

# Named Entity Recognition For Humans and Species With Domain-Specific and Domain-Adapted Transformer Models

Alejandro Vaca-Serrano<sup>1</sup>

<sup>1</sup>*Instituto de Ingeniería del Conocimiento, Francisco Tomás y Valiente st., 11 EPS, B Building, 5th floor UAM Cantoblanco, 28049 Madrid, Spain*

## Abstract

This work presents different solutions to the tasks proposed at the LivingNER challenge, as part of the Iberlef 2022 Conference, with a special focus on the NER task. For that, a general domain large model was adapted to the biomedical domain, showing that this process improves the posterior fine-tuning on a majority of tasks. However, although achieving similar results, it is not able to outperform two base size models specific of the biomedical domain. A careful analysis of the reason for this gap in performance is carried out, showing that the tokenizers' vocabulary has a great impact on the aggregation of predictions both at the word level and the word group level. This highlights the effectiveness of using domain specific models for tasks very specific to a concrete linguistic domain. Official test results show a very good performance on the NER task, where all the submissions made are clearly above the average results. However, results for tasks 2 and 3 are very poor, which indicates that a deeper understanding of the underlying nature of those tasks is needed.

## Keywords

NER, Entities normalization, Impact Analysis, Clinical Reports Analysis, Biomedical Domain, Domain Adaptation of Language Models

## 1. Introduction

In this work we present different approaches for the different subtasks of the LivingNER challenge [1], as part of the Iberlef conference. LivingNER aims to improve the existing automatic systems to detect, classify and analyze the impact of living entities. Some identified areas possibly impacted by such systems are medicine, biology, ecology/biodiversity, nutrition and agriculture.

The work focuses mainly in the first task, that is the NER task, while less effort was used for the second and third tasks. For the NER task, exhaustive experiments were carried out with different Transformer models in Spanish, thus expanding the existing benchmarks of language models in Spanish, such as those in [2], [3] or [4].

Furthermore, a detailed analysis on the effect of the domain of the corpus each tokenizer is trained on is done. With this analysis, this work tries to contribute to the existing knowledge regarding the effect of the domain over language models, specifically in Spanish. Then, different ways of mixing the predictions of those models are explored.

---


*IberLEF 2022, September 2022, A Coruña, Spain.*

✉ [alejandro\\_vaca0@hotmail.com](mailto:alejandro_vaca0@hotmail.com) (A. Vaca-Serrano)

🌐 <https://www.linkedin.com/in/alejandro-vaca-serrano/> (A. Vaca-Serrano)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

In section 2, other similar challenges are explored, and openly available language models in Spanish relevant to these tasks are reviewed. Then, in section 3, the different tasks of LivingNER are briefly explained. The whole system for the three tasks is described in section 4.

## 2. Related Work

### 2.1. NER for the Health Sector in Spanish

The Biomedical Text Unit at BSC (Barcelona Supercomputing Center) has released several corpora related to the biomedical/health domain in the previous years. One such example is Meddocan [5] [6], an anonymization track in which systems have to identify several entities such as NOMBRE\_SUJETO\_ASISTENCIA or CALLE. This serves for removing this important information from clinical reports, so that they could be better anonymized. The project that obtained the highest results on that challenge for the NER task was [7]. In [7], it is shown that very promising results in NER tasks can be obtained by using pre-trained Embeddings such as FastText [8] or FLAIR [9]. Then, BiLSTMs [10] were used to mix those Embeddings. The task is approached as a token classification task, where for each input token an output label must be predicted.

Other challenge of similar nature as LivingNER is Meddoprof [11] [12], a task consisting on classifying and normalizing professions and occupations in medical texts. The best results in NER were obtained again by the same team as Meddocan [12] [6], NLNDE (Neither-Language-Nor-Domain-Experts)[7]. In [13], the NLNDE team use the multilingual XLM-R [14] model in several settings. Some interesting experiments such as domain adaptation of the language model and language-adaptive pre-training were carried out. Note that this competition took place on 2021 Spring, when not many Spanish language models had been developed, a possible reason for such language adaptation. Other boosting strategies such as transfer learning and strategic datasplits were tried out. These methods account, according to [13], for an improvement of 5.3 F1 points compared to the fine-tuned XLM-R model.

### 2.2. Transformer Models in Spanish

The first language model released in Spanish was BETO [2], a Spanish BERT [15]. Then, in the context of the MarIA project [3], Spanish RoBERTa [16] and GPT-2 [17] models were released, both base and large. The most suitable architecture for the nature of LivingNER NLU tasks is the encoder-only, as full attention is generally better than masked attention in those cases. For this reason, RoBERTa-base and large from MarIA were selected, named MarIA-base and MarIA-large along the paper. Additionally, BERTIN model was released this year [18]. It is also a version of RoBERTa in Spanish, trained with less resources than [3] but with novel techniques.

Additionally, there are two domain-specific models in Spanish. Both of them use the RoBERTa-base architecture. They are trained using corpora from the biomedical and biomedical-clinical domains respectively [19], and will be called BioMedical and BioClinical along the paper.

By looking at different models' comparisons in [4], it was decided that MarIA-large, MarIA-base, BioMedical and BioClinical would be used for the different subtasks of LivingNER. Additionally, a domain adaptation for MarIA-large to the biomedical domain is carried out, which is

explained later.

However, as a matter of comparison, in a first step, BETO [2] and BERTIN [18] were also trained. When their scores were checked, it was decided to not use them for the next step. This experiment shows that the model comparisons above are consistent with the LivingNER task 1 results.

### **3. Challenge Description**

In this section the challenge is explained, by defining the different tasks that form it.

#### **3.1. Task 1**

In the first task, systems must identify humans and species in clinical texts. Systems are expected to provide the spans, together with the offsets of the whole entities in those texts. This is a typical NER task with two entities (plus the null entity, which is the entity assigned to those tokens not being any other entity).

#### **3.2. Task 2**

In task 2 systems have to classify the entities predicted in the first task, according to their NCBI taxonomy ids (National Center for Biotechnology Information) [20]. There are many different labels for this task, and patterns are very difficult to find, as there is little information in the entity name or the text surrounding it that could be related to the concrete class of that entity.

This can be solved as a multiclass classification task. Additional information is provided for task 2, such as the following tags for the entities: isH, isN, iscomplex. These can easily be used for training, as they are already included in the provided dataset, but for prediction these have to be first predicted, as they are not available for the test set.

#### **3.3. Task 3**

For task 3, systems have to classify each clinical record according to several dimensions: is food, is nosocomial, is animal injury, is pet (binary classification in each case). Additionally, when any of these labels are detected in the document, systems have to search for the supporting evidence, that is, the codes of the entities being food, nosocomial, etc.

#### **3.4. Corpus Description**

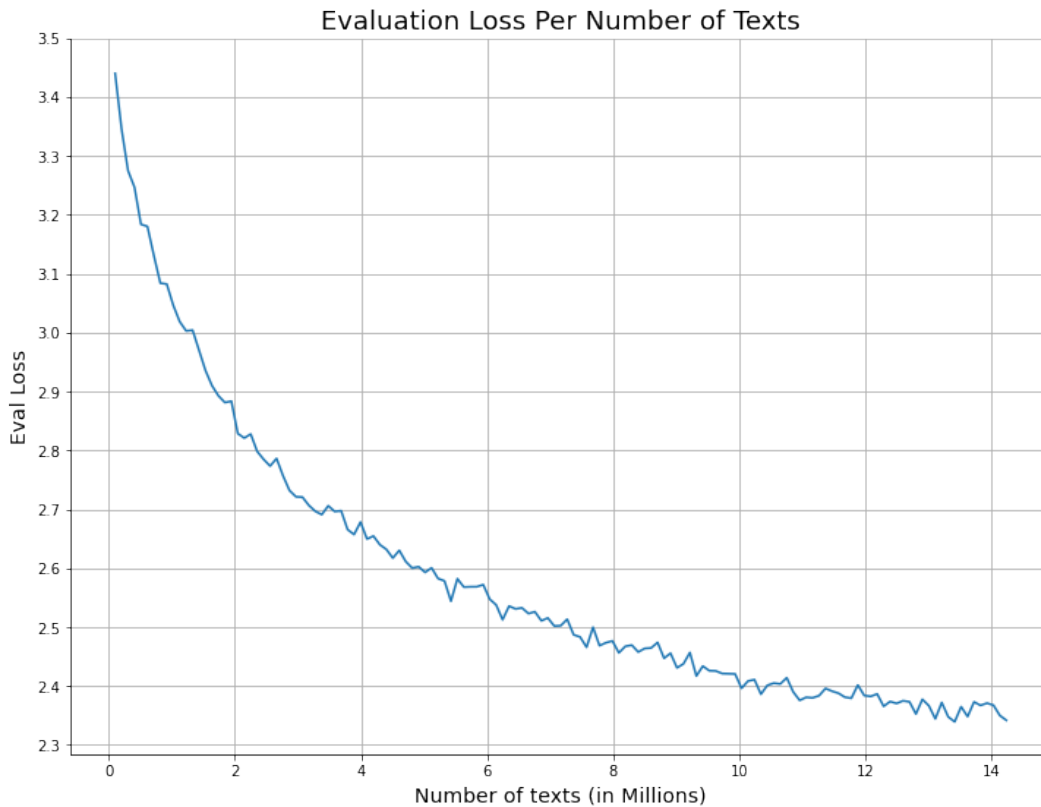
For this competition, [1] created a high quality dataset by having experts in microbiology annotating it, following well defined annotation guidelines, together with a consistency analysis. Furtherly, clinical cases resembling the content of clinical records from a variety of disciplines were manually selected [1]. This all enabled the creation of a dataset that eases the task for the models trained on it, thus it is expected that these are able to easily find patterns, specially in tasks 1 and 3.

## 4. System Description

### 4.1. Domain Adaptation for MarIA-Large

As stated previously, a domain adaptation strategy was used for MarIA-large [3]. There is existing evidence, as shown in [13], that adapting the domain of a language model before fine-tuning it to specific tasks of that domain can provide improvements. In [21], a domain adaptation strategy for language models is described, together with its promising results. Due to these works, it was decided to try the same with MarIA-large. Domain adaptation consists on repeating the pre-training procedure with small adaptations as described in [21] to adjust the model to texts of a specific domain.

For that, the Spanish Biomedical Crawled Corpus [22] was used, together with the texts from the train split of the LivingNER corpus [1]. This is a collection of about 3M documents crawled from more than 3,000 URLs belonging to Spanish biomedical and health domains. The hyperparameters for training the model were inspired on those by the original RoBERTa-large model [16], albeit they were adapted for a model that has already been sufficiently pre-trained, following the methodology described in [21]. For the validation split in the domain adaptation step, the validation data of the LivingNER corpus [1] was used.



**Figure 1:** Evolution of the Eval Loss during the Domain Adaptation Training for MarIA-large.

Hyperparameter	Values
Learning Rate	(1e-5, 7e-5, log)
Num Train Epochs	{3, 5, 7, 10, 15, 20}
Train Batch Size	{16, 32, 48, 64}
Warmup Steps Ratio	(0.01, 0.10, log)
Weight Decay	(1e-10, 0.3, log)
Adam Epsilon	(1e-10, 1e-6, log)
Number of trials	200
Initial Random Trials	40

**Table 1**

Hyperparameter space for BETO, BERTIN, MarIA-base, BioMedical and BioClinical models on Task 1.

Figure 1 shows the evolution of the eval loss in the pre-training task, which is a good sign since the validation split is formed of texts from the LivingNER challenge, therefore the model is expected to be able to learn more easily the future tasks it is used for. The total improvement is around 1.2 in terms of loss, related to MarIA-large [3]. This model is called from this point on *BioLarge*, and has been made publicly available. The model was trained on a 24 GB NVIDIA RTX 3090, and took around a week to complete the 14M texts training.

## 4.2. Models Training

### 4.2.1. Task 1: NER

For training the models, the Transformers library [23] is used, together with Datasets [24], both from Huggingface. As several texts were too large to be processed at once due to the models' maximum sequence length, 512 in all cases, they were processed with overflow. This way, more than one example was created from each text, with a window of 128 overlapping tokens in each case. This enabled the models to learn from labels at the end of the documents also. As will be seen later, this is also crucial for prediction.

For optimizing the models, Optuna [25] was used. The hyperparameter setups are depicted in tables 1 and 2. All data from the challenge, training and validation, was concatenated and then a small random split was splitted for validation purposes. Models are evaluated over this subset every epoch, stopping when they stop improving. Models are evaluated with tools from the Scikit Learn library [26]. All predicted labels and all correct labels are concatenated into a single array, thus not taking into account the texts to which each token belong. The metric used for deciding the best model is F1-score macro [27].

### 4.2.2. Meta-Ensembling the predictions for Task 1

Generally, when more than one model is used for the same task, the final predictions are more accurate. For this task, however, no real ensemble was proposed, due to the computing cost of training such a model. Nevertheless, a strategy was followed that used the predictions of several models and puts them together. Duplicate start offsets are removed, so that no duplicate entities are finally submitted. The longest of the overlapped entities (predicted by more than one model

Hyperparameter	Values
Learning Rate	(5e-6, 7e-5, log)
Num Train Epochs	{3, 5, 7, 10, 15, 20}
Train Batch Size	{16, 32, 48, 64}
Warmup Steps Ratio	(0.01, 0.10, log)
Weight Decay	(1e-10, 0.3, log)
Adam Epsilon	(1e-10, 1e-6, log)
Number of trials	200
Initial Random Trials	40

**Table 2**

Hyperparameter space for *BioLarge* and MarIA-large models on Task 2.

but with possible different ends) is chosen as the one to stay. Additionally, a filter was designed to only leave in entities that were predicted by more than one model, which increases precision but reduces recall. This filter is optional and takes place before the deduplication step.

#### 4.2.3. Task 2

For task 2, an approach is proposed to augment the information for the models. As the task consists on predicting the label for each entity, examples were created that were composed of the past and post full sentences to the sentence in which an entity appears, along with the current one being processed. The entity text is decorated with two <e> marks.

Additionally, to use all the information contained in the dataset, several binary classification tasks were developed to assign the isN, isH, iscomplex tags in the LivingNER dataset. For the binary classification tasks, the same preprocessing as for task 2 is used.

Far less resources were assigned to this task, so only *BioLarge* and MarIA-large were trained for it, with a simpler hyperparameters setting, depicted in table 3. Apart from those language models, and in order to compare the results of such a model against a classical Machine Learning model, it was decided to use a TFIDF vectorization [28] together with a Linear Support Vector Machine (SVM) [29]. The balanced weight for the loss function was used following the Scikit-learn configuration, and the hyperparameter space for this pipeline (vectorizer and model) is depicted in table 4

#### 4.2.4. Task 3

In task 3, the classification subtasks are at document-level, but it is useful to take the entities predicted into account, as the labels refer to them. For this reason, each label is highlighted in the text in the form "label: HUMAN. <e> persona <e>". This way, it is intended to provide the models with more information useful for classifying the text along the different axes.

Given that an English-translated version of the corpus was available, backtranslation [30] was carried out, with the use of MarianMT from the University of Helsinki [31].

In this case, again only *BioLarge* and MarIA-large are used. The hyperparameter space is the same as for task 2, shown in table 3. The second part of the task, consisting on choosing

Hyperparameter	Values
Learning Rate	(1e-5, 5e-5, log)
Num Train Epochs	{5, 10, 15, 20}
Train Batch Size	{16, 32, 48, 64}
Warmup Steps Ratio	(0.01, 0.10, log)
Weight Decay	(1e-2, 0.1, log)
Adam Epsilon	(1e-10, 1e-6, log)
Number of trials	15
Initial Random Trials	8

**Table 3**  
Hyperparameter space for *BioLarge* and MarIA-large models.

Hyperparameter	Values
TFIDF min df	{1, 2}
TFIDF max df	(0.1, 1.0)
TFIDF analyzer	{word, char, char wb}
TFIDF lowercase	{True, False}
SVM alpha	(10e-3, 10e3)
SVM max iter	{100, 500, 1000}
SVM eta0	(10e-6, 10e-1)
SVM power t	{0.1, 0.5, 2.0}

**Table 4**  
Hyperparameter space for TFIDF and Linear SVM.

which tags support the evidence for the labels assigned, is done by naively choosing all the codes detected.

As it is a sequence classification task, the aggregation of predictions is easier. *BioLarge* and MarIA-large logits are concatenated and averaged before getting the final predicted labels, therefore producing more reliable predictions.

## 5. Experiment and Evaluation Results

### 5.1. Results in the Validation Set for Task 1

#### 5.1.1. Results in terms of token-to-token

As models are evaluated in terms of their token-to-token f1-macro score, it is interesting to check how this differs from the official evaluation, that takes into account the full entities, and the documents to which they belong. This can provide some insights into the inner workings of the models and tokenizers depending on their domain adaptation.

Table 5 shows the results for each model in the validation set in terms of f1-macro [27], in the token-to-token task, which does not take into account the full entities that systems need to predict. The problem with this evaluation, although easier to implement for training

**Table 5**

F1-Score Results for LivingNER Eval MiniSplit in the Token-to-Token task.

Model	F1	Rank
MarIA-large	0.98743	4
MarIA-base	0.98610	5
BETO	0.97824	6
BERTIN	0.92432	7
BioClinical	<b>0.98910</b>	1
BioMedical	0.98892	3
<i>BioLarge</i>	0.98899	2

**Table 6**

F1-Score Results for LivingNER Eval MiniSplit for the Real Entities task.

Model	F1
MarIA-large	0.8769
MarIA-base	0.8827
BioClinical	0.9108
BioMedical	<b>0.9142</b>
<i>BioLarge</i>	0.9037

optimization purposes, is that it is dependent on each model’s tokenizer.

### 5.1.2. Results in terms of Real Entities for Task 1

For evaluating the best of these models in terms of the real entities predicted, the LivingNER evaluation library [1] was used. In the prediction phase, we need to predict entities even when they are more than 512 tokens away from the start of a text. For this reason, we use the same strategy as in training, explained previously. This has an additional advantage, because for the same token, several predictions can be obtained, if it is located in the overlapping sequences.

Furthermore, predicted labels are aggregated at the word level, and the final label is chosen by majority. In this way, if the model has failed to predict one part of the word as an entity but it has predicted correctly the other parts, it is selected as an entity. Words marked as entities are then aggregated with neighbored words having the same label, obtaining grouped entities.

Entities are cleaned when they start with an space, as well as with "(" or ";". The same is done with the end of the entities, replacing "(" by ")".

For this step, only the best performing models in the token-to-token task were used. These are MarIA, large and base, *BioLarge*, BioClinical and BioMedical. Table 6 shows their results in the minidev split used for validating the models’ training.

As can be seen in table 6, domain-specific models clearly outperform general models, as BioMedical and BioClinical are the two best performing models. The only exception is *BioLarge*, which is a previously general, domain-adapted model, and performs only slightly worse than BioMedical and BioClinical. The difference between domain-adapted and domain-specific models is greater in the Real Entities task. In fact, table 5 shows that *BioLarge* outperforms



**Table 7**  
F1-Score Results for LivingNER Eval MiniSplit for Task 2.

Model	Task	F1
MarIA-large	isN	0.6835
<i>BioLarge</i>	isN	<b>0.7839</b>
MarIA-large	isH	0.5974
<i>BioLarge</i>	isH	<b>0.6231</b>
MarIA-large	iscomplex	<b>0.5857</b>
<i>BioLarge</i>	iscomplex	0.5775
MarIA-large	NCBITax	0.0541
<i>BioLarge</i>	NCBITax	<b>0.0757</b>

BioMedical on the token-to-token task.

This is arguably due to the domain specificity of the tokenizer. When general domain tokenizers process biomedical texts, they tend to split entities into many tokens, as these are usually rare in a general domain corpus. This makes the task harder for them, as they have more subtokens to predict. Even when they predict a higher percentage of those tokens right, when aggregating the predictions into real entities they have a handicap against the domain-specific tokenizers, which is shown by the results in table 6.

## 5.2. Results in the Validation Set for Task 2

For task 2, several subtasks were needed, as explained previously. Table 7 shows the results in terms of F1 macro [27] for *BioLarge* and MarIA-large, using again Scikit-Learn library [26].

Although results are very poor for the NCBITax task, that is the final objective of task 2, a positive outcome of these results is that the domain adaptation carried out seems to be useful. As in tables 5 and 6, in 7 *BioLarge* performs better than MarIA-large on most tasks. Moreover, in those tasks in which it performs worse, the difference is not very significant.

When evaluated with the LivingNER Evaluation Library [1], *BioLarge* obtains 0.1555 f1 macro, which can be improved to 0.1965 if all HUMAN entities are set the code 9606.

## 5.3. Results in the Validation Set for Task 3

The results for all subtasks of task 3 in the minidev split are shown in table 8. In this case, the difference between *BioLarge* and MarIA-large is not so significant, in fact, in general terms MarIA-large has better metrics, although the differences are very small, since their f1 scores are very close to 1.

## 6. Official Test Results

For submitting the final predictions, systems had to evaluate over 13,472 clinical cases, as the test and background splits were mixed to avoid manual annotation of samples. In this section,

**Table 8**

F1-Score Results for LivingNER Eval MiniSplit for the Task 3.

Model	Task	F1
MarIA-large	isPet	<b>0.9938</b>
<i>BioLarge</i>	isPet	<b>0.9938</b>
MarIA-large	isAnimalInjury	<b>0.9938</b>
<i>BioLarge</i>	isAnimalInjury	0.9877
MarIA-large	isNosocomial	<b>0.9938</b>
<i>BioLarge</i>	isNosocomial	0.9877
MarIA-large	isFood	0.9938
<i>BioLarge</i>	isFood	<b>1.0</b>

**Table 9**

F1-Score Results for LivingNER Test Set (Official Results).

Run Name	Precision	Recall	F1-Macro
run1-BioLargeBioBaseBioClinicalNoFilter	0.9213	0.8637	0.8916
run2-BioMedicalBase	0.9334	0.8914	0.9119
run3-BioBaseBioClinicalFilter	0.9435	0.8685	0.9045
run4-BioLargeBioBaseBioClinicalFilter	0.9432	0.8373	0.8871
run5-BioBaseBioClinicalNoFilter	0.9228	0.908	0.9153

the official results from the competition are shown.

### 6.1. Results in the Test Set for Task 1

Table 9 shows the Precision, Recall and F1-Macro on the official test set for the different runs submitted. Run 1 uses *BioLarge*, BioMedical and BioClinical models. The filter explained previously when mixing their predictions was not used for that run; it was used though in run 4, with the same models. Run 2 is the BioMedical model alone. Run 3 is composed of BioMedical and BioClinical, with the filter activated, which increases the precision compared to the same but without filter, which is run 5.

Results in table 9 show the same pattern observed in the minidev split, as domain-specific models perform better, also when mixed, without using the domain-adapted model. This is specially significant in terms of recall, which can be explained by the tokenizer issue described previously, while less significant differences are appreciated in their precision scores.

The mean F1-score for this task was 0.8239, while the standard deviation was 0.2371, showing a great difference between submissions. The results obtained by this work are clearly above the mean, showing good performance for all the runs submitted.

**Table 10**

F1-Score Results for LivingNER Test Set (Official Results).

run name	Precision	Recall	F1
run1-BioLarge	0.512	0.4799	0.4954
run2-svm	0.4545	0.426	0.4398

## 6.2. Results in the Test Set for Task 2

Due to the amount of texts to predict for submitting, the MarIA-large model could not be used, as several models have to be used for each task 2 prediction (one for each binary task to get tags like isH for the entities plus task 2 itself). For this reason, only *BioLarge* and TFIDF+SVM were submitted. Table 10 shows the results for those models in terms of precision, recall and f1-score for this task.

The mean f1-score for this task among all participants was 0.8267 with a standard deviation of 0.1508, therefore it is clear that the results obtained by both *BioLarge* and SVM are deceptive, with very low scores. More work should be put into this task to come closer to the performance of other models presented at the workshop.

## 6.3. Results in the Test Set for Task 3

In the case of task 3, it is possible that some conceptual or coding mistake has been made, as all results in terms of precision, recall and f1 are 0 except for the animal injuries, that is very close to 0. The mean scores for this task are also close to 0, but the results of this work are not even there.

## 7. Conclusions and Future Work

In this work different solutions were presented for the tasks of the LivingNER challenge, with more focus on the NER task, in which official results show that all the systems presented perform better than average, around 0.91 vs 0.82. The results for tasks 2 and 3 are below average, showing that more effort should be put into those tasks to get at least an average performance.

Specifically, clearly understanding the nature of the tasks at hand could help, as it is possible that some proposed solutions do not perfectly match the nature of the data. A clear conclusion in this regard is that, specifically when the task is very specific to a domain, it is crucial to understand the tasks and challenges of that domain, which is left as future work.

Additionally, a new biomedical domain-adapted model was trained and released: *BioLarge*. It was trained using MarIA-large model and a biomedical corpus in Spanish. Results in evaluation, in all tasks and subtasks, show that *BioLarge* clearly outperforms MarIA-large on almost all tasks, thus proving the advantages of adapting a model to a domain before fine-tuning. However, BioClinical and BioMedical outperform *BioLarge* on the real entities task, with similar results in the token-to-token task.

This can be partially explained by the domain fit of their tokenizers, which has a crucial impact in terms of grouping subwords and group of words into entities. The vocabulary of

general domain tokenizers does not include many of the terms specific to a domain, which is a handicap for general domain models, that must predict more tokens right to have the same number of complete words correct.

This work provided a complete review of openly available language models in Spanish for different tasks, increasing the existing knowledge about their general performance. Results in the NER task exhibit a high correspondence to those of [3] [4]. However, in order to have a complete assessment, it would be necessary to include in the comparison process the BioMedical and BioClinical models in tasks 2 and 3.

## References

- [1] 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.6576488>. doi:10.5281/zenodo.6576488, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).
- [2] J. e. a. Cañete, Spanish pre-trained bert model and evaluation data, 2020. URL: <https://users.dcc.uchile.cl/~jperez/papers/pml4dc2020.pdf>.
- [3] A. Gutiérrez-Fandiño, J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodríguez-Penagos, A. Gonzalez-Agirre, M. Villegas, Maria: Spanish language models, *Procesamiento del Lenguaje Natural* 68 (2022) 39–60. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405>.
- [4] A. Vaca Serrano, G. G. Subies, H. M. Zamorano, N. A. Garcia, D. Samy, D. B. Sanchez, A. M. Sandoval, M. G. Nieto, A. B. Jimenez, Rigoberta: A state-of-the-art language model for spanish, 2022. URL: <https://arxiv.org/abs/2205.10233>. doi:10.48550/ARXIV.2205.10233.
- [5] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrenondo, H. Rodríguez, J. A. Lopez Martin, M. Villegas, M. Krallinger, MEDDOCAN corpus: gold standard annotations for Medical Document Anonymization on Spanish clinical case reports, 2020. URL: <https://doi.org/10.5281/zenodo.4279323>. doi:10.5281/zenodo.4279323, Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).
- [6] M. Marimon, A. Gonzalez-Agirre, A. Intxaurrenondo, H. Rodríguez, J. L. Martin, M. Villegas, M. Krallinger, Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results, in: *IberLEF@SEPLN*, 2019.
- [7] L. Lange, H. Adel, J. Strötgen, NLNDE: the neither-language-nor-domain-experts' way of spanish medical document de-identification, *CoRR abs/2007.01030* (2020). URL: <https://arxiv.org/abs/2007.01030>. arXiv:2007.01030.
- [8] L. Mouselimis, fastText: Efficient Learning of Word Representations and Sentence Classification using R, 2022. URL: <https://CRAN.R-project.org/package=fastText>, r package version 1.0.2.
- [9] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*,

- Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 54–59. URL: <https://aclanthology.org/N19-4010>. doi:10.18653/v1/N19-4010.
- [10] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Computation* 9 (1997) 1735–1780.
- [11] S. Lima-López, E. Farré-Maduell, A. Miranda-Escalada, V. Brivá-Iglesias, M. Krallinger, Nlp applied to occupational health: Meddoprof shared task at iberlef 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts, *Procesamiento del Lenguaje Natural* 67 (2021) 243–256. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6393>.
- [12] S. L.-L. y Eulàlia Farré-Maduell y Antonio Miranda-Escalada y Vicent Brivá-Iglesias y Martin Krallinger, Nlp applied to occupational health: Meddoprof shared task at iberlef 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts, *Procesamiento del Lenguaje Natural* 67 (2021) 243–256. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6393>.
- [13] L. Lange, H. Adel, J. Strötgen, Boosting transformers for job expression extraction and classification in a low-resource setting, in: *Proceedings of The Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings, 2021. URL: [http://ceur-ws.org/Vol-2943/meddoprof\\_paper1.pdf](http://ceur-ws.org/Vol-2943/meddoprof_paper1.pdf).
- [14] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, *CoRR abs/1911.02116* (2019). URL: <http://arxiv.org/abs/1911.02116>. arXiv:1911.02116.
- [15] J. e. a. Devlin, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: <https://arxiv.org/pdf/1810.04805.pdf>.
- [16] Y. e. a. Liu, Roberta: A robustly optimized bert pretraining approach, 2019. URL: <https://arxiv.org/pdf/1907.11692.pdf>.
- [17] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners (2018). URL: <https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf>.
- [18] J. D. la Rosa y Eduardo G. Ponferrada y Manu Romero y Paulo Villegas y Pablo González de Prado Salas y María Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, *Procesamiento del Lenguaje Natural* 68 (2022) 13–23. URL: <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403>.
- [19] 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, *CoRR abs/2109.03570* (2021). URL: <https://arxiv.org/abs/2109.03570>. arXiv:2109.03570.
- [20] C. Schoch, S. Ciufu, C. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. McVeigh, K. O’Neill, B. Robbertse, S. Sharma, V. Soussov, J. Sullivan, L. Sun, S. Turner, I. Karsch-Mizrachi, Ncbi taxonomy: A comprehensive update on curation, resources and tools, *Database* 2020 (2020). doi:10.1093/database/baaa062.
- [21] C. Karouzos, G. Paraskevopoulos, A. Potamianos, UDALM: unsupervised domain adaptation through language modeling, *CoRR abs/2104.07078* (2021). URL: <https://arxiv.org/abs/2104.07078>. arXiv:2104.07078.

- [22] C. P. Carrino, J. Armengol-Estapé, O. de Gibert Bonet, A. Gutiérrez-Fandiño, A. Gonzalez-Agirre, M. Krallinger, M. Villegas, Spanish biomedical crawled corpus: A large, diverse dataset for spanish biomedical language models, 2021. [arXiv:2109.07765](https://arxiv.org/abs/2109.07765).
- [23] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface’s transformers: State-of-the-art natural language processing, *CoRR abs/1910.03771* (2019). URL: <http://arxiv.org/abs/1910.03771>. [arXiv:1910.03771](https://arxiv.org/abs/1910.03771).
- [24] Q. Lhoest, A. Villanova del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, J. Davison, M. Šaško, G. Chhablani, B. Malik, S. Brandeis, T. Le Scao, V. Sanh, C. Xu, N. Patry, A. McMillan-Major, P. Schmid, S. Gugger, C. Delangue, T. Matussière, L. Debut, S. Bekman, P. Cistac, T. Goehringer, V. Mustar, F. Lagunas, A. Rush, T. Wolf, Datasets: A community library for natural language processing, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 175–184. URL: <https://aclanthology.org/2021.emnlp-demo.21>. [arXiv:2109.02846](https://arxiv.org/abs/2109.02846).
- [25] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, *CoRR abs/1907.10902* (2019). URL: <http://arxiv.org/abs/1907.10902>. [arXiv:1907.10902](https://arxiv.org/abs/1907.10902).
- [26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* 12 (2011) 2825–2830.
- [27] J. Opitz, S. Burst, Macro F1 and macro F1, *CoRR abs/1911.03347* (2019). URL: <http://arxiv.org/abs/1911.03347>. [arXiv:1911.03347](https://arxiv.org/abs/1911.03347).
- [28] A. Aizawa, An information-theoretic perspective of tf-idf measures, *Information Processing Management* 39 (2003) 45–65. URL: <https://www.sciencedirect.com/science/article/pii/S0306457302000213>. doi:[https://doi.org/10.1016/S0306-4573\(02\)00021-3](https://doi.org/10.1016/S0306-4573(02)00021-3).
- [29] N. Cristianini, E. Ricci, *Support Vector Machines*, Springer US, Boston, MA, 2008, pp. 928–932. URL: [https://doi.org/10.1007/978-0-387-30162-4\\_415](https://doi.org/10.1007/978-0-387-30162-4_415). doi:[10.1007/978-0-387-30162-4\\_415](https://doi.org/10.1007/978-0-387-30162-4_415).
- [30] S. Edunov, M. Ott, M. Auli, D. Grangier, Understanding back-translation at scale, *CoRR abs/1808.09381* (2018). URL: <http://arxiv.org/abs/1808.09381>. [arXiv:1808.09381](https://arxiv.org/abs/1808.09381).
- [31] J. Tiedemann, S. Thottingal, OPUS-MT – Building open translation services for the World, in: *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal, 2020.