MLlab4CS at SemEval-2023 Task 2: Named Entity Recognition in low-resource language Bangla using Multilingual Language Models

Shrimon Mukherjee*, **Madhusudan Ghosh***, **Girish**, **Partha Basuchowdhuri**
Indian Association for the Cultivation of Science, Kolkata
{shrimonmukherjee, madhusuda.iacs, girishg1502}@gmail.com,
partha.basuchowdhuri@iacs.res.in

Abstract

Extracting of NERs from low-resource languages and recognizing their types is one of the important tasks in the entity extraction domain. Recently many studies have been conducted in this area of research. In our study, we introduce a system for identifying complex entities and recognizing their types from low-resource language Bangla, which was published in SemEval Task 2 MulitCoNER II 2023. For this sequence labeling task, we use a pre-trained language model built on a natural language processing framework. Our team name in this competition is **MLlab4CS**. Our model **MuRIL$_{\textbf{CRF Bangla}}$** produces a macro average F-score of 76.27%, which is a comparable result for this competition.

1 Introduction

An interesting and well-studied area of natural language processing (NLP) is the extraction of entities and detecting relationships between entities from unstructured text. Named entity recognition (NER) is a sub-task of information extraction, where entities are extracted from a given text and classified into some pre-defined classes.

NER has many practical use cases in various NLP tasks, including text summarization (Toda and Kataoka, 2005), information retrieval, machine translation (Babych and Hartley, 2003), and question answering (Mollá et al., 2006). A wide range of methodologies have been investigated by researchers to perform NER tasks including rule-based techniques (Krupka and IsoQuest, 2005), supervised feature-based approaches (Liao and Veeramachaneni, 2009), unsupervised methods (Feldman and Rosenfeld, 2006), deep learning-based methods (Li et al., 2020) and transformer-based approaches (Vaswani et al., 2017).

Recent applications of pre-trained language models (PLMs) have shown significant improvement in entity extraction and relation extraction tasks. PLMs can learn highly accurate linguistic, semantic, and factual information from a sizable amount of unlabeled data through self-supervised pre-training (Wang et al., 2022).

The second edition of the shared task was conducted in SemEval 2023 and was named Multilingual Complex Named Entity Recognition (MultiCoNER II). The goal of this task was to build a system, which is capable of extracting fine-grained complex entities of different *categories*, as provided in the data. The complete list of *categories* is - Facility, OtherLOC, HumanSettlement, Station, VisualWork, MusicalWork, WrittenWork, ArtWork, Software, MusicalGRP, PublicCorp, PrivateCorp, AerospaceManufacturer, SportsGRP, CarManufacturer, ORG, Scientist, Artist, Athlete, Politician, Cleric, SportsManager, OtherPER, Clothing, Vehicle, Food, Drink, OtherPROD, Medication/Vaccine, MedicalProcedure, AnatomicalStructure, Symptom, and Disease (Fetahu et al., 2023b).

The organizers (Fetahu et al., 2023b) invited the participants to submit NER models for each of the 13 languages listed on the website. They provided a different dataset (Fetahu et al., 2023a) for each language. Each dataset was made up of three parts: training, development, and testing. During the evaluation period of this competition, only the tokens were provided in the testing dataset and the named entity tags were provided for the training and development datasets. After the end of the evaluation phase, the organizers released the gold-standard test data and we used it to analyze our system. We participated in the task to build a NER model in the Bangla language track.

We participated in SemEval MultiCoNER II 2023 and developed a neural framework to extract the entities from a Bangla language dataset (Track 11). Note that Bangla is a low-resource Asian language used primarily in Bangladesh and

*These authors contributed equally to this work

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1388–1394
July 13-14, 2023 ©2023 Association for Computational Linguistics

in some parts of eastern India. Our proposed system performs entity extraction from a low-resource Bengali dataset using MuRIL language model. It uses a transfer learning strategy by fine-tuning MuRIL (Khanuja et al., 2021). Furthermore, we apply a conditional random field (CRF) decoder to implement the entity extraction task. Experimental results show that the proposed system achieves comparable results on this task.

2 Related Work

Named entity recognition (NER), also known as entity extraction, is a fundamental task in natural language processing. Early works in NER systems used handcrafted features (Zhou and Su, 2002), but even state-of-the-art systems developed for one domain did not generally perform well for other domains. These models also struggle with multilingual datasets, and while they may obtain better precision, it comes at the cost of a lower recall and higher cost of work. Since the early 2000s, NER has attracted much attention and has seen various applications in bioinformatics, molecular biology, natural language processing, and information extraction. In recent years, deep-learning methods have been employed in NER tasks (Chiu and Nichols, 2015) and have shown considerable improvement over conventional methods. Deep learning models can perform better than feature-based approaches as NER benefits from non-linear transformations, which help the model to learn complex features of the dataset.

Pre-trained contextual embeddings like BERT, LUKE, and XLM-R have achieved near-human performance in certain NER tasks. These embeddings are generally trained on large datasets like Wikipedia in order to improve the contextual representation of entities. However, these models significantly underperform on unseen entities and noisy datasets (which are more "real-world"-like) Meng et al. (2021). Some models implement knowledge bases or gazetteers, into neural architectures in order to incorporate external information. But, such systems have only reported limited improvements (Rijhwani et al., 2020). Models that use gazetteer features have been shown to suffer feature undertraining, where the model relies heavily on either context or gazetteer. Hence, such models are not reliable as utility of gazetteer is variable.

Recent studies (Lin et al., 2018), (Yang et al., 2016), (Xie et al., 2018) have developed multilingual, multi-task models that achieve state-of-the-art performance in a cross-lingual setting, with much lower resource requirements than earlier models. In some cases, these models are able to perform at the level of systems that use data augmentation, by using transfer learning. They also provide comparable results in scenarios with significantly fewer cross-lingual resources.

The SemEval 2023 task 2 is a continuation of MultiCoNER 2022 (Malmasi et al., 2022b) which also featured the aforementioned tracks along with a code-mixed track. The 2022 datasets also featured complex entities with limited context (Malmasi et al., 2022a). Methods that augmented external knowledge into transformer models achieved the best results, but they were unable to perform at the same level for Indian languages. NER in multi-lingual domain has benefited from advances in pre-trained multi-lingual transformers, but these systems are bottlenecked by limited entity knowledge across all languages (Fetahu et al., 2021). Another key challenge for pre-trained models is the presence of emerging entities in real-world scenarios like web-queries. Fine-tuning pre-trained embeddings is a significantly effective approach for many downstream NLP tasks including NER. Recent approaches propose to train already fine-tuned embeddings on specific training data. Shi et al.(Shi and Lee, 2021) proposed to first train on a general multilingual model and then fine-tune it for a specific language. Multi-stage fine-tuning has also shown to accelerate the training process.

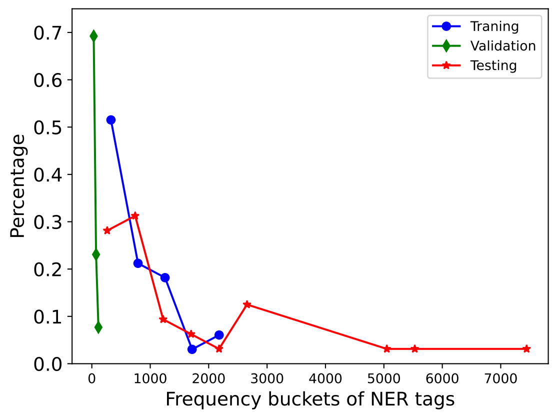

Figure 1: Distribution of NER tags in the training, validation, and test dataset, mentioned in Table 3.

3 Data

There are 33 different fine-grained NER entity types in the Bangla dataset (Fetahu et al.,

Bangla Dataset (Track 11)
ট্রাম[OtherPROD] স্টেশনটি কোচি[HumanSettlement] কোওচি[ORG] প্রশাসনিক[ORG] অঞ্চল[ORG] সূর্য[HumanSettlement] উদয়ের[HumanSettlement] দেশ[HumanSettlement] ।
তিনি একটি চলচ্চিত্র পরিবার থেকে এসেছিলেন যেখানে আটটি চাচা এবং খালা অন্তর্ভুক্ত ছিল এবং তিনি তার বাবা -মায়ের পাশাপাশি marion[Athlete] robert[Athlete] morrison[Athlete] এবং জন[Artist] ফোর্ড[Artist] এর সাথে মুভি সেটগুলিতে বড় হয়েছিলেন।

Table 1: A sample snippet of the tagged Bangla data.

LOC	CW	GRP	PER	PROD	MED
Facility	VisualWork	MusicalGRP	Scientist	Clothing	Medication/Vaccine
OtherLOC	MusicalWork	PublicCorp	Artist	Vehicle	MedicalProcedure
HumanSettlement	WrittenWork	AerospaceManufacturer	Athlete	Food	AnatomicalStructure
Station	ArtWork	SportsGRP	Politician	Drink	Symptom
	Software	CarManufacturer	Cleric	OtherPROD	Disease
		ORG	SportsManager		
		PrivateCorp	OtherPER		

Table 2: Fine-grained entity types present in the Bangla dataset.

2023a). The dataset is in standard CoNLL format, which uses BIO (Beginning-Inside-Outside) tagging scheme. The datasets provided were of three types namely, training, development, and testing. The training and development data contained tokens with tags, whereas testing data contained only tokens during the evaluation phase. After the completion of the competition, updated test data containing tokens and tags was released. The dataset statistics have been shown in Table 3. Fine-grained entity-type statistics have been presented in Table 2. In Table 1, a snippet of the dataset has been displayed. In Figure 1, the distribution of entity tags was shown for the training, development, and testing data.

From Table 3, we can clearly see that the testing data consists of double the number of instances present in the training data. By checking the frequency counts of the entity tags, we found that they are present in all three types of data in similar proportions. In Figure 1, the distribution of entity tags in terms of frequency buckets has been shown. It shows that only a few of the entity tags are highly frequent.

Dataset	**# of samples**
`Training Data`	9,708
`Validation Data`	507
`Test Data`	19,859

Table 3: Bangla (track 11) dataset statistics

4 Methodology

In this section, we introduce the problem definition followed by the proposed methodology of our neural system called (MuRIL + CRF + Bangla) $\textbf{MuRIL}_{\textbf{CRF Bangla}}$ [1].

4.1 Problem Definition

The primary task of our work may be stated as follows. Given a set of Documents $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_n)$, our objective is to identify entity spans $\mathcal{E}_s$ as well as the categories of the identified entities from the sentences $(x_1, x_2, x_3...x_n)$ of a particular document $\mathcal{D}_i \in \mathcal{D}$. Here, we consider our tasks as a sequence labeling problem such that, given a sequence of tokens in any sentence $x = (v_1, v_2, v_3...v_n)$ from a document $\mathcal{D}_i \in \mathcal{D}$, our proposed neural framework aims to learn certain parameters of f_θ, which can map an input sequence of word vectors to a sequence of output labels $f : (V_i, \theta) \rightarrow y_i$, where each $V_i \in \mathrm{R}^d$ denotes an embedding vector for the token V_i and each $y_i \in \{B- < type >, I- < type >, O\}$ represents a label indicating whether the token is the beginning, continuation, or end of an entity span.

4.2 System Description

In our system, we use **MuRIL Large** (Khanuja et al., 2021) model as the underlying language model and fine-tune it using a task-specific loss function. **MuRIL** is specifically designed to deal with the diverse linguistic characteristics of Indic

[1] `https://github.com/MLlab4CS/MurilCRFBangla.git`

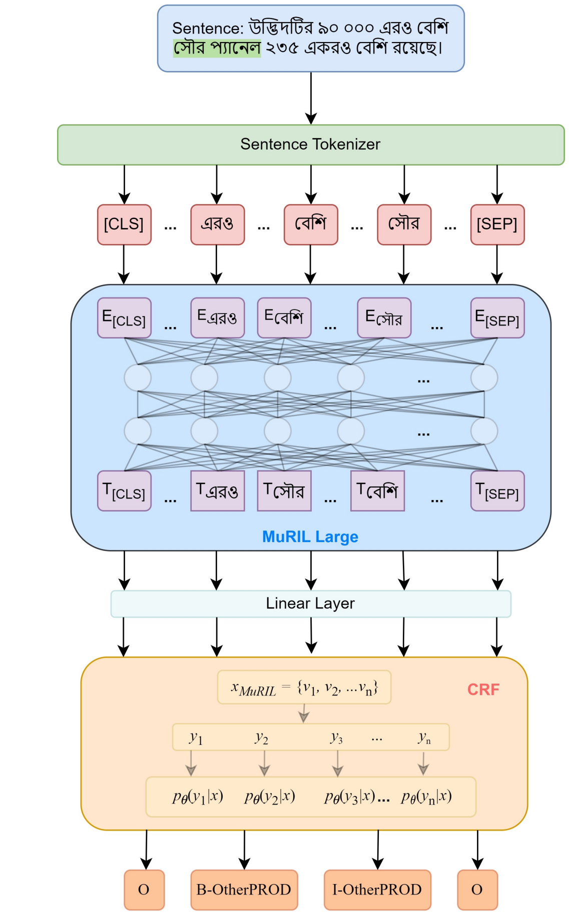

Figure 2: The overall architecture of our proposed model **MuRIL$_{\text{CRF Bangla}}$**

languages. Therefore, we chose this model for the NER task in the Bangla language. We take a set of sequence pairs, $D = \{(x, y)\}$, and input the word sequences into the tokenizer module. Subsequently, we pass them through **MuRIL Large** PLM. The output representation produced by the module ($x_{MuRIL} = \{v_1, v_2, ..v_n\}, v_i \in \mathrm{R}^{d_{MuRIL}}$) is fed into a linear-chain CRF layer to obtain the conditional probability $p_\theta(y|x)$:

$$\phi(y', y,) = \exp(W_y^T v_i + b_{y',y})$$
$$p_\theta(y|x) = \frac{\prod_{i=1}^{n} \phi(y_{i-1}, y_i, v_i)}{\sum_{y' \in \gamma(x)} \prod_{i=1}^{n} \phi(y'_{i-1}, y'_i, v_i)} \quad (1)$$

The set of model parameters is referred to as θ, while $\gamma(x)$ stands for the collection of all feasible label sequences corresponding to x. In the potential function $\phi(y', y,)$, $W_y^T v_i$ is the emission score and $b_{y',y}$ is the transition score, where $W^T \in \mathrm{R}^{t \times d_{MuRIL}}$ and $b \in \mathrm{R}^{t \times t}$ are parameters and the subscripts y' and y are the indices of the matrices. We train our proposed model using negative log-likelihood loss $\mathcal{L}_{NLL}(\theta) = -\log (p)_\theta(y^\star|x)$. During inference, the model prediction $\hat{y}_\theta$ is given by Viterbi decoding.

The overall architecture diagram of our proposed model, **MuRIL$_{\text{CRF Bangla}}$** has been shown in Figure 2.

Models	Precision	Recall	F-score
IndicBert$_{\text{Bangla}}$	47.93	49.19	47.61
XLMR-Large$_{\text{Bangla}}$	86.24	86.88	86.21
MuRIL$_{\text{Bangla}}$	86.95	86.88	86.28
IndicBert$_{\text{CRF Bangla}}$	61.26	60.63	60.03
XLMR-Large$_{\text{CRF Bangla}}$	86.95	87.85	87.09
MuRIL$_{\text{CRF Bangla}}$	**92.33**	**89.93**	**90.80**

Table 4: Results of our model vs. the baselines on Bangla validation data.

Models	Precision	Recall	F-score
IndicBert$_{\text{Bangla}}$	31.20	35.46	31.83
XLMR-Large$_{\text{Bangla}}$	69.95	75.07	71.35
MuRIL$_{\text{Bangla}}$	71.78	77.82	74.31
IndicBert$_{\text{CRF Bangla}}$	42.35	48.21	43.37
XLMR-Large$_{\text{CRF Bangla}}$	70.31	75.38	71.74
MuRIL$_{\text{CRF Bangla}}$	**74.47**	**80.44**	**76.27**

Table 5: Results of our model vs. the baselines on Bangla Test data.

ArtWork	**Gold annotation**	ক্রাইস্ট চাইল্ড ওয়াকিং ফ্রেম সহ এই পেইন্টিংয়ের পিছনে আঁকা।
	MuRIL$_{\text{CRF Bangla}}$	ক্রাইস্ট চাইল্ড ওয়াকিং ফ্রেম সহ এই পেইন্টিংয়ের পিছনে আঁকা।
Artist, HumanSettlement	**Gold annotation**	জন্ম : আরিথা ফ্রাঙ্কলিন সোল গায়ক মেমফিস এ (ডি। ২০১৮)
	MuRIL$_{\text{CRF Bangla}}$	জন্ম : আরিথা ফ্রাঙ্কলিন সোল গায়ক মেমফিস এ (ডি। ২০১৮)

Table 6: Comparative evaluation of two sentences from the test corpus. Green color indicates that our model predicts the entity correctly, whereas, red indicates that our model predicts the entities incorrectly.

5 Results and Analysis

We explore the effectiveness of several available multilingual PLMs for our downstream entity extraction task.

- **IndicBert$_{\text{Bangla}}$**: In this model, we fine-tune `indic-bert` [2] (Kakwani et al., 2020) on our downstream task.

- **XLMR-Large$_{\text{Bangla}}$**: Here, we apply fine-tuning strategy using `xlm-roberta-large` [3] (Conneau et al., 2020) for our downstream task.

[2] https://huggingface.co/ai4bharat/indic-bert
[3] https://huggingface.co/xlm-roberta-large

Entity Type	Freq in training data	Precision	Recall	F-score
ArtWork	356	47.00	10.33	16.94
AerospaceManufacturer	313	43.75	14.43	21.71
Scientist	551	42.86	60.00	50.00
OtherPER	1414	50.32	56.04	53.03
Clothing	293	40.54	88.24	55.56
SportsManager	494	53.63	78.28	63.66
MusicalWork	617	64.77	75.66	69.80
Cleric	564	70.08	77.08	73.41
Politician	1377	77.11	70.56	73.69
OtherPROD	718	74.72	74.72	74.72
Food	461	70.77	81.24	75.64
OtherLOC	449	70.23	87.79	78.04
Artist	2411	77.13	80.61	78.83
Athlete	1178	78.70	80.22	79.45
Facility	837	76.49	84.45	80.28
VisualWork	498	83.22	77.90	80.47
PublicCorp	580	73.04	91.30	81.16
Vehicle	324	74.68	88.94	81.19
Symptom	230	73.19	96.19	83.13
Drink	251	79.26	89.17	83.92
Medication/Vaccine	408	82.58	87.23	84.84
MusicalGRP	553	85.86	87.00	86.42
WrittenWork	1278	87.52	85.38	86.44
MedicalProcedure	367	82.15	91.73	86.68
AnatomicalStructure	365	87.18	89.47	88.31
ORG	1599	90.13	87.27	88.68
Software	738	86.16	92.00	88.98
Station	628	84.94	94.63	89.52
PrivateCorp	96	86.86	93.70	90.15
Disease	483	88.25	93.50	90.80
CarManufacturer	363	87.37	98.81	92.74
HumanSettlement	2076	94.32	93.73	94.03
SportsGRP	1192	92.62	96.97	94.75
Macro avg		74.47	80.44	76.27

Table 7: Entity-wise Precision, Recall, and F-score by **MuRIL$_{\textbf{CRF Bangla}}$**, sorted in terms of F-score.

- **MuRIL$_{\textbf{Bangla}}$**: Similarly, we fine-tune `muril-large-cased` [4] for our entity extraction task.

- **IndicBert$_{\textbf{CRF Bangla}}$**: Similar to IndicBert$_{\text{Bangla}}$, we fine-tune `indic-bert` follwed by a CRF decoder layer.

- **XLMR-Large$_{\textbf{CRF Bangla}}$**: Here, we fine-tune `xlm-roberta-large` along with a CRF decoder layer.

- **MuRIL$_{\textbf{CRF Bangla}}$**: Additionally, we also fine-tune `muril-large-cased` with a CRF decoder for our downstream entity extraction task.

We conduct the required training experiments on these six proposed models using Bangla training data. Table 4 shows the performance of our five proposed models on Bangla development data. From Table 4, we observe that **MuRIL$_{\textbf{CRF Bangla}}$** outperforms all the models in terms of macro average precision, recall, and F-score. So, we fixed **MuRIL$_{\textbf{CRF Bangla}}$** as our best model to submit the prediction on the test dataset. Table 5 shows the proposed model's performance in terms of macro average precision, recall, and F-score. We observe that **MuRIL$_{\textbf{CRF Bangla}}$** outperforms other models in the macro average precision, recall, and F-score.

Table 6 shows a dataset snippet, where our model captures NERs from the test dataset. Table 7 produces fine-grained entity-wise precision, recall, and F-score and also macro average on precision, recall, and F-score by our best model **MuRIL$_{\textbf{CRF Bangla}}$**. From Table 7, we clearly see that our model **MuRIL$_{\textbf{CRF Bangla}}$** performs well for some low-frequency entities like PrivateCORP but fails to do so, for some other low-frequency entities like ArtWork, and AerospaceManufacture.

Throughout our experiments, we utilize the FLAIR (Akbik et al., 2019) framework and train all the neural models for 30 epochs with a batch size of 64, using AdamW (Loshchilov and Hutter, 2019) optimizer with a very small initial learning rate of $5e^{-5}$ and a stopping criterion as mentioned in (Conneau et al., 2020). To carry out our experiments, we used NVIDIA A100 and NVIDIA RTX A600 GPUs on the Paperspace cloud.

[4] `https://huggingface.co/google/muril-large-cased`

6 Conclusion

This paper discusses a transformer-based deep neural network for extracting complex named entities from a Bangla (low-resource language) dataset released by SemEval MultiCoNER II 2023 shared task team. Our system, **MuRIL$_{\textbf{CRF Bangla}}$**, has achieved a macro F-score of 76.27% on the test data. Our future work will include few-shot learning strategies to extract complex NERs from low-resource languages. We also plan to look into data-driven factored modeling techniques for dealing with class imbalance issues.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59.

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

Jason P. C. Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns.

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 ACL*, pages 8440–8451, Online. ACL.

Ronen Feldman and Benjamin Rosenfeld. 2006. Boosting unsupervised relation extraction by using ner. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 473–481.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023a. Multi-CoNER 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.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, NC Gokul, Avik Bhattacharyya, Mitesh M Khapra, and Pratyush Kumar. 2020. Indicnlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

GR Krupka and K IsoQuest. 2005. Description of the nerowl extractor system as used for muc-7. In *Proc. 7th Message Understanding Conf*, pages 21–28.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Wenhui Liao and Sriharsha Veeramachaneni. 2009. A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 58–65.

Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. 2018. A multi-lingual multi-task architecture for low-resource sequence labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 799–809, Melbourne, Australia. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

Diego Mollá, Menno Van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian language technology workshop 2006*, pages 51–58.

Shruti Rijhwani, Shuyan Zhou, Graham Neubig, and Jaime Carbonell. 2020. Soft gazetteers for low-resource named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8118–8123, Online. Association for Computational Linguistics.

Tianze Shi and Lillian Lee. 2021. TGIF: Tree-graph integrated-format parser for enhanced UD with two-stage generic- to individual-language finetuning. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 213–224, Online. Association for Computational Linguistics.

Hiroyuki Toda and Ryoji Kataoka. 2005. A search result clustering method using informatively named entities. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 81–86.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Liwen Wang, Rumei Li, Yang Yan, Yuanmeng Yan, Sirui Wang, Wei Wu, and Weiran Xu. 2022. Instructionner: A multi-task instruction-based generative framework for few-shot ner. *arXiv preprint arXiv:2203.03903*.

Jiateng Xie, Zhilin Yang, Graham Neubig, Noah A. Smith, and Jaime Carbonell. 2018. Neural cross-lingual named entity recognition with minimal resources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 369–379, Brussels, Belgium. Association for Computational Linguistics.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *CoRR*, abs/1603.06270.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 473–480, USA. Association for Computational Linguistics.