silp_nlp at SemEval-2023 Task 2: Cross-lingual Knowledge Transfer for Mono-lingual Learning

Sumit Singh and **Uma Shanker Tiwary**
Indian Institute of Information Technology, Allahabad
sumitrsch@gmail.com
ust@iiita.ac.in

Abstract

Our team silp_nlp participated in SemEval2023 Task 2: MultiCoNER II. Our work made systems for 11 mono-lingual tracks. For leveraging the advantage of all track knowledge we chose transformer-based pretrained models, which have strong cross-lingual transferability. Hence our model trained in two stages, the first stage for multi-lingual learning from all tracks and the second for fine-tuning individual tracks. Our work highlights that the knowledge of all tracks can be transferred to an individual track if the baseline language model has cross-lingual features. Our system positioned itself in the top 10 for 4 tracks by scoring 0.7432 macro F1 score for the Hindi track (7th rank) and 0.7322 macro F1 score for the Bangla track (9th rank).

1 Introduction

Multilingual Complex Named Entity Recognition) (Fetahu et al., 2023b) targets the recognition of complex named entities in multilingual and multi-domain texts, encouraging researchers to develop new approaches to extract diversified entity types. The semantic structure of data (Fetahu et al., 2023a), which is also provided by organizers for training and validation and testing purposes, has six types of coarse-grained entities: person, location, group, product, medical and creative work with more than 30 fine-grained entities of these 6 coarse-grained entities. The dataset was provided for 12 mono-lingual tracks, and, in addition, this task also provided multi-language data encouraging the development of more generic and adaptive systems. NER task is used in many other NLP tasks. Our team participated in 11 mono-lingual tracks (Hindi, Bangla, English, German, Spanish, French, Italian, Portuguese, Swedish, Ukrainian, and Chinese). Our system scores 7th rank in Hindi and is secure in the top 10 in 4 languages (Fetahu et al., 2023b).

Our system leveraged the advantage of cross-lingual learning through multi-stage training similar to (Wang et al., 2022) and (He et al., 2022). For this in the first stage, we have selected Cross-lingual pretrained models which cover a group of languages of our task and we fine-tune these pretrained language models for a combined dataset of all tracks in which languages are covered by these language models for 10 epoch. It creates a model checkpoint which has knowledge of multiple languages for our task. In the second stage, we fine-tuned the model checkpoints which are generated in the first stage for each track individually. For the Hindi and Bangla tracks our best model was achieved by Muril large model (Khanuja et al., 2021) which is combinedly fine-tuned for three tracks (English, Bangla, Hindi) in the first stage and thereafter the model checkpoints fine-tuned for Hindi and Bangla. For the best model for all other languages, we fine-tuned xlm-r (Conneau et al., 2020) for all the tracks combinedly in the first stage and in the second stage fine tuned all mono-lingual tracks.

In addition to multilingual, there were two key challenges in this year's task, fine-grained entity taxonomy with over 30 different classes, and noisy testing data which is generated by adding simulated errors to the test set. Our best prediction model is available here[1].

2 Related Work

Transformer-based language models like Bert (Devlin et al., 2019) shows the competitive result for several NLP tasks like NER but the problem of low-context situations, long-tail entities, emerging entities and complex entities outlined by the Meng et al. (2021) hard to handle by these models since these models generate contextual word embeddings and pretrained on existing data only.

[1]https://huggingface.co/sumitrsch

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1183–1189

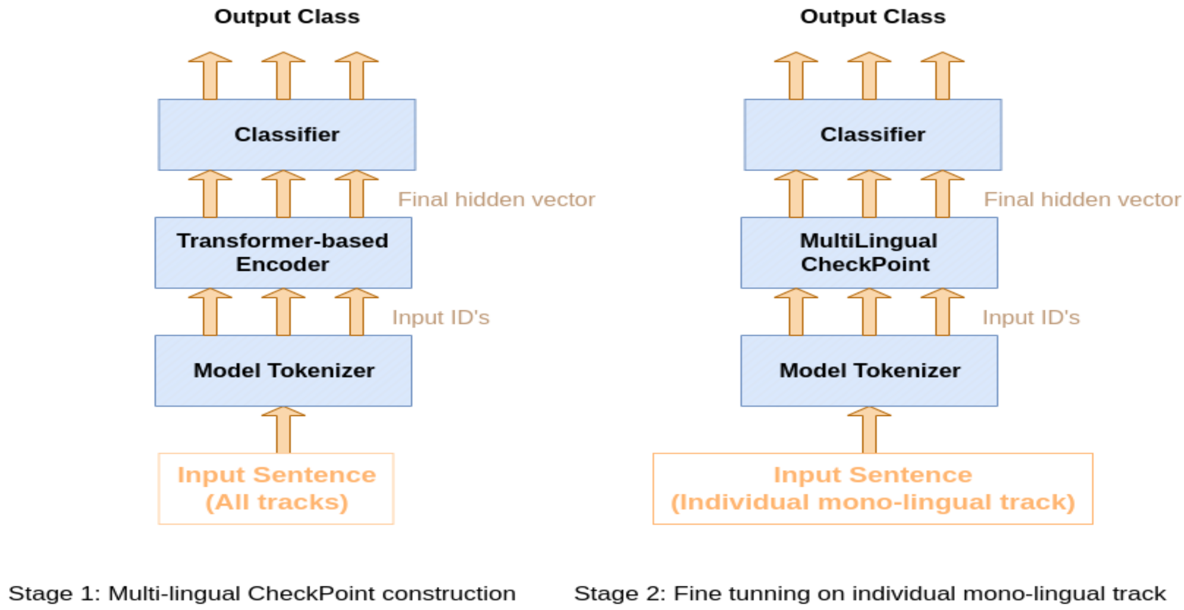

Figure 1: General Architecture of our system

The problem of emerging entities for Web queries in multi-Lingual Code-Mixed context addressed by (Fetahu et al., 2021). These challenges also occur in (Malmasi et al., 2022b) and (Fetahu et al., 2023b). (Fetahu et al., 2021) and Meng et al. (2021) integrated a gazetteer with language models as a solution for these challenges. There are some challenges with the use of a gazetteer, one of them is gazetteer should cover the entities in a sufficient ratio and also for new domain and entity types a specific gazetteer is required, which is subject to the availability of the gazetteer. Another challenge described in Meng et al. (2021) is weak learning of the model due to the model knowing the entity type in advance for the entities, which are covered by the gazetteer.

Our work is the continuation of the multilingual NER task started in 2022 (Malmasi et al., 2022b), which has only 6 entities present for 11 mono-lingual tracks with the multi-language and code-mixed track (Malmasi et al., 2022a). The Challenge of (Malmasi et al., 2022b) was to predict complex entities like Creative works with low context for the multi-lingual track. One of the leading approaches to generate a system for (Malmasi et al., 2022b) the problem is based on a Knowledge-based System (Wang et al., 2022) for lower context examples by augmentation of text from Wikipedia. (Chen et al., 2022) also performed well for all tracks by integrating language models like BERT (Devlin et al., 2019) with a gazetteer network by minimizing the KL divergence. (He et al., 2022) and (Wang et al., 2022) apply multi-stage training for leveraging the advantage of multi-lingual tracks. (He et al., 2022) also used Ensemble architecture and data augmentation for making a more robust model for noisy environments. (Singh et al., 2022) fine tuned muril-large and achieve 0.69 and 0.59 F1-Scores for Hindi and Bangla tracks.

3 Data

Language	Train Size	Dev Size	Test Size
Bangla	9708	507	19859
German	9785	512	20145
English	16778	871	249980
Spanish	16453	854	246900
French	16548	857	249786
Hindi	9632	514	18399
Italian	16579	858	247881
Portuguese	16469	854	229490
Swedish	16363	856	231190
Ukrainian	16429	851	238296
Chinese	9759	506	20265
Farsi	16321	855	219168
Multi	170823	8894	358667

Table 1: Data set distribution

Hyper parameters	Muril setup	xlm-r setup
Baseline language model for first stage	google/muril-large-cased	xlm-roberta-large
Hidden size for language model	1024	1024
Classification layer	Linear layer with cross-entrophy loss function	Linear layer with cross-entrophy loss function
Learning rate for language models	5e-06	1e-06
Learning rate for the classification layer	5e-06	1e-06
First-stage training epochs	10	10
Second-stage training epochs	20	20
Batch size	64	64
Dropout rate	0.1	0.1
Activation function	Relu	Relu
Optimizer	AdamW	AdamW

Table 2: Hyper-parameters for Muril and xlm-r setups

This shared task provide 11 mono-lingual track and 1 multi-lingual track (Fetahu et al., 2023a) to encourage NLP researchers to develop NER systems. All tracks have split training, development and testing data however testing data is provided after the final ranking is done. The size of training, development and testing data of all tracks are tabulated in Table 1.

Along with complex entities like creative works, there are 6 coarse-grained entities and more than 30 fine-grained entities are there for all tracks, which are challenging as they are harder to recognize. Organizers (Fetahu et al., 2023a) provided around 9700 sentences only for training for Hindi, Bangla, Chinese and German tracks, this made these tracks more challenging than other tracks as for other mono-lingual tracks around 16500 training data were available.

4 Methodology

All track data are divided into training and validation data by the task organizer. Similar to (He et al., 2022) and (Wang et al., 2022) our system selected strong cross-lingual transfer models (XLM-R (Conneau et al., 2020), pre-trained on more than 100 languages and Muril (Khanuja et al., 2021), pre-trained on all Indic and English languages) for two-stage training as the base model. In the first stage, we performed multi-lingual training in which training data of all tracks were utilized in selected language models. Figure 1 showed that in the first stage, our model utilized the annotations of all the tracks. This experiment was done with different hyperparameters for 10 epochs and the best multi-lingual checkpoint was selected based on average loss on validation data. Thereafter in the second stage, we use the fine-tuned multilingual checkpoint of the first stage as the initial model for fine-tuning each mono-lingual track. In the second stage, training of each mono-lingual track is performed with different hyperparameters and we select the final model based on the loss of validation data of the corresponding track.

As shown in Figure 1, first of all, the model tokenizer tokenizes the input sentence further preprocessing is done like (Singh et al., 2022). model_max_length of tokenizers set to 92 for better GPU utilization. Thereafter language model generates word embeddings for each token. A linear layer is applied on each token embedding that projects this embedding into logits, a vector of 67 dimensions since the total unique annotations in the dataset are 67. Further, for classification, we have used the cross-entropy loss function.

5 Experimental setup

Our best score was achieved with Muril setup for the Hindi and Bangla track and the other track's best score was achieved with the XLM-R setup. Detail about both the setup defined in the next subsection. Implementation of our task with the Cross entropy classification layer is done by xxxtoken-classificaion class defined in (Wolf et al., 2020), where xxx refers to the selected model.

Model	BN	HI	EN	ES	UK	IT	FR	PT	ZH	DE	SV
xlm-r (baseline)	14.00	21.00	42.18	37.89	46.98	48.0	44.87	50.16	10.0	27.38	49.60
muril (baseline)	68.00	71.00	46.85	-	-	-	-	-	-	-	-
xlm-r setup	62.37	64.25	**60.8**	**62.9**	**63.18**	**63.11**	**62.39**	**61.05**	**51.65**	**64.92**	**65.0**
muril setup	**73.22**	**74.32**	54.88	-	-	-	-	-	-	-	-

Table 3: F1 score(%) for xlm-r setup, Muril setup, xlm-r (baseline) and muril large (baseline) of 11 tracks.

5.1 Hindi and Bangla track

Muril model selected for Hindi and Bangla track referring to (Singh et al., 2022). Muril (Khanuja et al., 2021) pretrained on text in 17 languages which include English, Hindi and Bangla with explicitly augmented monolingual text corpora with translated and transliterated document pairs. Therefore for this setup, in the first stage, we combined the English, Hindi and Bangla track datasets for training and validation. Based on validation error after fine-tuning for 10 epochs, Muril multi-lingual checkpoint was constructed. Further, in the second stage, Muril multi-lingual checkpoint is utilized for fine-tuning three mono-lingual tracks separately. Again best model was selected by validation loss after 20 epochs of training. We trained the model in both stages with 1e-6, 5e-6 and 1e-5 learning rates and 16, 32 and 64 batch sizes. Hyperparameters for the best model showed in Table 2.

5.2 Other mono-lingual track

In this setup XLM-R (Conneau et al., 2020) was selected as the baseline model since this model has strong cross-linguality transferability over more than 100 languages and it covers all mono-lingual tracks. Similar to the previous muril setup, this setup was also done in two stages. In the first stage, all tracks are combinedly and trained for 10 epochs and the xlm-r multilingual checkpoint is constructed. In the second stage xlm-r multilingual checkpoint is utilized for fine-tuning all individual mono-lingual tracks for our task. We trained mode in both stages with 1e-6, 5e-6 and 1e-5 learning rates and 16,32 and 64 batch sizes. Hyperparameters for the best model for this setup are shown in Table 2 .

6 Results and Analysis

Our best macro F1 score with official rank for 11 tracks is tabulated in Table 4. Our system scored 7th rank for the Hindi track with 74.32% F1 score and 9th rank for the Bangla track with 73.22% F1 score. Both the scores were achieved by muril setup. For all the other tracks best score was achieved by xlm-r setup.

Track	F1(FG)	F1(CG)	Rank
Hindi	74.32	87.02	7
Bangla	73.22	86.91	9
Swedish	65.0	82.54	10
German	64.92	83.08	10
Spanish	62.9	79.26	11
Ukrainian	63.18	80.46	11
Italian	63.11	80.53	12
French	62.39	77.99	12
Portugese	61.05	80.35	12
Chinese	51.65	69.59	17
English	60.85	77.34	19

Table 4: Fine-grained and course-grained macro F1 score(%)

Our model vs baseline We have also finetuned XLM-R (Conneau et al., 2020) for all tracks and Muril (Khanuja et al., 2021) for Hi, Bn and En tracks individually with the hyperparameters tabulated in Table 2. We considered these finetuned checkpoints as baseline models. Comparison between our models with baseline models is shown in Table 3. The improvement over the baseline model evidently shows that our two-stage strategy learns knowledge of multi-tracks in the first stage and transfers this learning during the second stage of training of the monolingual track.

Muril setup vs xlm-r setup for Hindi, Bangla and English Results shown in Table 3 indicate that Muril setup is better than xlm-r setup for the Hindi and Bangla track while for the English track xlm-r setup performs better compared to the Muril setup.

Clean vs noisy test data Table 6 showed that our system performs lower on noisy data compared to clean data. This might be possible because training data was structured data and we didn't augment other data for robust training.

Fine-grained vs coarse-grained Table 4 showed that the macro F1 score for coarse-grained classes is better than for fine-grained classes. This infers that the model mispredicted the fine-grained classes while correctly predicting the

Entity	Average F1-score
Facility	0.6825
OtherLOC	0.5768
HumanSettlement	**0.8814**
Station	0.7678
VisualWork	0.7756
MusicalWork	0.6692
WrittenWork	0.7235
ArtWork	0.4121
Software	0.7677
MusicalGRP	0.7700
PublicCorp	0.6533
PrivateCorp	0.2969
AerospaceManufacturer	0.4472
SportsGRP	0.8306
CarManufacturer	0.6987
ORG	0.6786
Scientist	0.4262
Artist	0.7829
Athlete	0.7590
Politician	0.5807
Cleric	0.5686
SportsManager	0.5357
OtherPER	0.4739
Clothing	0.5278
Vehicle	0.6177
Food	0.61991
Drink	0.6151
OtherPROD	0.6036
Medication/Vaccine	0.7503
MedicalProcedure	0.6887
AnatomicalStructure	0.7186
Symptom	0.4471
Disease	0.7285

Table 5: Average of entity-wise F1-score for 11 monolangual tracks

Track	Clean F1	Noisy F1	F1
Swedish	67.15	60.87	65.0
Spanish	64.88	58.77	62.9
Italian	64.53	60.13	63.11
French	64.4	58.04	62.39
Portugese	63.07	57.23	61.05
Chinese	54.65	42.11	51.65
English	62.59	56.96	60.85

Table 6: Clean and Noisy set macro F1 score(%)

coarse-grained classes for some inputs.

Entity-wise evaluation Entity-wise F1 score for all tracks tabulated in Table 7. Table 5 showed the average of entity-wise F1-score for 11 monolingual tracks. The average F1 score of entity HumanSettlement is the best (0.8814) among all other entities, this entity is part of the Location entity which showed good results in the previous year's multiconer task but it is also has observed that the OtherLOC entity has 0.5768 F1 score and it is also part of Location Entity. With empirical results, we can infer that the model is confused during the prediction of OtherLOC with other Location entities.

7 Conclusion

In this paper, we performed multi-lingual track knowledge utilization for mono-lingual track for the MultiCoNER II shared task. Our team achieved the top 10 on 4 tracks. We show that the performance of monolingual models can be enhanced by two-stage fine-tuning by allowing them to learn from the training data of all the languages. Also, we have compared two cross-lingual language models, Muril and XLM-R for multi-lingual knowledge transferability for the Hindi, Bangla and English track of this task. Muril performed better for Hindi and Bangla tracks although XLM-R performed better for the English track. For future work, we plan to explore more language models for utilizing their knowledge for NER and other NLP tasks.

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.

Class	BN	ZH	DE	ES	UK	IT	FR	PT	HI	EN	SV
Facility	0.7615	0.5762	0.6832	0.6434	0.6621	0.7241	0.657	0.6752	0.7358	0.6471	0.7428
OtherLOC	0.6459	0.4346	0.4666	0.4673	0.5722	0.5021	0.4933	0.6768	0.6654	0.5256	0.8959
HumanSettlement	0.9404	0.7704	0.8995	0.8668	0.882	0.8637	0.8317	0.8872	0.9411	0.8796	0.9334
Station	0.8924	0.7843	0.7695	0.7308	0.7318	0.7108	0.7402	0.7392	0.8792	0.7649	0.703
VisualWork	0.8213	0.5986	0.7636	0.7425	0.766	0.8972	0.8365	0.7588	0.8634	0.6745	0.8097
MusicalWork	0.6221	0.4736	0.7833	0.7045	0.6626	0.8162	0.6692	0.7689	0.3957	0.702	0.7632
WrittenWork	0.8457	0.6636	0.7848	0.7148	0.6949	0.678	0.7188	0.675	0.8495	0.6536	0.6802
ArtWork	0.3448	0.3772	0.6647	0.4564	0.3255	0.5793	0.5001	0.132	0.3986	0.4951	0.2601
Software	0.8963	0.5294	0.7788	0.8032	0.8133	0.7401	0.7437	0.7697	0.8773	0.7008	0.7921
MusicalGRP	0.8486	0.5863	0.7461	0.7717	0.8393	0.7966	0.7436	0.7744	0.883	0.6765	0.804
PublicCorp	0.7825	0.4448	0.6101	0.6774	0.6921	0.6358	0.6294	0.7356	0.7804	0.5862	0.6128
PrivateCorp	0.7982	0.2086	0.3086	0.1928	0.0705	0.1919	0.3675	0.0062	0.6462	0.2193	0.2564
AerospaceManufacturer	0.5556	0.5455	0.7016	0.4172	0.3909	0.3166	0.5204	0.2647	0.5422	0.483	0.1823
SportsGRP	0.9413	0.7579	0.882	0.7765	0.8391	0.7537	0.7609	0.7972	0.9624	0.8137	0.8522
CarManufacturer	0.8632	0.4691	0.6578	0.7184	0.7457	0.5974	0.678	0.706	0.8981	0.6668	0.6852
ORG	0.873	0.5829	0.6892	0.6394	0.6693	0.5917	0.5992	0.6635	0.8659	0.614	0.6769
Scientist	0.435	0.3367	0.401	0.4428	0.4491	0.4516	0.4301	0.3912	0.474	0.4515	0.4254
Artist	0.7908	0.7062	0.762	0.7888	0.7611	0.8619	0.8085	0.7855	0.7772	0.7931	0.7768
Athlete	0.7542	0.6894	0.7362	0.7525	0.8288	0.8323	0.7566	0.667	0.8361	0.7747	0.7221
Politician	0.668	0.461	0.5694	0.561	0.4772	0.5562	0.6182	0.5545	0.7203	0.5596	0.6424
Cleric	0.6323	0.357	0.4525	0.5946	0.5566	0.6448	0.5737	0.6404	0.6864	0.5241	0.5926
SportsManager	0.5386	0.4206	0.46	0.5624	0.5367	0.6454	0.5341	0.5157	0.6855	0.5521	0.4425
OtherPER	0.4216	0.4063	0.47	0.5278	0.5604	0.4355	0.4622	0.4963	0.4906	0.4044	0.5386
Clothing	0.3529	0.4361	0.4902	0.5662	0.5579	0.4784	0.5458	0.4298	0.7513	0.5628	0.6344
Vehicle	0.7773	0.5376	0.6267	0.6115	0.6277	0.5831	0.4854	0.5758	0.8276	0.5144	0.6279
Food	0.7074	0.5157	0.6528	0.5965	0.6589	0.5714	0.5633	0.6093	0.7249	0.5597	0.6592
Drink	0.7687	0.2814	0.5	0.6633	0.6543	0.6202	0.5877	0.6612	0.7516	0.5787	0.6994
OtherPROD	0.7128	0.4736	0.6208	0.5573	0.6088	0.5516	0.5659	0.6591	0.7093	0.508	0.6731
Medication/Vaccine	0.8064	0.6206	0.7924	0.7356	0.7923	0.7564	0.7325	0.7403	0.8126	0.7343	0.7299
MedicalProcedure	0.846	0.5545	0.7603	0.6569	0.6317	0.6898	0.6747	0.6566	0.8006	0.6399	0.6653
AnatomicalStructure	0.8564	0.6289	0.7716	0.7062	0.7498	0.6405	0.625	0.6663	0.8377	0.6782	0.7447
Symptom	0.7602	0.2024	0.3818	0.4225	0.3482	0.4375	0.4805	0.3489	0.61	0.4602	0.4663
Disease	0.9003	0.6142	0.7857	0.6881	0.6913	0.6734	0.6549	0.718	0.8473	0.6829	0.758

Table 7: Entity-wise F1-score for 11 monolangual tracks

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

Jianglong He, Akshay Uppal, Mamatha N, Shiv Vignesh, Deepak Kumar, and Aditya Kumar Sarda. 2022. Infrrd.ai at SemEval-2022 task 11: A system for named entity recognition using data augmentation, transformer-based sequence labeling model, and EnsembleCRF. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1501–1510, Seattle, United States. Association for Computational Linguistics.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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)*, pages 1412–1437, Seattle, United States. 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 Asso-*

ciation for Computational Linguistics: Human Language Technologies, pages 1499–1512.

Sumit Singh, Pawankumar Jawale, and Uma Tiwary. 2022. silpa_nlp at SemEval-2022 tasks 11: Transformer based NER models for Hindi and Bangla languages. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1536–1542, Seattle, United States. 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. 2022. 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.

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. Huggingface's transformers: State-of-the-art natural language processing.