CAIR-NLP at SemEval-2023 Task 2: A Multi-Objective Joint Learning System for Named Entity Recognition

Sangeeth N[1,2], **Biswajit Paul**[2], **Chandramani Chaudhary**[1]
[1] National Insititute of Technology, Tiruchirappalli, India
[2] Centre for Artificial Intelligence and Robotics, CV Raman Nagar, Bangalore, India
406120005@nitt.edu, biswajit.cair@gov.in, chandramani@nitt.edu

Abstract

This paper describes the NER system designed by the CAIR-NLP team for submission to Multilingual Complex Named Entity Recognition (MultiCoNER II) shared task, which presented a novel challenge of recognizing complex, ambiguous, and fine-grained entities in low-context, multi-lingual, multi-domain dataset and further evaluation on the noisy subset. We propose a Multi-Objective Joint Learning System (MOJLS) for NER, which aims to enhance the representation of entities and improve label predictions through joint implementation of a set of learning objectives. Our official submission MOJLS implements four objectives. These include, representation of the named entities should be close to its entity type definition, low-context inputs should have representation close to their augmented context, and also minimization of two label prediction errors, one based on CRF and another biaffine based predictions, where both are producing distributions over the output labels. The official results ranked our system 2^{nd} in five tracks (Multilingual, Spanish, Swedish, Ukrainian, and Farsi) and 3^{rd} in three tracks (French, Italian, and Portuguese) out of 13 tracks. Also evaluation on the noisy subset, our model achieved relatively better ranks. Official ranks indicate the effectiveness of the proposed MOJLS in dealing with the contemporary challenges of NER.

1 Introduction

Named Entity Recognition (NER) is an established natural language processing task that aims to locate entity mentions by marking entity-span boundaries and further classify them into corresponding entity types from a set of pre-defined entity categories in unstructured text (Wang et al., 2022b). NER is a critical building block in Information Extraction, and Knowledge-base construction pipeline (Lample et al., 2016). The SemEval 2023 Task 2: Multilingual Complex Named Entity Recognition (MultiCoNER-II) presented a novel challenge of building a NER system covering 33 fine-grained categories over 12 languages, namely English, French, Spanish, German, Italian, Portuguese, Ukrainian, Swedish, Chinese, Farsi, Hindi, Bangla and also in multilingual setting in 13 tracks (Fetahu et al., 2023b). MultiCoNER-II task presents a host of real-world contemporary challenges in terms of (1) Low-context sentences (short and uncased text), (2) Syntactically complex and ambiguous fine-grained entities, (3) Entities having a large long-tail distribution and evolving nature, (4) Multi-linguality, (5) Evaluation on a noisy subset, where inputs are corrupted with noise either on context tokens or entity tokens, and (6) Limited training examples per language.

In this paper, we propose a Multi-Objective Joint Learning System (MOJLS) for NER, which aims to learn an enhanced representation of low-context, fine-grained entities and thereby improves recognition. We considered four training objectives. Minimization of (1) Representation gaps between the entities to the corresponding entity type definition (ETD) using KL divergence loss, where ETD is extracted from external knowledge bases, (2) Representation gaps between input sentence and input augmented with entity context, extracted through external information (3) Conditional Random Field (CRF) label prediction loss using negative log-likelihood loss function (4) Biaffine layer label prediction loss using Cross-Entropy (CE) loss function. Our multilingual NER model is trained by fine-tuning a large multilingual pre-trained language model (PLM) over the multilingual NER dataset by minimizing the combined losses from all the four loss functions. Further, monolingual NER models are trained by fine-tuning the multilingual model using the corresponding monolingual data. We make the following observations based on experiments carried out.

1. MOJLS produced improved representation

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1926–1935

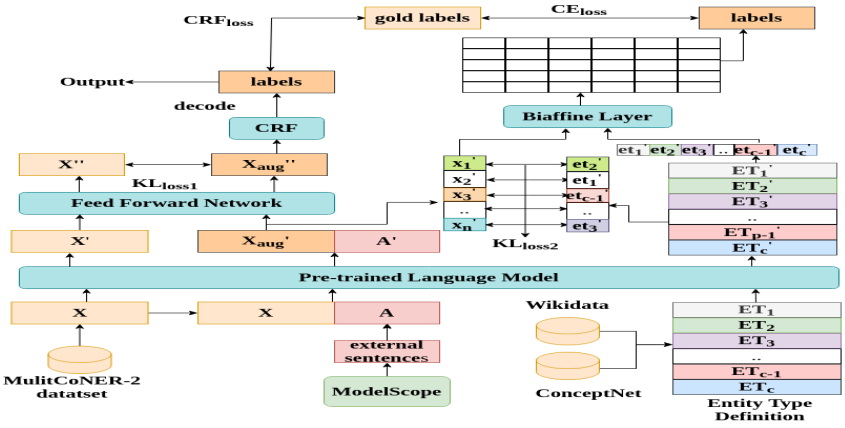

Figure 1: Architecture of our NER system

and prediction performance.

2. The objective of having the representation of entities close to its entity type definition by jointly minimizing representation gaps between an entity and the corresponding type-definition representation enhanced performance by 2.5% and made the representation of entities within an entity type close to each other.

3. MOJLS could handle noise better and generalize well. Our model showed an average degradation of 6.39% on the noisy subset compared to the clean subset in contrast to the 7.26% degradation seen in the 1st-ranked submissions.

4. Adding relevant external information to the context-deprived entities (entity context augmentation) improved performance by 7.35%.

The official results show that the CAIR-NLP system ranked 2^{nd} in five tracks (Multilingual, Spanish, Swedish, Ukrainian, and Farsi) and 3^{rd} in three tracks (French, Italian, and Portuguese) out of 13 tracks. Also, our model achieved better ranks on noisy subset evaluation, 2nd in five tracks (Spanish, Swedish, French, Italian, and Portuguese) and 3rd in one track (English) out of the 8 tracks.

Related work is presented in the appendix section.

2 Task Setup and Dataset

SemEval 2023 Task 2: MultiCoNER II Multilingual Complex Named Entity Recognition presented the challenge of developing NER systems for 33 fine-grained categories across 12 languages, namely English, French, Spanish, German, Italian, Portuguese, Ukrainian, Swedish, Chinese, Farsi, Hindi, Bangla and also multilingual setting, focusing on recognizing semantically ambiguous and complex entities in low-context setting. Table 2 and Table 3 in the appnedix section show the list of coarse, fine-grained entity types and statistics of MultiCoNER II dataset.

3 System Description

This section introduces our Multi-objective Joint Learning System for NER. The overall architecture of the proposed system is shown in Figure 1. The details of each building block are described below.

3.1 Entity Context Augmentation

The average length of the training and development data set is around 40 tokens and lacks adequate entity context. So to have a better representation entities, we suitably augmented entity context and used a search engine based text retrieval approach (Wang et al., 2021), where input sentence is placed as query, and retrieved text is collected. Further, to retrieve most semantically similar texts, BERTScore (Zhang* et al., 2020) is used to estimate the relatedness of each retrieved text to the input sentence. Finally, top-k-ranked texts are appended to the input sentence with the separator token [EOS].

$$A = [a_1, a_2, .., a_k]$$
$$XA = [X[EOS]A] \quad (1)$$

where XA is the input sentence augmented with external context, X is the input sentence, A is the related texts retrieved for augmentation, a_i represents the augmented sentence i and k is the number of sentences selected for augmentation.

To reduce retrieval time, we utilized the augmented data readily available from modelscope package[1], which implemented the above strategy and maintained a public repository of entity context for the MultiCoNER-II data set.

3.2 Entity Type Definition

To have a better representation and further disambiguation of fine-grained entities, we setup the objective that all entities belonging to a particular entity type should have representation close to each other. To realize this objective, we prepared an external knowledge base called Entity Type Definition (ETD) for each entity type capturing basic definitions, alias names, and relation types with other concepts. Each entity's definition and alternative names are taken from Wikidata, while relation - "ISAType" - are taken from ConceptNet (Speer et al., 2017). Each "ISAType" relation of the entity in ConceptNet has a weighted score assigned to it. We chose only the types with weight score greater than 1. A sample ETD for the SportsManager type is shown in Figure 2.

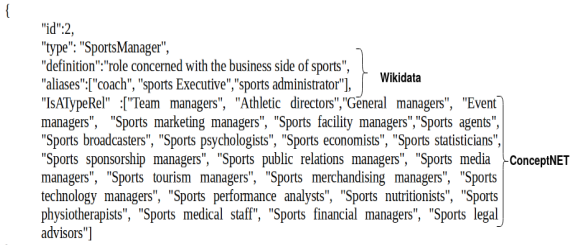

Figure 2: Entity Type Information of SportsManager

3.3 Multi-Objective Joint Learning

We propose a multi-objective joint learning strategy for the MultiCoNER-II task. A pre-trained language model(PLM) XLM-R (Conneau et al., 2020) is used to get the initial embeddings for all 13 tracks.

Let the input sentence be X of n tokens $X = x_1, .., x_n$ and the related augmented sentences be A of length k, $A = a_1, .., a_k$ extracted from modelscope package. ET is the entity type definition retrieved from Wikidata and ConceptNet. $ET = ET_1, ET_2, .., ET_c$ where c is the number of entity types, including the non-entity type (Others). We concatenate the input sentence X and related sentence A using separator token [EOS] to generate extended input $XA = X[EOS]A$. We feed the input sentence X, extended input XA and Entity Type Definition ET to the PLM individually to get the embeddings X', $X'_{aug}A'$, and $ET^{'}$. The module then feeds the embeddings X' and the input sentence embeddings X'_{aug} from $X'_{aug}A'$ into a feed-forward layer to obtain the projected embeddings X'' and X''_{aug}.

(A) Objective-1 : Representation loss minimization between input and augmented input To have a representation of input (X'') close to its augmented input (X''_{aug}), we set up the objective of minimizing KL divergence between X'' and X''_{aug}. This setting aims to make sentence representation without external context close to the representation with external context and is expected to improve model performance when no external contexts are available. This strategy is similar, though not identical, to the one used in cooperative learning methodology (Wang et al., 2021). Contrary to the cooperative learning approach, our model backpropagates the gradient for fine-tuning the representations.

$$KL_{loss_1} = D(X'', X''_{aug}) = KL(X'', X''_{aug}) \quad (2)$$

where D is the distance function, X'' and X''_{aug} are the output embeddings of the input sentence with and without external context. Softmax function is used to convert embedding values to positive numbers between 0 and 1, before applying KL loss.

(B) Objective-2: Representation loss minimization between the entity and its type

To realize the idea that all entities belonging to a particular entity type should have a representation close to each other, we set up the objective of minimizing the distance between the representations of the entity tokens and their corresponding entity types, obtained from the representation of ETD. The entity type ET_i = [CLS] defintion <EOS> alias <EOS> relation_types <EOS>. We use the [CLS] token embeddings to represent a particular entity type $et'_i = ET'_i[CLS]$. The loss is given by:

$$KL_{loss_2} = D(x'_i, et'_i) = KL(x'_i, et'_i) \quad (3)$$

where D is the distance function, the x'_i and et'_i are the representations of the token x_i, and its type ET_i, and p is the number of entity types in the dataset. In our model, KL divergence is used as the distance function, with embedding values converted between 0 and 1 using softmax function.

[1] $https://modelscope.cn/$

(C) Objective-3: Label loss minimization from Biaffine model predictions

We employ a Biaffine model (Yu et al., 2020) over the sentence and entity type embeddings to create an $l * c$ scoring tensor t_c, where the l is the length of sentence and c is the number of entity types. We compute the score by using the following equation:

$$t_c = X'_{aug} U_c E T'^T + V_c X'_{aug} + b_c \quad (4)$$

where X'_{aug} and ET' are the embeddings of the input sentence and the entity types, respectively. U_c and V_c are weight tensors of dimension $d \times d$ tensor and $d \times c$ where d is the feature dimension of the input sentence. b_c is the bias vector of dimension c, respectively. The tensor t_c provides scores for all possible types a token can be tagged. We select the type having the maximum score.

$$y(i) = argmax\ t_c(i) \quad (5)$$

The model is trained to minimize the cross-entropy loss of the correct label sequence:

$$p_c(i_c) = \frac{exp(t_c(i_c))}{\sum_{j=1}^{c} exp(t_c(i_j))}$$

$$CE_{loss} = -\sum_{i=1}^{n}\sum_{j=1}^{c} y_{i_c}\ log\ p_c(i_c) \quad (6)$$

(D) Objective-4: Label loss minimization from CRF predictions To setup the CRF label error minimization objective, the representation X''_{aug} is fed to linear chain CRF layer to obtain the conditional probability $p_\theta(y|Xaug'')$. CRF's, conditional probability is modeled by defining a feature map that maps an entire input sequence $Xaug''$ paired with an entire state sequence y to some d-dimensional feature vector. Then we can model the probability as a log-linear model with the parameter vector. For a series of tokens $Xaug'' = (x''_1, x''_2, .., x''_n)$, we obtain a series of predictions $y = (y_1, y_2, ..., y_n)$

As described in (Lample et al., 2016), the score of the entire sequence is defined as

$$S(X''_{aug}, y) = \sum_{i=0}^{n} T_{y_i, y_{i+1}} + \sum_{i=0}^{n} P_{i,y_i} \quad (7)$$

where element T_{ij} signifies the transition score from the i^{th} label to the j^{th} label and P_i signifies the emission scores from the i^{th} token to the j^{th} label. The two additional states in T are a sequence's start and end states.

The model is trained to maximize the log probability of the correct label sequence:

$$log(p(y|X''_{aug})) = S(X''_{aug}, y) - log(\sum_{\tilde{y} \in Y_x} e^{S(X''_{aug}, \tilde{y})}) \quad (8)$$

$$CRF_{loss} = log(p(y|X''_{aug})) \quad (9)$$

where Y_x are all possible label sequences in the dataset.

The total loss in training is the summation of all the loss in Eq.2, 3, 6 and 9

$$Total_{loss} = KL_{loss_1} + KL_{loss_2} + CE_{loss} + CRF_{loss} \quad (10)$$

and the overall training objective is to minimize the total loss.

4 Experimental Setup

4.1 Data Resources

The official data set (Fetahu et al., 2023a) is only used to train our NER models for all 13 tracks. Table 3 in the appendix section shows statistics of MultiCoNER II train, dev and test sets. For context augmentation, the modelscope package is used. We also created an Entity Type Definition knowledge base for each entity using Wikidata and ConceptNet.

4.2 Training

The MultiCoNER-II task comprises training NER models covering 12 languages along with a multilingual model.

Base Model (Configuration 1) : Our base model is trained with CRF loss and adapting multistage fine-tuning strategy (Wang et al., 2022a), where we initially trained the multilingual model. Then the monolingual models are trained by fine-tuning the best multilingual model. This training strategy aided in reducing of training time and enhanced model performance. The multilingual model is trained for 30 epochs with an early stop criterion after the 10^{th} epoch, and the monolingual models are trained for only 5 epochs over the multilingual model. The early stopping criterion stops the training when the performance on the validation set continues to degrade for 5 epochs.

Lang/ Dataset	EN	ES	SV	UK	PT	FR	FA	DE	ZH	HI	BN	IT	MULTI
Clean Subset F1	81.29	85.03	84.54	81.29	81.73	84.67	77.5	74.71	62.89	72.23	69.46	85.08	79.16
Noisy Subset F1	74.89	80.66	79.75	-	77.1	79.54	-	-	44.74	-	-	81.0	-
Overall F1	79.33	83.63	82.88	81.29	80.16	83.08	77.5	74.71	58.43	72.03	69.46	83.78	79.16
Overall Rank (Official)	4	2	2	2	3	3	2	7	13	8	10	3	2
Rank on Noisy Subset (Official)	**3**	2	2	-	**2**	**2**	-		**10**	-	-	**2**	-

Table 1: Official evaluation results on clean, noisy subsets, overall macro F1 score and ranks

Context Augmentation (Configuration 2): We enhanced our base model by adding context information and minimizing the representation gaps between input and input augmented with context, using KL divergence loss.

Entity Type Definition and Biaffine predictions (Configuration 3): We further enhanced our model by adding Entity Type Definition and the learning objective that all entities belonging to a particular entity type should have a representation close to each other and used KL divergence loss along with additional biaffine-based label prediction loss. The hyper-parameters used for the training are shown in Table 4 in the appendix section.

4.3 Evaluation

The evaluation measure used by the task organizer is the macro averaged F1 score at the entity level, unlike CoNLL NER dataset. Also, evaluation is carried out for clean and noisy subsets, and teams are ranked based on the overall macro F1 scores.

5 Results and Analysis

5.1 Official Results

CAIR-NLP team's official results on the SemEval 2023 Task 2 : MultiCoNER II for all the 13 tracks are captured in Table 1. The official results placed our system 2^{nd} in five tracks (Multilingual, Spanish, Swedish, Ukrainian, and Farsi) and 3^{rd} in three tracks (French, Italian, and Portuguese) out of 13 tracks.

5.2 Performance on Clean and Noisy Subsets

Language-wise performance on the official clean, noisy evaluation subsets and overall macro F1 scores are captured in Table 1. Official results (Fetahu et al., 2023b) show that our model performed relatively better on noisy subsets than other teams. On account of performance degradation on the noisy subset as compared to the clean subset, our model showed an average degradation of 6.39% as compared to the 7.26% degradation seen in the top-ranked submissions.

Due to space constraint, extended analysis of the results is presented in the appendix section.

6 Conclusion

For low-context, complex, ambiguous, and often noisy entities, relying solely on pre-trained language models to achieve competitive NER performance is insufficient. Enhancement of entity context and refined representation of fine-grained entities using external knowledge bases and information sources are critical for achieving enhanced performance. A multi-objective joint learning strategy with suitable objectives for the overall label predictions along with set of intermediate objectives can enhance the performance. Specifically for MultiCoNER II shared task, context augmentation improved the performance by 7.47%. The objective of having a representation of all fine-grained entities close to each other within an entity type, additionally enhanced performance by 2.78% and aided well in dealing with input noise and better generalization. Adding other suitable objectives in our MOJLS formulation, better augmentation strategies, and use of external knowledge bases are likely to enhance performance further.

Acknowledgements

The authors would like to thank Prasanna Kumar KR, Chitra Viswanathan, Prashant Banjare, Abhinav Mishra, Pavanpankaj Vegi, Sivabhavani J, Dinakara K for their inputs and Director, CAIR and DG (MCC) for their constant support and encouragement.

References

Beiduo Chen, Jun-Yu Ma, Jiajun Qi, Wu Guo, Zhen-Hua Ling, and Quan Liu. 2022. USTC-NELSLIP at SemEval-2022 task 11: Gazetteer-adapted integration network for multilingual complex named entity

recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1613–1622, Seattle, United States. Association for Computational Linguistics.

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.

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, 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.

Abbas Ghaddar and Phillippe Langlais. 2018. Robust lexical features for improved neural network named-entity recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1896–1907, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1800–1812, Online. Association for Computational Linguistics.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, Wei Lu, and Yong Jiang. 2022a. DAMO-NLP at SemEval-2022 task 11: A knowledge-based system for multilingual named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1457–1468, Seattle, United States. Association for Computational Linguistics.

Yu Wang, Hanghang Tong, Ziye Zhu, and Yun Li. 2022b. Nested named entity recognition: A survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(6):1–29.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Appendix

6.1 Dataset

MultiCoNER II dataset comprises of 6 coarse-grained and 33 fine-grained categories across 12 languages. Table 2 show the list of coarse, fine-grained entity types. The statistics of the train, dev and test data sets are given in Table 3. The evaluation test set comprised both clean and noisy subsets. Noisy subsets include sentences corrupted with noise on context or entity tokens.

Coarse Level	Fine Level
Person (PER)	Scientist, Artist, Athlete, Politician, Cleric, SportsManager, OtherPER
Product (PROD)	Drink, Food, Vehicle, Clothing, OtherPROD
Group (GRP)	MusicalGRP, ORG, PublicCorp, SportsGRP, AerospaceManufacturer, CarManufacturer, PrivateCorp
Creative Work (CW)	VisualWork, WrittenWork, MusicalWork, Software, ArtWork
Location (LOC)	Facility, HumanSettlement, Station, OtherLOC
Medical (MED)	Medication/Vaccine, MedicalProcedure, AnatomicalStructure, Symptom, Disease

Table 2: List of coarse and fine-grained entity types

Code	Language	Train	Dev	Test
BN	Bangla	9,708	507	19,859
DE	German	9,785	512	20,145
EN	English	16,778	871	2,49,980
ES	Spanish	16,453	854	2,46,900
FA	Farsi	16,321	855	2,19,168
FR	French	16,548	857	2,49,786
HI	Hindi	9,632	514	18,399
IT	Italian	16,579	858	2,47,881
PT	Portuguese	16,469	854	2,29,490
SV	Swedish	16,363	856	2,31,190
UK	Ukrainian	16,429	851	2,38,296
ZH	Chinese	9,759	506	20,265
Multi	Multilingual	1,70,824	8,895	3,58,688

Table 3: Statictics of MultiCoNER 2 dataset

6.2 Hyper-Paramters

The CAIR-NLP (official) used the hyperparameters described in Table 4 for monolingual models. The CAIR-NLP (Updated) used the exact configuration of the multilingual model with an early stopping criterion. The updated model shows an average improvement of 3.44% for the low-training resource languages (Hindi, Bengali, Chinese, and German).

Parameters		Multilingual	Monolingual
learning rate	xlm-roberta	2e-5	2e-6
	biaffine layer	5e-3	5e-3
	crf layer	5e-2	5e-2
Optimizer		AdamW	
Training Epochs		30	5
Train Batch Size		16	16
Eval Batch Size		256	256
Maximum Sequence Length		512	512
Dropout		0.5	0.5

Table 4: Major Hyperparameters of our model

6.3 Extended Analysis fof results

This section covers effect of different learning objectives used.

6.3.1 Effect of Entity Type Definition

The idea of using Entity Type Definition (ETD) is to have a better representation of fine-grained entities in the MultiCoNER 2 data set. Table 6 shows that this strategy improves the performance by 2.78% on top of the context augmentation. We also analyzed the result of using ETD without external context on English, Spanish, and Multilingual data sets. The experiments are done using the XLM-Roberta base model. The results captured in Table 5 show an average improvement of only 2.4%. The potential reason could be due to the low context of the input sentence; the ETD-based representation strategy alone failed to compensate for the context. However, the ETD-based strategy could do well with the augmented context.

Model/Lang	EN	ES	MULTI
Baseline	65.05	69.84	68.16
W/ETI and W/o EC	67.55	72.31	70.52
W/ ETI and EC	79.63	87.35	84.87

Table 5: The result compares the usage of Entity Type definition with and without external context on English, Spanish and multilingual dataset.

6.3.2 Effect of Context Augmentation

Relative improvements achieved by including different configurations implementing different objectives and associated loss functions are presented below. This evaluation was done on the development set. Table 6 shows an average improvement of 7.47% in F1 scores when using the external context

Lang/Model	Loss/ Objectives	EN	ES	SV	UK	PT	FR	FA	DE	ZH	HI	BN	IT	MULTI
Our base model	CRF_{loss}	73.72	78.84	81.28	76.43	79.28	77.36	68.48	79.14	71.59	85.44	84.4	81.09	76.06
+ External Context	CRF_{loss}	81.34	87.27	85.68	85.56	87.69	85.62	79.25	81.16	75.88	88.12	88.22	88.36	84.26
	CRF_{loss}+ KL_{loss1}	82.11	88.17	86.55	86.35	88.57	86.47	80.07	82.02	76.57	89.48	89.35	89.52	85.01
+ Entity Type Def + Biaffine	$Total_{loss}$	84.18	90.95	89.85	89.09	90.39	88.01	87	84.19	79.65	91.87	92	91.24	87.58

Table 6: Results capturing effects of External Context(EC), Entity Type Definition (ETD) and Bi-affine Label Predictions and associated objectives/losses

Lang/Coarse Entity Type	EN	ES	SV	UK	PT	FA	FR	DE	HI	BN	ZH	IT	MULTI
LOC	93.96	94.77	97.46	94.44	95.43	91.25	93.57	89.07	89.42	89.07	82.43	95.04	92.05
PER	97.29	97.78	98.07	97.09	97.32	92.42	97.6	93.6	90.57	89.33	89.35	98.32	95.08
Medicine	88.84	93.64	92.36	90.28	91.83	88.75	91.66	81.62	81.29	80.43	70.7	92.88	87.1
GRP	90.88	92.73	93.06	91.59	92.45	88	91.59	86.85	88.97	82.25	77.7	93.29	88.26
PROD	83.24	89.95	90.34	89.22	89.79	85.6	88.69	74.36	70.44	65.3	64.59	90.27	83.1
CW	91.73	93.2	94.04	91.05	93.56	89.54	93.77	85.95	77.7	76.35	73.07	96.52	89.19

Table 7: Result for each coarse entity type for each language and multilingual

Lang/Fine Grained Entity Type	EN	ES	SV	UK	PT	FA	FR	DE	HI	BN	ZH	IT	MULTI
AerospaceManufacturer	84.65	81.54	75.86	70.89	60.41	84.49	83.27	84.45	28.32	29.63	64.03	64.72	71.99
Athlete	86.96	85.2	87.59	86.33	86.09	72.54	87.38	79.9	82.08	72.43	73.25	91.81	83.48
Facility	82.65	84.72	87.94	84.59	86.66	82.46	84.63	72.38	67.87	73.86	66.82	87.87	80.42
PublicCorp	78.29	90.03	85.66	90.38	89.71	77.31	84.89	72.33	79.25	77.49	62.67	88.51	81.3
CarManufacturer	84.74	92.45	85.19	87.82	88.77	85.47	88.25	68.5	82.71	81.66	63.37	89.54	84.58
MedicalProcedure	85.56	90.97	87.14	85.45	89.83	84.97	88.38	82.27	79.41	77.68	65.78	90.53	82.94
Cleric	70.48	77.68	73.79	72.62	80.32	67.64	75.77	60.37	75.89	67.84	48.22	82.87	72.6
Vehicle	75.67	83.28	82.72	82.57	83.43	77.32	81.68	70.85	76.88	69.74	65.91	81.04	76.88
MusicalGRP	87.69	91.05	92.4	92.32	90.09	84.09	91.01	82.64	81.17	77.47	68.71	93.72	85.75
Station	90.58	92.42	93.81	90.23	93.42	93.82	93.82	83.25	88.72	88.96	82.94	93.76	88.85
Politician	72.25	75.37	78.41	68.72	76.85	70.83	75.67	66.07	73.79	69.18	51.74	76.14	70.7
HumanSettlement	95.5	95.61	98.04	95.02	96.13	91.97	94.26	91.58	91.06	89.59	84.19	95.66	93
Drink	82.25	86.9	87.76	85.39	90.12	82.4	87.31	58.33	70.83	79.67	54.4	88.68	81.74
OtherPER	60.73	67.8	66.28	64.5	67.04	59.86	64.39	57.66	55.29	48.72	48.78	67.38	59.06
Artist	86.8	86.79	86.39	82.63	88.54	83.76	88.67	80.72	78.47	75.64	73.5	91.84	84.28
Medication/Vaccine	88.03	92.92	90.95	88.39	91.66	88.29	90.57	82.4	79.32	80.53	68.78	91.42	85.77
Clothing	78.97	82.11	82.84	79.27	79.22	58.71	81.67	70.42	62.67	35.82	52.8	80.09	71.72
OtherPROD	76.89	83.04	86.27	83.37	87.4	81.36	81.17	69.73	64.72	60.95	59.74	83.82	77.89
SportsManager	69.56	70.29	70.44	71.76	74.5	68.94	73.65	60.97	55.93	57.73	51.03	76.8	67.21
ArtWork	83.36	76.82	63.68	55.61	46	27.63	83	77.63	26.99	17.83	49.39	87.56	60.75
OtherLOC	72.63	72.78	97.49	78.81	86.06	59.63	73.34	64.62	73.47	65.56	50.48	74.57	79.33
AnatomicalStructure	87.41	92.27	92.23	90.85	90.42	87.05	88.41	75	76.38	77.12	69.63	91.84	85.46
WrittenWork	86.64	88.36	91.33	87.76	87.47	82.57	90.4	81.24	78.32	78.34	72.6	88.72	84.69
MusicalWork	89.74	89.37	90.73	85.52	90.15	83.31	90.12	84.13	48.12	56.56	57.7	94.15	85.43
Software	89.07	94.69	94.13	94.01	94.44	85.77	93.03	83.67	82.55	84.84	69.73	93.82	88.5
Food	78.91	87.85	86.27	85.4	86.25	82.94	85.35	60.64	72.02	58.96	59.76	89.14	78.11
PrivateCorp	68.21	80.31	72.34	47.37	7.06	73.53	85.18	77.99	70.42	84.62	69.51	57.67	72.11
SportsGRP	93.57	94.8	96.27	94.26	94.4	91.08	92.54	89.81	94.7	92.41	82.57	94.49	91.42
VisualWork	89.76	90.38	93.06	90.59	91.01	91.07	95.39	84.93	80.27	75.6	67.44	97.12	88.79
Disease	85.9	91.82	89.78	88.21	90.65	86.49	89.04	81.12	82.03	78.61	70.08	90.03	85.26
Symptom	79.01	83.84	77.32	77.65	81.2	78.96	86.27	67.86	73.68	76.02	39.08	83.08	76.54
ORG	81.01	85.75	87.06	84.39	85.42	79.57	82.74	75.85	86.25	77.26	67.69	83.27	79.93
Scientist	59.18	66.71	58.56	59.87	66.31	51.53	62.86	48.68	63.92	53.95	42.98	66.15	55.79

Table 8: Result for each fine grained entity type for each language and multilingual

compared to the baseline model on the development set. External augmentation of entity context significantly improved the performance of NER in context-deprived MultiCoNER-II data set. The addition of KL_{loss1} shows an average improvement of 0.91% in F1 scores compared to the model using external context with the CRF loss alone. As all the test sentences are context augmented, so effect

of KL_{loss1} function is not quite apparent.

6.3.3 Performance on Coarse and Fine-grained Entities

Table 7 and 8 above show detailed results of our model on coarse and fine-grained entity types. Our model performed well on coarse entity types having an average of more than 88% for 4 entity types out of 6. Specifically, Location (LOC) and Person(PER) achieved an F1 score of 92% and 95%, respectively. 19 out of 33 fine-grained entity types have an average F1 score of more than 80%. Human Settlement and SportsGroup entity types have an F1 Score of 93% and 92.4%, respectively.

Language	Code	No. of Team participated	Overall Rank
Multilingual	MULTI	18	2
Spanish	ES	18	2
Swedish	SV	16	2
Ukrainian	UK	14	2
Farsi	FA	14	2
French	FR	17	3
Portuguese	PT	17	3
Italian	IT	15	3
English	EN	34	4
German	DE	17	7
Hindi	HI	17	8
Bangla	BN	18	10
Chinese	ZH	22	13

Table 9: Official ranking of CAIR-NLP team on SemEval 2023 Task 2 : MultiCoNER II

6.4 Comparision with other teams

Among all the teams participated as summarized in Table 9, we selected to compare our results with the DAMO-NLP, PAI, USTC-NELSLIP, NLPeople, and IXA/Cogcomp teams. Table 10 shows the comparative figures of our model with these six teams. Detailed results are available on the official site [2] and also analysis of teams' relative performance and other insightful details and findings in the task description paper (Fetahu et al., 2023b). The CAIR-NLP (Updated) used the exact configuration of the multilingual model with an early stopping criterion. The updated model shows anaverage improvement of 3.44% for the low-training resource languages (Hindi, Bengali, Chinese, and German)

6.5 Related Work

Deep neural models have produced state-of-the-art performance on the traditional benchmark NER datasets like CoNLL03/OntoNotes (Peters et al., 2018; Ghaddar and Langlais, 2018). Pre-trained language models and the conditional random fields layer are dominantly used (Devlin et al., 2019; Yamada et al., 2020). A graph based novel approach was proposed for addressing the challenge of nested entity recognition (Yu et al., 2020). It used graph-based dependency parsing and a biaffine model to score start and end token pairs and provide a global view of the input and enhanced NER predictions.

Contemporary NER poses additional set of challenges as highlighted by the MultiCoNER (Malmasi et al., 2022a,b) and MultiCoNER-II (Fetahu et al., 2023a,b) shared tasks and involves the detection of semantically ambiguous and complex named entities in low-context setting across multiple languages and sometimes in code mixed setting (Fetahu et al., 2021). A knowledge-based system (Wang et al., 2022a) was proposed for the MultiCoNER, which used Wikipedia to build a multilingual knowledge base for providing relevant context information to entities and achieved state-of-the-art results on MultiCoNER. For a given input sentence, it searches the knowledge base for related context and appends this information to the input sentence. As a result, contextualized token representation and entity recognition have shown significant improvement. Earlier search engine-based context retrieval was proposed (Wang et al., 2021) that showed recognition improvement compared to systems without external context. Also, when no external contexts are available, this model could enhance performance using cooperative learning strategy (Wang et al., 2021). USTC-NELSLIP (Chen et al., 2022) developed a gazetteer-adapted integration network for solving MultiCoNER task. The method begins by adapting the representations of gazetteer networks to those of language models by minimizing the KL divergence between them. These two networks are then combined for backend-supervised NER training after adaptation.

We used external information to enhance entity context, aiding in disambiguation as in (Wang et al., 2021) (Wang et al., 2022a). In addition to that, we made use of an external knowledge base to get information about the entity types instead of using a gazetteer as in (Chen et al., 2022). We also used the Biaffine model to get prediction scores between words and the entity type instead of finding the tags' start and end as in (Yu et al., 2020). Finally, we set up a multi-objective joint learning task for better representation and prediction.

[2] https://multiconer.github.io/results

Lang/ Model	EN	ES	SV	UK	PT	FR	FA	DE	ZH	HI	BN	IT	MULTI
DAMO-NLP	**83.33**	**89.78**	**89.57**	**89.02**	**85.97**	**89.59**	**87.93**	84.97	**75.98**	78.56	81.6	**89.79**	**84.48**
PAI	80.0	71.67	72.38	71.28	81.61	86.17	68.46	**88.09**	74.87	80.96	**84.39**	84.88	77.0
USTC-NELSIP	72.15	74.44	75.47	74.37	71.26	74.25	68.86	78.71	66.96	**82.14**	80.59	75.7	75.62
IXA/Cogcomp	72.82	73.81	76.54	75.25	72.28	74.25	69.49	80.35	64.86	79.56	78.95	74.67	78.17
NLPeople	71.81	72.76	75.08	73.41	70.16	72.85	70.76	77.67	65.96	78.5	78.24	73.71	78.38
CAIR-NLP(Official)	79.33	83.63	82.88	81.29	80.16	86.17	77.5	74.71	58.46	72.23	69.46	83.78	79.16
CAIR-NLP(Updated)	79.56	84.19	83.52	81.57	81.19	83.38	79.14	76.53	61.73	76.23	73.9	84.79	81.80
Our Model Rank	4	2	2	2	3	3	2	7	13	8	10	3	2

Table 10: Part of the official results. The bold score represents the top-rank model in each language