LLM-RM at SemEval-2023 Task 2: Multilingual Complex NER using XLM-RoBERTa

Rahul Mehta
IIIT Hyderabad,India
rahul.mehta@research.iiit.ac.in

Vasudeva Varma
IIIT Hyderabad,India
vv@iiit.ac.in

Abstract

Named Entity Recognition(NER) is a task of recognizing entities at a token level in a sentence. This paper focuses on solving NER tasks in a multilingual setting for complex named entities.Our team, LLM-RM participated in the recently organized SemEval 2023 task, Task 2: MultiCoNER II,Multilingual Complex Named Entity Recognition. We approach the problem by leveraging cross-lingual representation provided by fine-tuning XLM-Roberta base model on datasets of all of the 12 languages provided - Bangla, Chinese, English, Farsi, French, German, Hindi, Italian, Portuguese, Spanish, Swedish and Ukrainian.

1 Introduction

Named Entity Recognition(NER) is the task of recognizing entities (e.g., person, location, organization) in a piece of text.

Most of the NER datasets like CoNLL2003 (Erik F. Tjong and Meulder, 2003) are focussed on high-resource languages like English and are specific to a given domain like news. The SemEval-2023 Task 2, Multilingual Complex Named Entity Recognition (MultiCoNER II) (Fetahu et al., 2023b) contains NER datasets curated for 12 languages which are low resource and from other domains as well. The task also provided a large number of 30 classes and simulated errors being added to the test set to make it more challenging. It contains the following languages: English, Spanish, Dutch, Russian, Turkish, Portugues, Korean, Farsi, German, Chinese, Hindi, and Bangla.

With the advent of Deep Learning based models, Transformer models like BERT, and bidirectional LSTM based ELMo became state of the art. BERT's multilingual counterpart, mBERT became state of the art in multilingual NER tasks. Another model XLM-RoBERTa(XLM-R) (Alexis Conneau and Stoyanov, 2020) has shown to outperform mBERT on low resource languages on the NER tasks. XLM-R is suitable for our task as it is pre-trained for more than 100 languages including the 4 languages which we worked upon. Therefore, we have used XLM-RoBERTa-base transformer model and fine-tuned it on each of the given 12 languages.

1.1 The Challenge

We participated in the shared task named MultiCoNER II, which is part of Semeval 2023. The task consists of building a named entity recognition system in 12 languages. There are 2 tracks to the tasks, one is the monolingual track for each language and the other is a multilingual track consisting of all languages. The condition of the competition was that one can only use the prediction of the monolingual model in the monolingual track and cannot use it for the multilingual track

The task is a successor to the 2022 challenge of MultiCoNER shared task, (Malmasi et al., 2022b) where the key challenges were the following 1) syntactically complex entities like movies titles e.g Shawshank Redemption 2) these complex entities having low context and 3) long-tail entity distributions.

These challenges of NER for recognizing complex entities and in low-context situations were mentioned by Meng et al. (2021). The authors mention that for complex entities, some particular types (e.g creative works) can be linguistically complex. They can be complex noun phrases (Eternal Sunshine of the Spotless Mind), gerunds (Saving Private Ryan), infinitives (To Kill a Mockingbird), or full clauses (Mr Smith Goes to Washington). For long-tail entities, they mention that many domain entities have large long-tail distribution, with potentially millions of values (e.g., location names). It's very hard to build representative training data, as it can only cover a portion of the potentially infinite possible values.

The current MultiCoNER II challenge, expands the previous challenge of 2022 with new tasks and

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 453–456

it emphasize the shortcomings of the current top models that are transformers based and depends on knowledge bases. It focusses on challenges like out of knowledge-base entities and noisy scenarios like the presence of spelling mistakes and typos.

2 Related Work

The earliest use of the neural networks to solve NER was by (Hammerton, 2003) who attempted to use an LSTM network to solve the problem. Recent transformer networks like BERT (Jacob Devlin and Toutanova, 2018) and ELMo (Matthew E. Peters and Zettlemoyer, 2018) have further led to the state of the art results in the NER task for English language.

CRF layer (Sutton and McCallum, 2012) was also proposed to be used as a final layer for token classification.

To expand NER in a multilingual setting, (Fetahu et al., 2021) introduced a multilingual and code-mixed dataset concerning the search domain. They also used an XLM-RoBERTa model combined with gazetteers to build their multilingual NER classification system.

In a paper for the first edition of this task in 2022, XLM-RoBERTa has been applied to Hindi and English dataset and have shown to perform better than mBERT with a similar set of languages as in this current task (Sumit Singh and Tiwary, 2022).

3 Data

The dataset was first released in the Semeval task in 2022, called MultiCoNER dataset (Malmasi et al., 2022a). In the 2nd edition of the task, The organisers provided a new dataset called MultiCoNER v2 (Fetahu et al., 2023a) comprising of individual language datasets in 12 languages.

Table 1 contains the number of sentences in the training, development and test datasets per language in MulticoNER v2 dataset.

The test dataset is used for the final evaluation of the leaderboard and is further split into corrupted and non-corrupted sets. It is to be noted that the uncorrupted test set size is quite large compared to the training set for all the languages.

Table 2 contains the list of 30 entities present across all datasets and their grouping into the corresponding entity types.

Language	Train	Dev	Test-1	Test-2
BN	9708	507	0	19,859
ZH	9,759	506	5696	14,569
EN	16,778	871	74,960	21,0267
DE	9,785	512	5,880	16,334
FA	16,321	855	0	219,168
FR	16,548	857	74918	174,868
HI	9,632	514	0	18,406
IT	16,579	858	74,334	173,547
PT	16,469	854	68,822	160,668
ES	16,453	854	74,050	252,257
SV	16,363	856	69,342	161,848
UK	16,429	851	0	238,296

Table 1: Sentences per split (Train,Test,Test-1 : Test-Corrupted, Test-2: Test-Uncorrupted) per language where BN is Bangla, ZH is Chinese, EN is English, FA is Farsi, FR is French, DE is German,HI is Hindi, IT is Italian, PT is Portuguese, ES is Spanish and UK is Ukranian language

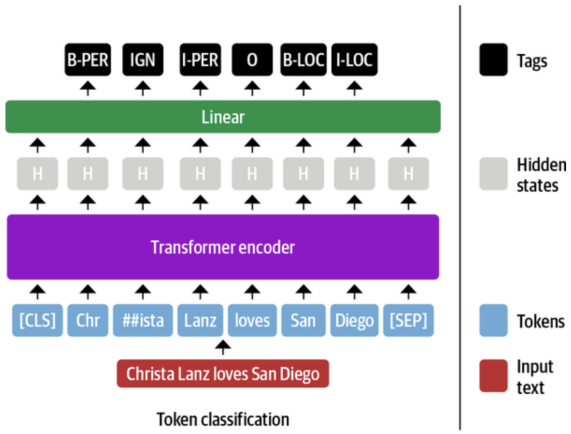

Figure 1: NER token classification using transformer encoder models like RoBERTa

4 Methodology

We utilised XLM-RoBERTa model for all 12 languages we participated in. XLM-RoBERTa is a massive Transformer trained on 100+ languages on 2TBs of CommonCrawl data.

Figure 1 describes the architecture of transformer encoder models like RoBERTa. It is trained with a multilingual MLM objective using only monolingual data. XLM-RoBERTa has already been shown to outperform mBERT on cross-lingual classification by upto 23 percent accuracy on low-resource languages. It also gave competitive results with respect to the state of the art monolingual models.

Entity Name	Entity Type
Facility, OtherLOC, HumanSettlement, Station	Location
VisualWork, MusicalWork, WrittenWork, ArtWork, Software	Creative Work
MusicalGRP, PublicCORP, PrivateCORP, ORG, AerospaceManufacturer, SportsGRP, CarManufacturer	Group
Scientist, Artist, Athlete, OtherPER Politician, Cleric, SportsManager,	Person
Clothing, Vehicle, Food, Drink, OtherPROD	Product
Medication/Vaccine, MedicalProcedure, AnatomicalStructure, Symptom, Disease	Medical

Table 2: List of Entities

4.1 Experimental Set up

We have used the PyTorch framework and HuggingFace's Tranformers library for our system.The training is done on a single GPU Nvidia Titan 2080 machine.

We used the XLM-RoBERTa-base model and fine-tuned it on each of the language datasets. We train the model in a batch size of 4 on the training dataset. A dropout of 0.2 is applied between the last hidden layer and the output layer to prevent the model from overfitting. We used a learning rate of 2e-5. For the optimizer, we used AdamW (Loshchilov and Hutter, 2019) optimizer with an epsilon of 1e-08. AdamW is a stochastic optimization method and it modifies the implementation of weight decay in Adam, by separating weight decay from the gradient update. The max length of the sequence used is 16. We trained the model on 15 epochs and select the best one based on f1-scores on the test set.

For token classification at the end, we used a linear layer as the last layer to classify into the given 30 named entities.

The final model is based on the checkpoints of the model which are selected based on best F1-scores on the development set.

5 Results

Table 3 shows the macro average scores of precision, recall and F-1.

Language	Precision	Recall	F1-Score
BN	62.18	61.60	60.46
ZH	26.51	27.34	25.50
EN	55.97	52.85	53.11
FA	53.41	54.87	53.13
FR	53.17	54.42	53.33
DE	66.56	59.92	61.62
HI	70.64	69.17	69.04
IT	60.48	62.48	60.97
PT	64.08	63.59	63.41
ES	61.02	58.19	58.32
SV	57.19	60.03	57.55
UK	55.59	49.22	49.06

Table 3: Scores of XLM-RoBERTa-base on 12 languages on development set where BN is Bangla, ZH is Chinese, EN is English, FA is Farsi, FR is French, DE is German,HI is Hindi, IT is Italian, PT is Portuguese, ES is Spanish and UK is Ukranian language

Entity-Type	HI	EN	DE	SP
Group	73.11	55.54	65.35	61.82
Medical	75.69	56.27	55.18	62.68
Person	58.87	46.99	51.80	48.66
Creative Work	74.37	59.80	66.57	64.12
Product	59.95	46.88	48.90	54.55
Location	76.15	55.08	64.17	55.86

Table 4: F1-Scores of XLM-RoBERTa-base on 4 languages by Entity types on the development set

From Table 4, we observe that the entities belonging to the Creative Work category consistently have the highest F1-score across all 4 languages. Also, it can be noted that the entities belonging to Person and Product group have the lowest F1-scores for the model.

Table 5 shows the macro average F1-scores of the final test set used for leaderboard

From table 3 and table 5, we observe that the performance for Hindi drops most from development to test set, while for English and Spanish, it just drops slightly.

Lang	Overall F1	Corrupted F1	Uncor. F1
HI	63.29	0	63.29
EN	52.08	46.3	54.73
DE	55.54	54.1	54.73
ES	54.81	49.32	57.42

Table 5: Scores of XLM-RoBERTa-base on 4 languages on the final test set where HI is Hindi, EN is English,DE is German and ES is Spanish language

Also, we couldn't submit results for languages other than Hindi,English,German and Spanish by the time the competition ended, therefore we only have F1,corrupted f1 and uncorrupted F1 scores for these languages for the test set.

6 Conclusion

In this paper, we presented using XLM-RoBERTa-base to solve the shared task of MultiCoNER.

Future work can include exploring more recent transformer-based models like XLM-V with very large vocabularies. Also, data augmentation techniques like entity replacement can be tried.

References

Naman Goyal Vishrav Chaudhary Guillaume Wenzek Francisco Guzmán Edouard Grave Myle Ott Luke Zettlemoyer Alexis Conneau, Kartikay Khandelwal 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.

Kim Sang Erik F. Tjong 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*, page 142– 147.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023a. MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition.

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.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023b. SemEval-2023 Task 2: Fine-grained Multilingual Named Entity Recognition (MultiCoNER 2). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

James Hammerton. 2003. Named entity recognition with long short-term memory. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4,Association for Computational Linguistics*, page 172–175.

Kenton Lee Jacob Devlin, Ming-Wei Chang and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint*, page arXiv:1810.04805.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization.

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.

Mohit Iyyer Matt Gardner Christopher Clark Kenton Lee Matthew E. Peters, Mark Neumann and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL 2018*.

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.

Pawankumar Jawale Sumit Singh and Uma Shanker Tiwary. 2022. Transformer based NER models for Hindi and Bangla languages. In *The 16th International Workshop on Semantic Evaluation*.

Charles Sutton and Andrew McCallum. 2012. An introduction to conditional random fields. In *Foundations and Trends® in Machine Learning*, page 4(4):267–373.