UC3M-PUCPR at SemEval-2022 Task 11: An Ensemble Method of Transformer-based Models for Complex Named Entity Recognition

Elisa Terumi Rubel Schneider[1], **Renzo M. Rivera-Zavala**[2], **Paloma Martinez**[2], **Claudia Moro**[1] and **Emerson Cabrera Paraiso**[1]

[1]Graduate Program on Informatics, Pontifícia Universidade Católica do Paraná, Brazil
paraiso@ppgia.pucpr.br

[2]Computer Science and Engineering Department, Universidad Carlos III de Madrid, Spain
pmf@inf.uc3m.es

Abstract

This study introduces the system submitted to the SemEval 2022 Task 11: MultiCoNER (Multilingual Complex Named Entity Recognition) by the UC3M-PUCPR team. We proposed an ensemble of transformer-based models for entity recognition in cross-domain texts. Our deep learning method benefits from the transformer architecture, which adopts the attention mechanism to handle the long-range dependencies of the input text. Also, the ensemble approach for named entity recognition (NER) improved the results over baselines based on individual models on two of the three tracks we participated in. The ensemble model for the code-mixed task achieves an overall performance of 76.36% F1-score, a 2.85 percentage point increase upon our individually best model for this task, XLM-RoBERTa-large (73.51%), outperforming the baseline provided for the shared task by 18.26 points. Our preliminary results suggest that contextualized language models ensembles can, even if modestly, improve the results in extracting information from unstructured data.

1 Introduction

Named Entity Recognition (NER) is a Natural Language Processing (NLP) technique to extract relevant information from unstructured natural language data, identifying and categorizing entities in texts and thereby supporting other NLP tasks.

However, processing complex and ambiguous entities is a challenging task that has not received sufficient attention from the research community (Meng et al., 2021). These complex entities can be formed by any linguistic constituent (long noun phrases and sometimes complete sentences), as titles of creative works such as books and movies, different from traditional entities like person names and locations. As the creative work titles can be semantically ambiguous, the challenge is recognizing entities based mainly on their context.

The SemEval 2022 Task 11: MultiCoNER (Multilingual Complex Named Entity Recognition) (Malmasi et al., 2022b) targets the recognition of lowercase complex entities in multilingual and multi-domain texts, encouraging researchers to develop new approaches to extract diversified entity types. The challenge involves the use of human language modeling in the NER task, on cross-domain texts. The semantic structure has six types of entities: person, location, group, corporation, product and creative work. The dataset was provided for 11 languages, and, in addition, this task also provided multi-language and code-mixed [1] features, encouraging the development of more generic and adaptive systems.

As the contextualized pre-trained language models based on the transformer architecture have reached the state-of-the-art in several NLP tasks (Vaswani et al., 2017), in our method, we explore transformer-based models and combined the results into an ensemble, which can make better predictions and have a superior performance than any single contributing model (as demonstrated by Copara et al. (2020), Knafou et al. (2020) and Hernandez et al. (2021)). The models were fine-tuned to NER task on the English, Spanish and code-mixed datasets.

The paper is organised as follows: Section 2 provides a brief explanation of related works, Section 3 provides details about the data used for training our models, Section 4 describes the proposed method with implementation details, Section 5 presents our results with some discussions, and in Section 6 we present the conclusions obtained from the observed results.

2 Related Works

Context-dependent representations models, pre-trained on large-scale unstructured data, particu-

[1] With entities in a different language than the rest of the query, as explained in (Fetahu et al., 2021).

Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), pages 1448 - 1456
July 14-15, 2022 ©2022 Association for Computational Linguistics

larly those supported by the transformer architecture (Vaswani et al., 2017), have been reached the state-of-the-art performance in NLP problems, including NER task.

These contextual word embedding models use the learned representations over the large data and, in a process called fine-tuning, have their last layers updated to adapt for a downstream task, using task-specific training data.

The Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) is a language representation model designed to pre-train deep bidirectional representations from unlabeled text conditioning on both left and right context in all layers. A Transformer is a encoder-decoder architecture, but in BERT-based models only the encoder is used, with the attention mechanism that learns contextual relationships between words in a text. The BERT-based models are pre-trained using two language generic tasks: masked language modeling (MLM) and next-sentence-prediction (NSP) tasks.

RoBERTa was built on BERT's language masking strategy, where the system learns to predict hidden sections of text within otherwise unannotated language examples. RoBERTa modifies key hyper-parameters in BERT, as removing next-sentence pretraining objective. Also, it was trained with much larger mini-batches and learning rates, improving the masked language modeling objective and leading to better downstream task performance (Conneau et al., 2020).

DistilBERT learns a distilled version of BERT, maintaining almost all performance but using only half the number of parameters with a technique called distillation, which approximates the large neural network by a smaller one (Sanh et al., 2019). We also chose DistilBERT on our ensemble because, although it has 40% less parameters than BERT-base and runs 60% faster, can preserve over 95% of BERT's performances (Sanh et al., 2019).

XLM-RoBERTa is a transformer-based masked language model trained on one hundred languages, which outperformed multilingual BERT on a variety of cross-lingual benchmarks (Conneau et al., 2020).

Some researches have focused on the combination of transformer-based models into an ensemble, for NLP tasks. In the work of Copara et al. (2020), the ensemble of contextualized language models resulted in an effective approach for NER in chemical patent documents, outperforming individual transformer models. Knafou et al. (2020) achieved high results with ensemble approach for NER and their study indicates that the more models were used in the ensemble, the more the performances tend to be high and stable. The research of Hernandez et al. (2021) shows that the combination of three BERT-based models obtained superior results than the individual models, in the classification of texts from social media.

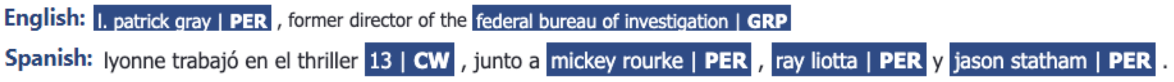
Motivated by the success of transformer-based models along with the ensemble approach, we trained several models and combined the results into an ensemble.

3 Data

This shared task encourages NLP researchers to develop complex NER systems for 11 languages, with semantically ambiguous and complex entities in short and low-context settings. Complex entities as creative works are really challenging as they are harder to recognize. An example of entities for English and Spanish can be seen in Figure 1.

According to the organizers (Malmasi et al., 2022a), 15,300 sentences were made available for training and 800 sentences for evaluation in the mono-lingual tracks. The training file has 15,000 sentences for code-mixed and the evaluation dataset has 500 sentences. Test files have varying sizes, with 219,652 sentences for English, 272,887 sentences for Spanish, and 100,000 for code-mixed tracks.

Data enrichment: To improve the performance of our models, we enriched our datasets with other NER annotated corpus containing similar entities. For English, we concatenated to the original MultiCoNER dataset the corpus CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003), emerging entities (Derczynski et al., 2017), and In Media Res (Braşoveanu et al., 2020). CoNLL-2003 is a NER dataset released in CoNLL-2003 shared task for



English: l. patrick gray | PER , former director of the federal bureau of investigation | GRP

Spanish: lyonne trabajó en el thriller 13 | CW , junto a mickey rourke | PER , ray liotta | PER y jason statham | PER .

Figure 1: Example of data for English and Spanish (Malmasi et al., 2022a).

language-independent NER. We used the English data (Tjong Kim Sang and De Meulder, 2003), provided from Reuters Corpus containing entities of location, person, organization, and miscellaneous. We converted this data into MultiCoNER format, transforming organization into corporation. The emerging entities (Derczynski et al., 2017) is a corpus released by WNUT2017 Shared Task with focus on rare and emerging entity recognition from noisy user-generated data. This corpus contains the entities person, location, corporation, product, creative-work, and group, the same entity types we have on MultiCoNER. In Media Res is a corpus for evaluating named entity linking with creative works (Braşoveanu et al., 2020), annotated from Wikipedia articles with entities like person, organization, location, (creative) work, event and other. We converted it to MultiCoNER format, excluding events and others and converting work to creative work.

For Spanish, we enriched the dataset with CoNLL-2002 (Tjong Kim Sang, 2002) and Wikiner (Nothman et al., 2012). CoNLL-2002 contains entities of type persons, organizations, locations, times and quantities. We use the Spanish texts, converting organization to corporation, and eliminating entities we do not use. Wikiner (Nothman et al., 2012) is an annotated dataset of Wikipedia texts for 9 languages. It has entities of person, location, organization, and other (misc). We used the texts in Spanish, and we did the conversion from organization to corporation and an analysis of entities of type misc, where could be categorized as products, creative works and others (ignored). Because of this analysis of misc-type entities, we only processed 10% of the corpus.

All the datasets were converted to lowercase format, and concatenated to the original MultiCoNER train dataset.

4 System Description

Our method consists on an ensemble of transformer-based models, trained for NER task and joined by the soft voting method (Figure 2). From the 11 languages, we decided to participate in the English and Spanish tracks, since we already have some related research on them (Rivera and Martinez (2021) and Akhtyamova et al. (2020)), and also in the multilingual and code-mixed tracks to assess our strategy on multilingual texts.

In this section, we detail the development of NER models with its parameters and training setup, the ensemble method, and metrics used in this shared task.

4.1 NER models

We use the transformer-based pre-trained models as checkpoint for fine-tuning on the NER task, with the enriched MultiCoNER datasets. In the fine-tuning step, the existing models were specialized for NER task, where a fully connected layer was added on top of the hidden states of each token to classify tokens according to the named-entities classes. For each track we

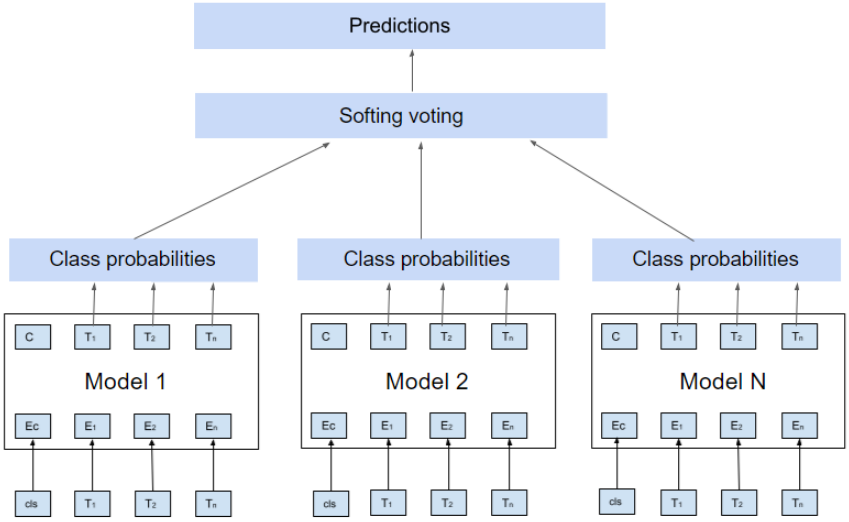


Figure 2: Architecture of our ensemble approach.

Pretrained Models	#Params	Track
Beto-uncased (Cañete et al., 2020)	110M	Spanish
BERT-base (Devlin et al., 2019)	110M	English
BERT-large (Devlin et al., 2019)	340M	English
BERT-base-multilingual(Devlin et al., 2019)	110M	English, Spanish, code-mixed
DistilBERT-base (Sanh et al., 2019)	66M	English, Spanish, code-mixed
Electra-discriminator (Clark et al., 2020)	110M	English
RoBERTa-base (Liu et al., 2019)	110M	English
RoBERTa-large (Liu et al., 2019)	340M	English
RoBERTa-base-bne (Gutiérrez-Fandiño et al., 2022)	125M	Spanish
RoBERTa-large-bne (Gutiérrez-Fandiño et al., 2022)	355M	Spanish
XLM-RoBERTa-base (Conneau et al., 2020)	270M	English, Spanish, code-mixed

Table 1: Pretrained models features.

participated, we performed the fine-tuning for NER on the pre-trained models. Table 1 shows all the models used in the ensembles and their number of parameters.

English track: The following pre-trained models, trained on a large corpus of English text, were fine-tuned on English train dataset: BERT-base-uncased (Devlin et al., 2019), BERT-large-uncased (Devlin et al., 2019), DistilBERT-base-uncased (Sanh et al., 2019), RoBERTa-base (Liu et al., 2019), RoBERTa-large (Liu et al., 2019), Electra-discriminator (Clark et al., 2020), in addition to the multilingual BERT-base-multilingual-uncased (Devlin et al., 2019) and XLM-RoBERTa-base (Conneau et al., 2020).

Spanish track: We fine-tuned on Spanish train dataset the following pre-trained models, trained on a large corpus of Spanish text: Beto uncased (Cañete et al., 2020), RoBERTa-base-bne (Gutiérrez-Fandiño et al., 2022), RoBERTa-large-bne (Gutiérrez-Fandiño et al., 2022), in addition to the multilingual BERT-base-multilingual-uncased (Devlin et al., 2019), XLM-RoBERTa-base (Conneau et al., 2020) and DistilBERT (Sanh et al., 2019).

For English and Spanish, we also pre-trained and fine-tuned a Megatron model (Shoeybi et al., 2019), however, it was not possible to incorporate it in the ensemble on time, and therefore it was separately evaluated as an unofficial result.

Multilingual and code-mixed track: We first fine-tuned models on the multilingual dataset provided by the shared task and then fine-tuned again on the code-mixed dataset, as the code-mixed training dataset is too small and training just with it could have lower performance. We fine-tuned these multilingual models[2]: BERT-base-multilingual-uncased (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020), and DistilBERT (Sanh et al., 2019). Due to time constraints, although we had trained models for the multilingual track, we could not run the predictions on the test dataset on time to submit to the multilingual track (since it was large files), so our results on multilingual track are unofficial.

All language models were fine-tuned on the training dataset explained in the previous section. We fine-tuned for 10 epochs, with a sequence length of a maximum of 128 tokens, a learning rate of 4e-5, warmup proportion of 0.06, and Adam optimizer for the deep learning training. We used the Pytorch implementation of Hugging Face library (Wolf et al., 2020).

Evaluation metrics: The metric returned by the shared task submission system is F1 macro average performance. For a better analysis and discussion, the performance is also examined per class label.

4.2 Ensemble method

Our ensemble method is based on a soft voting strategy (Schwenker, 2013), where each model returns its predicted probabilities and the class label is obtained by an argmax of the sum of all probabilities (in contrast to hard voting, where the system uses predicted class labels for majority rule voting).

In other words, for a given document, all models infer their predictions independently for each entity and return the probabilities of each label, i.e., their logits[3]. For each token, we sum the logits of all models of the ensemble for this given token and

[2] Which attends all languages of the multilingual and code-mixed tracks.

[3] Logits are the model outputs before application of an activation function (e.g. softmax).

then apply an argmax function, to assign a label to this passage.

The final label output for a token *n* is given by:

$$y_n(x_n) = argmax(\sum_{i=1}^{T} w_i h_i(x_n)) \qquad (1)$$

where *T* is the number of models participating in the ensemble, $\mathrm{w_i}$ is the weight of a model *i*, and $\mathrm{h_i(x_n)}$ is a list with the probabilities for all classes of the model *i* for the token *n*.

For each track, we performed three ensemble experiments.

Experiment 1: As explained above, the token label is an argmax of the sum of predicted probabilities of all models. In this experiment, all models have the same weight (w).

Experiment 2: Each model receives a score ranging from 0.1 to 1, which indicates its efficiency in the evaluation dataset (best overall F1-score). To identify the label of each token, we sum the output of all models, i.e., the probabilities of the token belonging to each class, in a weighted way. In other words, the predicted class probabilities for each classifier are multiplied by the classifier weight, summed across all models, and averaged. Thus, the final class label is then derived from the higher average probability (argmax).

Experiment 3: As in experiment 2, we also calculate weights for each model according to its effectiveness in the evaluation set, but on this experiment, we also calculate weights for each class as well (entity type). The model that performed better for that specific class has the highest weight, also values between 0.1 and 1.

5 Results and Discussion

Official results: The share-task proceedings reports the results (macro F1-score) by the participating systems in the SemEval 2022 Task 11 (Malmasi et al., 2022b), where our results are shown as `UC3M-PUCPR` team. These results, as well as the analysis performed in the discussion section, are exclusively computed on the test set. Our results refer to experiment 2 (where we used weights for each model), which performed better for all tracks. For the English track, our approach was ranked in 21th place from 30 participants, with a F-Score of 69.24, outperforming the baseline provided for the shared task by 8.04 points. On the other hand, for the Spanish track, our system had a significant drop, ranking in the last position (18th from 18 teams), with a F1-Score of 56.79, i.e., 33.15 points behind first place and 0.6 points behind baseline. Our Megatron-trained model performed better than the ensemble in this track, with 60.45 of F1 (unofficial result). In the code-mixed track, our approach ranked in 12th place from 22 participants, with an F1-Score of 76.36, outperforming the baseline by 18.26. For the multilingual track, we achieved 66.41 of F1, 12.31 points higher than baseline, but this result are not official since it was obtained after the competition deadline. Table 2 shows the official F1-scores of our submissions separated by classes, which allows the performance analysis of each entity individually, for all tracks. In table 3, we can assess the performance of each model separately, per class, for English track.

Named entity	English	Spanish	(*) Multilingual	Code-mixed
Location	0.6821	0.5807	0.7312	0.7925
Person	0.8470	0.7522	0.8017	0.8631
Product	0.6695	0.5102	0.6266	0.7692
Group	0.6953	0.5287	0.5945	0.7051
Creative work	0.6171	0.4350	0.6020	0.6937
Corporation	0.6437	0.6003	0.6741	0.7570
All	0.6924	0.5679	0.6641	0.7636

Table 2: Our F1 results for each class. (*) Multilingual results are unofficial.

Discussion: The ensemble method modestly improved the F1 performance in relation to the use of individual models for English (1.92 of ensemble improve in relation to the best individual model, BERT-large) and code-mixed (2.85 of ensemble improve in relation to the best individual model, XLM-RoBERTa-large). However, for Spanish track, no improvement was seen in the ensemble compared to Beto individually.

For English track, where we work with eight different pre-trained models, the ensemble model outperforms the other models for all classes, as can be seen in table 3, where the F1 seem relatively stable across all the pre-trained models.

To help understand the reason for the low results in Spanish, we present a confusion matrix, in figure 3. In this track, an analysis of the errors indicates that the model had difficulty recognizing the entities products, groups and creative works. By labeling them with "O" instead of the correct class, lowers the recall value for this entities.

When analyzing the performance of the entities

Language Models	Location	Person	Product	Group	Creative work	Corporation	F1
BERT-base	0.6703	0.8268	0.6283	0.6602	0.5777	0.6018	0.6608
BERT-large	0.6706	0.8309	0.6412	0.6766	0.5971	0.6228	0.6732
BERT-base-multilingual	0.6703	0.8268	0.6283	0.6602	0.5777	0.6018	0.6608
DistilBERT-base	0.6531	0.8189	0.5996	0.6445	0.5476	0.5869	0.6418
Electra-discriminator	0.6550	0.8323	0.6435	0.6570	0.5958	0.6130	0.6661
(*) Megatron	0.5860	0.7302	0.5242	0.5535	0.4374	0.5424	0.5623
RoBERTa-base	0.6203	0.7669	0.5565	0.6047	0.5256	0.5334	0.6012
XLM-RoBERTa-base	0.6623	0.7868	0.5930	0.6072	0.5355	0.5697	0.6258
Ensemble	**0.6821**	**0.847**	**0.6695**	**0.6953**	**0.6171**	**0.6437**	**0.6924**

Table 3: F1-score by model on the English track test set. (*) Megatron model was not part of the ensemble.

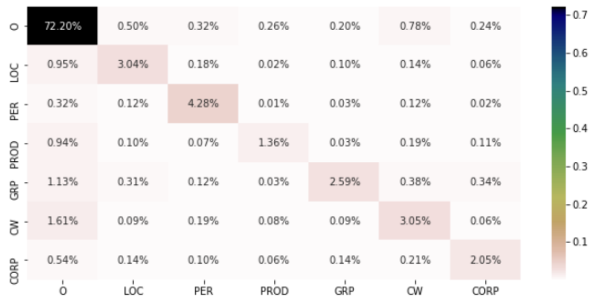


Figure 3: Confusion matrix on Spanish test data.

separately, we noticed that the creative work class had the worst F1 value for English, Spanish and code-mixed tracks, while the person class had the best performance. This is a sign that even in lower case, the trained models could still identify a person's name, i.e., they improved their ability to classify the token based on its context, which is a plus point for transformer-based contextualized models. The creative work entity was already expected to be the most challenging, since movie and book names are not simple noun phrases and are harder to recognize. As noted by Meng et al. (2021), the use of well-formed text with "easy" entities perform better than unseen entities or noisy text. Figure 4 shows how different models detected person and creative work entities, from English evaluation dataset, where we can see correct and incorrect detection in relation to creative work entity. Although some individual models incorrectly predicted the creative work entity, the ensemble achieved the correct answer.

Furthermore, we believe that the prediction of group, corporation, and product entities caused overlap between them because they can be semantically similar, which led to lower than expected results. The datasets we used to enrich the original MultiCoNER datasets do not have these entities so clearly.

For English and code-mixed tracks, we noticed that the large models performed better than the base models, which was already expected, agreeing with the results of previous research. For English, the best-isolated model was BERT-large-uncased, and for code-mixed, XLM-RoBERTa-large, both for evaluation and testing phases. On the other hand, on Spanish track, the uncased version of Beto model performed better than the new RoBERTa models generated from more than 201 million documents (570 Gb) from the Spanish National Library.

Despite improving performance, only the ensemble strategy may not be enough to have competitive results, when analyzing the results of the other teams, especially for the Spanish task, which had poor results.

The enrichment of the train databases may have helped for some entities, but it was not enough to significantly improve performance either, since our best results were precisely in code-mixed track, in which we did not use this technique. The entity types between the external databases may have minor semantic differences, for example, there is a fine line between group and organization. Maybe a closely analyze and even manual re-annotation on these entities could have improved performance.

Also, given such a large amount of test data (files with more than 200,000 sentences), the use of ensemble with transformer-based models may not be the most suitable for processing on regular computers, without much processing GPU power, due to the need to process this large amount of data with several models. In addition to the prediction process, the weighted sum of probabilities also consumes much memory. Our team could not process the test files for the multilingual track before the deadline, and we could not process the Megatron model on time to put into the ensembles (which could improve the results).

We realize that complex entities still represent a great challenge to the NLP community. In addition, lowercase texts, sentences without punc-

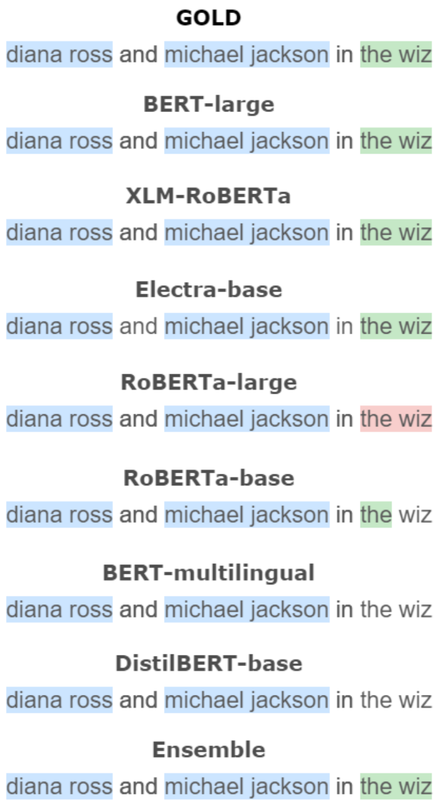

GOLD
diana ross and michael jackson in the wiz
BERT-large
diana ross and michael jackson in the wiz
XLM-RoBERTa
diana ross and michael jackson in the wiz
Electra-base
diana ross and michael jackson in the wiz
RoBERTa-large
diana ross and michael jackson in the wiz
RoBERTa-base
diana ross and michael jackson in the wiz
BERT-multilingual
diana ross and michael jackson in the wiz
DistilBERT-base
diana ross and michael jackson in the wiz
Ensemble
diana ross and michael jackson in the wiz

Figure 4: An example of predictions by different models from English evaluation dataset, where blue represents person entity type, green represents creative works and pink, group type.

tuation or grammatically incomplete, in a cross-domain dataset without any standardization of sentence length, make it even more challenging to extract information from unstructured data, as was the case in this task.

6 Conclusions

We presented our method for recognizing complex entities, which consists in a transformer-based models ensemble. Although our team's results are not among the first, the ensemble method worked for both English and code-mixed tracks, in which we obtained an improvement in the F1 value compared to individual transformer models. Our method outperformed in macro F1 the baseline provided by the organizers in 8.04 points on English track, 18.26 on code-mixed track and 12.31 on multilingual track (unofficial).

As our method is based only on machine learning, without fixed rules, this indicates that the transformer models were able to take advantage of natural language contexts to capture the most relevant features, on lowercase cross-domain texts.

Since extraction of complex entities represents a challenge and these entities are increasingly present, we would like to continue improving our method. Future work intends to train models using a different split (hold out) of the data, training models with more data and for more epochs. Also, since creative works had the worst performance on the three tracks, a hybrid approach that includes a dictionary with creative works can contribute to better results. Furthermore, an analysis of which models should participate in the ensemble can lead to better results and lower processing costs. For example, less robust models like DistilBERT that are faster and lighter, but do not improve performance, possibly do not contribute to the improvement of the results.

Despite having the lowest task evaluation score for the Spanish track, this method exhibited competitive performance at English and code-mixed tracks.

Acknowledgements

The authors thank the shared task organizers for the organization of this event. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Partially supported by Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with UC3M in the line of Excellence of University Professors (EPUC3M17) and PID2020-116527RB-I00 project (ACCESS2MEET).

References

Liliya Akhtyamova, Paloma Martínez, Karin Verspoor, and John Cardiff. 2020. Testing contextualized word embeddings to improve ner in spanish clinical case narratives. *IEEE Access*, 8:164717–164726.

Adrian M. P. Braşoveanu, Albert Weichselbraun, and Lyndon J. B. Nixon. 2020. In media res: A corpus for evaluating named entity linking with creative works. In *CONLL*.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Pre-training transformers as energy-based cloze models. In *EMNLP*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jenny Copara, Nona Naderi, Julien Knafou, Patrick Ruch, and Douglas Teodoro. 2020. Named entity recognition in chemical patents using ensemble of contextual language models. In *Working notes of the CLEF 2020, 22-25 September 2020*.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. Gazetteer Enhanced Named Entity Recognition for Code-Mixed Web Queries. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1677–1681.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. MarIA: Spanish language models. *Procesamiento del Lenguaje Natural*, 68(0):39–60.

Luis Alberto Robles Hernandez, Rajath Chikkatur Srinivasa, and Juan M Banda. 2021. A pharmacovigilance application of social media mining: An ensemble approach for automated classification and extraction of drug mentions in tweets. In *NeurIPS 2021 Workshop LatinX in AI*.

Julien Knafou, Nona Naderi, Jenny Copara, Douglas Teodoro, and Patrick Ruch. 2020. BiTeM at WNUT 2020 shared task-1: Named entity recognition over wet lab protocols using an ensemble of contextual language models. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 305–313, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: a Large-scale Multilingual dataset for Complex Named Entity Recognition.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Association for Computational Linguistics.

Tao Meng, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2021. GEMNET: Effective gated gazetteer representations for recognizing complex entities in low-context input. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1499–1512.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. 2012. Learning multilingual named entity recognition from Wikipedia. In *Artificial Intelligence*, volume 194, pages 151–175. Elsevier.

Renzo Rivera and Paloma Martinez. 2021. Analyzing transfer learning impact in biomedical cross-lingual named entity recognition and normalization. *BMC Bioinformatics*, 22.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC*$^2Workshop$.

Friedhelm Schwenker. 2013. Ensemble methods: Foundations and algorithms [book review]. *Computational Intelligence Magazine, IEEE*, 8:77–79.

Mohammad Shoeybi, Mostofa Ali Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *ArXiv*, abs/1909.08053.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.