021).

[15] 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).

[16] M. Honnibal, I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017.

[17] A. Elangovan, J. He, K. Verspoor, Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 1325–1335. URL: https://aclanthology.org/2021.eacl-main.113. doi:10.18653/v1/2021.eacl-main.113.