SampurNER: Fine-grained Named Entity Recognition Dataset for 22 Indian Languages

Prachuryya Kaushik and Ashish Anand

Indian Institute of Technology Guwahati, Assam, India
k.prachuryya@iitg.ac.in, anand.ashish@iitg.ac.in

Abstract

We introduce SampurNER, a fine-grained named entity recognition (FgNER) dataset encompassing all 22 scheduled Indian languages spoken by more than two billion people across various countries. While manual annotation for FgNER resources is often labor-intensive and expensive, distant supervision methods have been employed as a viable solution. However, such datasets are often noisy, with entity mentions tagged with multiple types, requiring computationally intensive noise-aware models for effective FgNER. Moreover, resources for both coarse-grained and fine-grained named entity recognition tasks in Indian languages remain scarce. To address this, we propose an entity-anchored machine translation (EaMaTa) framework that leverages the largest manually annotated English FgNER dataset, *FewNERD*, to create a large-scale FgNER dataset in 22 languages. On average, the dataset comprises over 153k sentences, 354k entities, and 3.3M tokens in each language. The languages covered are: *Assamese (as)*, *Bengali (bn)*, *Bodo (brx)*, *Dogri (doi)*, *Gujarati (gu)*, *Hindi (hi)*, *Kannada (kn)*, *Kashmiri (ks)*, *Konkani (gom)*, *Maithili (mai)*, *Malayalam (ml)*, *Manipuri (mni)*, *Marathi (mr)*, *Nepali (ne)*, *Odia (or)*, *Punjabi (pa)*, *Sanskrit (sa)*, *Santali (sat)*, *Sindhi (sd)*, *Tamil (ta)*, *Telugu (te)*, and *Urdu (ur)*. Various rigorous analyses and human evaluations confirm the high quality of the dataset and demonstrate the effectiveness of the entity-anchored machine translation framework with up to 9% increase in F1-score against the current state-of-the-art. Additionally, we extend our analysis to zero-shot, multilingual, and cross-lingual settings, investigating the influence of language family and script similarity on cross-lingual FgNER performance.

Datasets — hf.co/datasets/prachuryyaIITG/SampurNER
Extended version — tinyurl.com/SampurNER

Introduction

Named Entity Recognition (NER), the task of detecting and classifying mentions of entities from unstructured texts, has evolved from early rule-based systems (Rau 1991) to powerful neural architectures today (Huang et al. 2025). NER plays a crucial role in various applications such as recommendation systems, knowledge base construction, relation extraction, and information retrieval. While, in general, NER focuses on coarse-grained categories like *Person*, *Location*, and *Organization*, fine-grained named entity recognition (FgNER) extracts richer, domain-specific information with entity types such as *Athlete*, *Island*, *Sports League*, *Food*, etc. Sekine, Sudo, and Nobata (2002) first introduced fine-grained entity classification with 150 entity types in a multi-level hierarchy. Subsequent FgNER resources vary widely in entity type granularity: ACE (52 types) (Doddington et al. 2004), OntoNotes (88 types) (Gillick et al. 2014), BBN (93 types) (Weischedel and Brunstein 2005), FIGER (113 types) (Ling and Weld 2012), and HYENA (505 types) (Yosef et al. 2012). Large-scale resources like FINET (Del Corro et al. 2015), TypeNet (Murty et al. 2017), and UFET (Choi et al. 2018) proposed thousands of entity types. Abhishek et al. (2019) improved quality with language-specific heuristics and refined selections, whereas Ding et al. (2021) provides a large, manually annotated FewNERD dataset covering 66 fine-grained types.

Early Indian-language NER began with IJCNLP-2008 for Bengali, Hindi, Odia, Telugu, and Urdu (Singh 2008), and further enhanced by Gali et al. (2008), Gupta and Bhattacharyya (2010), Ekbal and Saha (2011), and Devi et al. (2014). Al-Rfou et al. (2015) and Pan et al. (2017) extended coverage to many languages, including a few Indian languages. Mhaske et al. (2023) further expanded coarse-grained NER for 11 Indian languages via the entity-projection framework (Ruder et al. 2021). Moreover, manually annotated NER datasets in several languages exist, including Telugu (Reddy et al. 2018), Maithili (Priyadarshi and Saha 2021), Hindi (Venkataramana et al. 2022), Assamese (Pathak, Nandi, and Sarmah 2022), Marathi (Litake et al. 2022), Nepali (Niraula and Chapagain 2022), Bodo (Narzary et al. 2024) etc.

Although coarse-grained NER for Indian languages has made notable progress, fine-grained NER (FgNER) is a relatively recent area of focus. The MultiCoNER2 (Fetahu et al. 2023a) shared task at SemEval-2023 (Fetahu et al. 2023b) provided FgNER datasets for Hindi and Bengali by translating the annotated dataset from English. Recently, the TAFSIL initiative by Kaushik, Mishra, and Anand (2025) utilized Wikipedia links and Wikidata to produce noisy FgNER datasets covering six Indian languages (Hindi, Marathi, Sankrit, Tamil, Telugu, and Urdu) across four taxonomies. Despite these advances, comprehensive

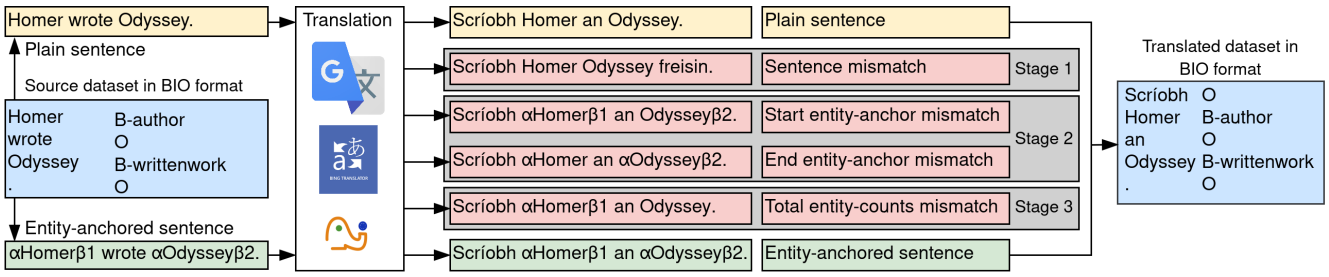

Figure 1: Entity-anchored machine translation: The source dataset is translated both as 'Plain sentence' and 'Entity-anchored sentence' to the target languages. The cleaning process includes the removal of sentences with sentence mismatch, entity-anchor boundaries mismatch (both start and end), and total entity counts mismatch between the source and translated sentences.

and high-quality FgNER resources remain scarce for most low-resource Indian languages.

To bridge this gap, we introduce **SampurNER**[1], a new initiative aimed at expanding FgNER datasets for Indian languages. Our approach leverages an **entity-anchored machine translation** (EaMaTa) framework to efficiently translate the high-quality FewNERD dataset into 22 scheduled Indian languages. FewNERD is the only large-scale, human-annotated FgNER dataset, comprising 188,238 sentences with 4,601,160 words, featuring a two-level hierarchy of 8 coarse-grained and 66 fine-grained entity types, annotated by 70 linguistically trained annotators with an inter-annotator agreement of 0.76 (κ).

Our contributions can be summarized as follows:

1. Development of an entity-anchored machine translation (EaMaTa) framework to create high-quality multilingual FgNER datasets from English annotations.

2. Construction of SampurNER, a large-scale dataset, comprising over 153,000 sentences, 354,000 entities, and 3.3 million tokens on average for each of the 22 scheduled Indian languages spoken by more than two billion people worldwide. The languages included are *Assamese (as)*, *Bengali (bn)*, *Bodo (brx)*, *Dogri (doi)*, *Gujarati (gu)*, *Hindi (hi)*, *Kannada (kn)*, *Kashmiri (ks)*, *Konkani (gom)*, *Maithili (mai)*, *Malayalam (ml)*, *Manipuri (mni)*, *Marathi (mr)*, *Nepali (ne)*, *Odia (or)*, *Punjabi (pa)*, *Sanskrit (sa)*, *Santali (sat)*, *Sindhi (sd)*, *Tamil (ta)*, *Telugu (te)*, and *Urdu (ur)*. Among these languages, *Bodo* and *Manipuri* are considered to be vulnerable languages (UNESCO 2017).

3. Manually annotated gold test set having 1000 sentences for ten languages: *as*, *bn*, *brx*, *hi*, *mr*, *ne*, *sa*, *ta*, *te*, and *ur*.

4. Rigorous human evaluations and extensive experiments with state-of-the-art suggest the good quality of the datasets created through the proposed EaMaTa framework.

5. Cross-lingual zero-shot analysis to examine the influence of multilingualism, language families, and script similarities on cross-lingual FgNER performance.

Related Works

Annotation projection-based and annotation preserving translation methods have been extensively used for NER across multiple languages (Mayhew, Tsai, and Roth 2017; Ugawa et al. 2018; Vavekanand, Das, and Kumar 2025), leading to the creation of several NER resources (Liu et al. 2021; Yang et al. 2022; Tulajiang et al. 2025; Kaushik and Anand 2025). Similarly, in MultiCoNER-1 (Malmasi et al. 2022) and MultiCoNER2, machine translation was used to develop coarse-grained and fine-grained NER datasets for multiple languages, including Hindi and Bengali, which were further enriched by Ma et al. (2023a,b); Tan et al. (2023).

Entity-Anchored Machine Translation

We propose an entity-anchored machine translation (EaMaTa) framework to preserve NER information during translation. As shown in Figure 1, we convert the source NER dataset from the BIO format into both plain sentence and entity-anchored sentence formats. In the entity-anchored format, each entity mention e_{ji} for the jth token is marked with a unique start symbol (α_j) and a distinct end symbol (β_{ji}) which varies based on the entity type i of e_{ji}. The design of these special anchor symbols is guided by an evaluation of specific translation systems to ensure that the anchors remain intact at entity boundaries after the translation process without the loss of crucial information. As shown in Algorithm 1, both the plain sentence and entity-anchored sentence are then independently translated from the source language into the target language. In Stage 1, we first consider the entity-anchored sentence without the anchor symbols and compare it with the corresponding plain translated sentence. If discrepancies exist between these two versions, then the sentence is discarded. Next, in Stage 2, the entity boundary errors are identified: sentences in which an entity mention contains a start anchor but lacks a corresponding closing anchor (or vice versa) are removed. Finally, in Stage 3, if the total count of entity types in the translated sentence does not match that of the original source sentence, then the translated sentence is discarded, as such discrepancies indicate instances where a valid entity mention has been incorrectly omitted. After filtering out erroneous sentences, we obtain a clean, FgNER dataset for the target language.

Experimental Setup

The state-of-the-art approach for sequence labeling tasks involves fine-tuning pre-trained language models (PLM) with

[1] '*Sampurna*' means 'entire' in many Indian languages

Language	BLEU			TER			chrF			COMET		
	Google	Bing	IT2	Google	Bing	IT2	Google	Bing	IT2	Google	Bing	IT2
Gujarati (gu)	**36.49**	34.33	33.60	**46.08**	48.98	50.15	**63.67**	63.24	62.50	**0.909**	0.901	0.905
Bodo (brx)	–	**18.72**	18.14	–	**64.06**	64.19	–	**57.95**	57.49	–	**0.647**	0.645
Santali (sat)	24.69	–	**27.49**	57.46	–	**56.04**	59.86	–	**61.57**	0.753	–	**0.777**

Table 1: Translation metrics for comparison of Google, Bing, and IndicTrans2 (IT2) translators (**best values** are in **bold**).

Algorithm 1: Entity-anchored machine translation

Input: S^{src}: Source dataset sentences in BIO format
Required: $\mathcal{M}$: Translation service
Output: S^T_{clean}: FgNER dataset in target language T

Preprocessing:
for each tokenized sentence $s \in S^{\text{src}}$ **do**
  Convert to plain format: $s^{\text{src}}_{\text{plain}}$
  Convert to entity-anchored format: Create $s^{\text{src}}_{\text{anchor}}$, by replacing entity mention e_{ji} belonging to entity type i with $\alpha_j e_{ji} \beta_{ji}$ where α_j: start anchor, β_{ji}: end anchor.
end for
Translation:
for each $s \in S^{\text{src}}$ **do**
  $s^T_{\text{plain}} = \mathcal{M}(s^{\text{src}}_{\text{plain}}), \quad s^T_{\text{anchor}} = \mathcal{M}(s^{\text{src}}_{\text{anchor}})$
end for
Cleaning:
for each s^T_{anchor} **do**
  Stage 1: Sentence mismatch detection:
  Convert (s^T_{anchor}) to $\overline{s}^T_{\textbf{plain}}$ by ignoring the anchors
  if $\overline{s}^T_{\textbf{plain}} \neq s^T_{\text{plain}}$ **then**
    Discard s^T_{anchor}
  end if
  Stage 2: Entity-anchors mismatch detection:
  if For all e in s^T_{anchor}, there exists e_{ji} such that $\alpha_j \notin s^T_{\text{anchor}}$ OR $\beta_{ji} \notin s^T_{\text{anchor}}$ **then**
    Discard s^T_{anchor}
  end if
  Stage 3: Total entity counts mismatch detection:
  if $\sum (e \in s^{\text{src}}_{\text{anchor}}) \neq \sum (e \in s^T_{\text{anchor}})$ **then**
    Discard s^T_{anchor}
  end if
end for
return $S^T_{\text{clean}} = \{s^T_{\text{anchor}} \mid s^T_{\text{anchor}}$ is not discarded$\}$

the NER datasets (Venkataramana et al. 2022; Litake et al. 2022; Mhaske et al. 2023; Tulajiang et al. 2025). Similarly, we have fine-tuned mBERT (bert-base-multilingual-cased) (Devlin et al. 2019) and IndicBERTv2 (IndicBERTv2-MLM-Sam-TLM) (Doddapaneni et al. 2023) for FgNER using the Hugging Face Transformers (Wolf et al. 2020). The models were fine-tuned as per the following hyperparameters: batch size: 64, epochs: 6, optimizer: AdamW, learning rate: 5e-5, and weight decay: 0.01. Fine-tuning was performed on an NVIDIA A100 GPU, with evaluation based on SeqEval metrics, and the best performance determined by the F1-score. To compare with the state-of-the-art noisy FgNER dataset TAFSIL, we considered their best-performing two variations of the DECENT model (Sierra-Múnera, Westphal, and Krestel 2023), where the base encoder RoBERTa-large (Liu 2019) is changed to XLM-RoBERTa-large (Conneau et al. 2020), and IndicBERTv2 to accommodate fine-tuning with Indian languages. The hyperparameters for DECENT-based models are: learning rate for encoder: 5e-6, learning rate for head: 5e-4, dropout probability for head: 0.5, epoch: 2, batch size: 16, negative oversampling rate: 31, and prediction threshold: 0.9.

Implementation of EaMaTa Framework

First, we selected three translation services based on their coverage of Indian languages: Google Translate (2025), Bing Translator (2025), and IndicTrans2 (2023). We randomly chose three sets of 1000 English-to-Indian language sentence pairs from the BPCC corpus (Gala et al. 2023). The English sentence from each pair was translated into the respective Indian language sentence. We assessed translation quality by comparing them to the corresponding Indian language sentences using BLEU, TER, chrF, and COMET metrics. The average performance across the three experimental sets is presented in Table 1 for three languages, whereas the results of all 22 languages are included in the extended version of the paper. Based on our findings, we selected the best translation service for each language: IndicTrans2 for Santali, Bing Translator for Bodo and Kashmiri, and Google Translate for the remaining 19 languages.

To evaluate the proposed EaMaTa framework against the state-of-the-art, we first applied it to the MultiCoNER2 English train set to generate FgNER datasets in Bengali, Hindi, Marathi, Tamil, Telugu, Sanskrit, and Urdu, with dataset details available in the extended version. Since MultiCoNER2 was created via distant supervision and contains induced noise, we alternatively used the large-scale, manually annotated FewNERD dataset to build the SampurNER dataset.

Comparison with State-of-the-Art

Following Kaushik, Mishra, and Anand (2025), we performed rigorous analysis against two state-of-the-art datasets: the MultiCoNER2 dataset in Bengali and Hindi, and the TAFSIL dataset in Hindi, Marathi, Tamil, Telugu, Sanskrit, and Urdu. As shown in Table 2 (and in the extended version), the dataset generated by the EaMaTa framework is of very high quality with up to 9% increase in F1-score against the current state-of-the-art. Hence, we continued to utilize the proposed EaMaTa framework to create the silver-standard SampurNER dataset using the FewNERD dataset.

Ln	Dataset Splits Test	Train	Macro P	R	F1	Micro P	R	F1(↑%)	Ln	Dataset Splits Test	Train	Macro P	R	F1	Micro P	R	F1(↑%)
hi	M2	M2(9.6k)	73.8	70.8	72.3	72.2	70.8	71.5	bn	M2	M2(9.7k)	72.7	69.8	71.2	71.3	69.9	70.6
	M2	Our(13k)	71.4	75.9	**73.6**	70.6	75.1	**72.8**(**2**)		M2	Our(13k)	69.7	74.1	**71.8**	69.1	73.4	**71.2**(**1**)
hi	TM	TM(644k)	64.6	76.2	70.0	64.5	76.2	70.0	mr	TM	TM(126k)	50.9	81.9	62.3	49.4	83.7	62.1
	TM	Our(13k)	68.7	73.0	**70.8**	68.9	73.1	**70.9**(**1**)		TM	Our(13k)	63.0	66.9	**64.9**	62.7	66.6	**64.6**(**4**)
ta	TM	TM(464k)	51.7	81.1	64.0	50.7	81.0	63.4	te	TM	TM(392k)	52.5	82.8	64.9	52.1	81.3	64.2
	TM	Our(12k)	63.2	67.2	**65.1**	63.6	65.6	**64.5**(**2**)		TM	Our(12k)	64.8	68.9	**66.8**	64.1	68.2	**66.1**(**3**)
sa	TM	TM(18k)	32.7	53.3	40.6	31.4	53.4	40.2	ur	TM	TM(604k)	63.0	77.1	69.4	60.2	78.3	68.1
	TM	Our(8.3k)	42.9	45.6	**44.2**	42.5	45.2	**43.9**(**9**)		TM	Our(13k)	68.3	72.6	**70.4**	67.1	71.3	**69.1**(**1**)

Table 2: Comparison of DECENT model performance using XLM-RoBERTa base encoder fine-tuned on different datasets. Train set sizes are in brackets. Abbreviations: Ln: Language, M2: MultiCoNER2 dataset, TM: TAFSIL dataset in MultiCoNER2 taxonomy, Our: dataset built with EaMaTa framework. **Best values** in **bold**. Percentage **F1** improvements are within (**brackets**).

Ln	Train set Sent	Entity	Token	Development set Sent	Entity	Token	Test set Sent	Entity	Token	XSTS	Entity Errors BM	ET	SE
en†	131,767	340,387	3,359,329	18,824	48,770	482,037	37,648	96,902	958,765	–	15.6	9.8	6.2
as	107,249	237,260	2,194,925	15,438	34,560	318,105	30,658	67,466	625,870	3.77	14.3	14.8	4.7
bn	119,296	287,264	2,484,304	17,513	42,877	368,063	33,374	79,340	689,690	4.14	13.6	13.9	4.9
brx	117,659	262,792	2,354,696	16,762	37,496	336,269	33,615	74,576	672,246	3.65	13.2	15.7	4.5
doi	112,329	264,154	2,885,149	17,619	42,526	459,537	34,931	82,597	903,796	3.72	18.8	10.1	7.3
gom	83,415	182,806	1,637,018	12,276	27,262	243,817	23,759	51,483	463,980	4.04	16.9	13.4	5.9
gu	126,581	315,919	2,828,298	18,122	45,431	406,929	28,959	69,207	619,889	4.23	13.6	13.7	4.9
hi	124,887	290,192	3,298,116	17,882	41,824	457,573	35,713	82,440	908,513	4.07	18.0	9.1	8.8
kn	115,565	266,523	2,083,241	16,962	39,781	308,326	26,327	59,365	453,817	3.63	14.3	13.5	5.1
ks	123,679	288,544	2,910,937	17,417	40,350	408,053	35,106	81,181	823,040	3.62	18.2	11.5	6.9
mai	108,826	256,701	2,763,005	10,224	22,706	245,657	19,899	43,530	472,498	3.59	17.3	10.9	7.1
ml	91,743	199,485	1,504,839	15,608	35,140	265,049	23,480	50,319	377,213	3.92	15.1	13.6	5.7
mni	110,068	246,084	2,264,925	15,561	34,869	321,556	31,463	69,739	644,709	3.42	18.9	17.7	5.6
mr	125,543	309,220	2,614,024	17,650	43,407	367,882	36,237	89,295	754,851	4.11	13.1	14.0	4.8
ne	125,695	311,439	2,661,064	18,252	45,778	389,382	35,498	87,112	747,802	4.08	13.1	14.7	3.9
or	118,633	289,943	2,427,051	18,090	45,247	376,152	32,477	78,893	657,395	4.02	14.8	13.5	5.2
pa	96,986	234,436	2,348,393	17,655	44,415	443,788	36,920	92,655	928,798	4.16	16.7	11.7	6.3
sa	69,581	152,269	1,214,021	10,043	22,175	176,574	19,729	42,643	341,208	3.59	14.9	13.3	6.7
sat	87,650	153,533	2,223,951	12,526	22,159	312,706	24,921	43,264	619,556	3.45	19.9	18.1	9.9
sd	90,362	214,371	2,218,078	17,221	42,845	440,340	32,159	78,317	809,085	3.58	19.1	10.9	7.3
ta	96,004	216,285	1,711,203	10,702	23,542	183,893	25,160	55,927	441,141	3.89	15.9	13.6	5.5
te	85,893	193,425	1,505,321	16,790	39,909	309,345	21,729	47,988	372,946	4.19	15.1	13.4	5.6
ur	122,794	298,069	3,229,867	17,570	43,205	465,417	35,198	85,785	929,427	4.05	17.8	11.1	6.6

Table 3: SampurNER and FewNERD (en†) datasets' statistics and evaluations. Abbreviations: Ln: Language, Sent: Sentences; XSTS: Crosslingual Semantic Text Similarity; BM: Boundary Mention error, ET: Entity Type error, SE: Spurious Entity error.

SampurNER Dataset

As presented in Table 3, the SampurNER dataset, on average, consists of over 153 thousand sentences, 354 thousand entities, and 3.3 million tokens for each of the 22 scheduled Indian languages. The SampurNER dataset is the first and the largest FgNER dataset across all 22 scheduled Indian languages. This dataset is a silver dataset as all the train, test, and development sets are translated and cleaned from the large-scale manually annotated FewNERD dataset.

Gold Test Set

From the silver test set, 1000 sentences have been randomly selected for ten languages and annotated by at least two annotators. All volunteer annotators have a minimum undergraduate degree and are selected based on their native language. The good quality of the manually annotated gold dataset can be ascertained with high inter-annotator agreement (IAA(κ)), above 0.8 for each language (Table 4).

Results & Analysis

This section covers the analyses performed on the SampurNER dataset, including human evaluations, analysis of PLMs' performance fine-tuned on SampurNER, cross-lingual zero-shot evaluation, impact of multilingualism and script similarity, entity error analysis, and ablation study. All experiments were first run on both silver and gold test sets,

	mBERT on silver test set						IndicBERTv2 on silver test set						IndicBERTv2 on gold test set						IAA
	Macro			Micro			Macro			Micro			Macro			Micro			
Ln	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	κ
en†	59.1	63.9	61.3	65.2	69.1	67.1	58.2	62.5	60.1	64.0	68.2	66.0				–			–
as	54.1	56.7	55.3	60.1	63.7	61.8	56.7	60.0	58.2	62.6	66.2	64.4	54.3	57.4	55.7	60.1	63.6	61.8	0.81
bn	57.3	59.5	58.3	62.3	65.5	63.8	59.0	62.4	60.5	64.4	67.7	66.0	56.5	59.7	57.9	61.8	65.0	63.5	0.83
brx	56.4	59.1	57.6	61.2	64.7	62.9	58.6	61.6	59.9	63.6	66.7	65.1	56.1	59.0	57.3	61.1	64.1	62.6	0.80
doi	52.5	55.4	53.6	58.1	61.8	59.9	54.1	57.6	55.5	59.3	63.6	61.4				–			–
gom	51.2	54.1	52.5	57.7	61.8	59.7	53.8	56.4	54.8	60.0	63.9	61.9				–			–
gu	55.9	57.8	56.7	62.2	65.2	63.7	58.7	61.3	59.8	64.8	67.9	66.3				–			–
hi	54.8	57.4	56.0	59.7	63.2	61.4	55.8	58.7	57.1	61.1	63.8	62.4	53.4	56.2	54.7	58.7	61.3	60.0	0.86
kn	55.2	58.2	56.5	61.4	64.9	63.1	58.0	61.2	59.3	63.8	67.0	65.4				–			–
ks	51.6	54.3	52.8	58.4	61.8	60.1	52.6	55.9	54.0	59.3	63.3	61.2				–			–
mai	51.4	53.8	52.5	58.4	61.9	60.1	53.5	57.0	55.0	60.2	64.3	62.2				–			–
ml	53.5	55.6	54.4	58.8	62.1	60.4	56.5	59.2	57.8	62.0	65.4	63.6				–			–
mni	18.6	4.4	6.4	25.3	7.3	11.3	49.7	50.1	49.6	55.6	58.8	57.2				–			–
mr	57.5	59.6	58.5	63.4	66.5	64.9	58.9	61.8	60.2	64.8	68.1	66.4	56.4	59.2	57.6	62.2	65.4	63.9	0.86
ne	58.7	61.2	59.8	64.2	67.1	65.6	60.7	63.1	61.7	65.7	68.5	67.1	58.1	60.4	59.1	63.1	65.8	64.5	0.83
or	24.9	9.1	12.4	32.0	11.4	16.8	57.2	60.6	58.8	63.1	66.4	64.7				–			–
pa	51.9	54.2	52.9	58.7	61.9	60.3	54.6	58.1	56.2	61.2	65.1	63.1				–			–
sa	53.4	55.5	54.2	59.6	63.0	61.3	53.4	57.1	55.0	60.7	64.7	62.6	51.1	54.7	52.7	58.3	62.1	60.2	0.80
sat	38.2	12.0	17.1	43.0	15.5	22.8	41.3	42.3	41.4	50.6	51.8	51.2				–			–
sd	35.6	34.9	35.0	43.3	42.7	43.0	50.7	53.8	52.1	57.9	61.6	59.7				–			–
ta	53.5	55.9	54.4	59.9	63.2	61.5	54.9	57.7	56.1	61.3	64.8	63.0	52.6	55.2	53.7	58.9	62.2	60.6	0.83
te	53.9	55.3	54.4	59.5	62.8	61.1	56.5	58.9	57.4	62.4	65.8	64.1	54.1	56.4	55.0	59.9	63.2	61.6	0.83
ur	55.1	58.1	56.5	60.3	63.8	62.0	55.2	58.7	56.8	60.7	64.5	62.6	52.8	56.2	54.4	58.3	61.9	60.3	0.82

Table 4: Performance of models fine-tuned on SampurNER train sets, evaluated on the silver test set (22 languages) and gold test set (10 languages). en† denotes the FewNERD dataset. IAA(κ): Inter-annotator agreement of SampurNER gold test set.

which differed by only $\pm 5\%$ (Table 4), and since silver test set exists for all 22 languages while gold test set covers only 10 languages, subsequent analyses use the silver test set.

Human Evaluation

To ensure the reliability and quality of the generated dataset, human evaluation was conducted as per the XSTS methodology (Licht et al. 2022), which focuses on meaning preservation by having annotators rate translated sentences from 1 to 5. Following Gala et al. (2023); Brahma, Maurya, and Desarkar (2023), two native-speaker annotators rated 100 sentences per language. All annotators are at least undergraduate degree holders. The consistently high XSTS scores (Table 3) confirm the good quality of the SampurNER dataset.

Performance of PLMs on Unseen Languages

As shown in the Table 4, the mBERT model fine-tuned on SampurNER dataset has performed fairly well in the Indian languages it has been pre-trained on: Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Punjabi, Tamil, Telugu, and Urdu. It is interesting to see that on some unseen languages such as Assamese, Bodo, Dogri, Konkani, Maithili, Sanskrit, and Kashmiri, fine-tuned mBERT model performed well. This is due to script similarity of Bengali with Assamese, and the use of Perso-Arabic script for Kashmiri and Urdu, and Devanagari script for Hindi and other unseen languages mentioned above. But, for the unseen languages having a different script, such as Manipuri, Odia, Santali etc., the fine-tuned mBERT model could not perform well in terms of F1-score. However, such variations are not observed in the case of IndicBERTv2 (Table 4) as the PLM is pre-trained on all 22 scheduled Indian languages. Hence, all the 22 fine-tuned IndicBERTv2 models on SampurNER datasets performed better than corresponding fine-tuned mBERT models for the respective language.

Cross-Lingual Zero-Shot Analysis

We have performed cross-lingual zero-shot analysis for every language pair, including the English dataset from FewNERD. As shown in Figure 2, the models are fine-tuned on the dataset of respective languages and tested on the test set of other languages. The pre-trained knowledge of mBERT in English is very proficient, due to which the zero-shot performance is very high across all languages. In fact, it is sometimes higher than the language-specific fine-tuned models for unseen languages such as Manipuri, Santali, Sindhi etc. The impact of script similarity depends on the languages covered during pre-training of the PLM. For example, when mBERT is fine-tuned with Assamese, on which it was not pre-trained, the zero-shot performance on Bengali is 49 micro-F1 score. Whereas, when it is fine-tuned with Bengali, on which it was pre-trained, the zero-shot performance in Assamese drops down to a 28 micro-F1 score, although both languages are written using Bengali-Assamese script. Whereas such variations are not observed when simi-

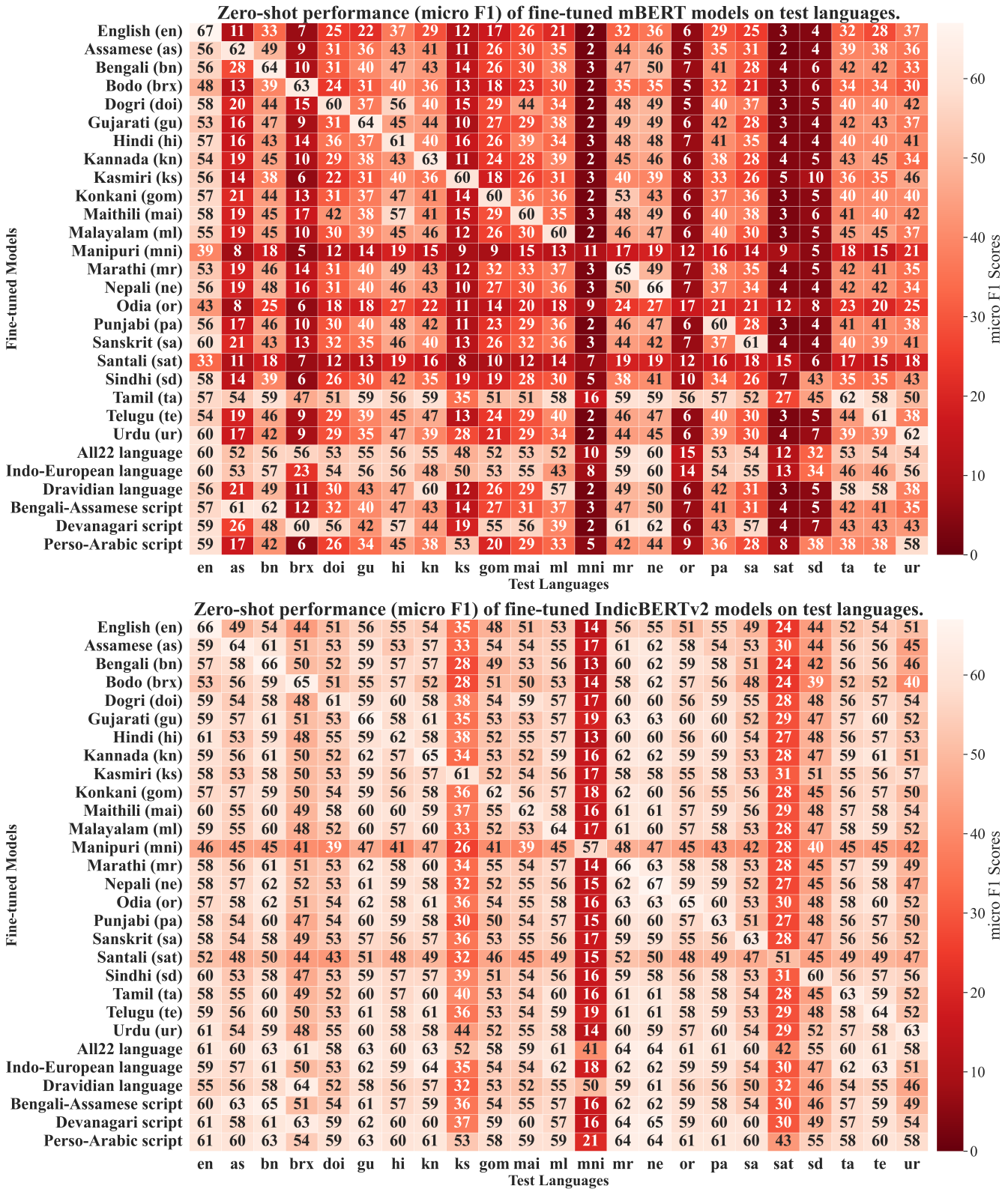

Figure 2: Zero-shot performance (micro F1) of fine-tuned mBERT and IndicBERTv2 models on different test languages.

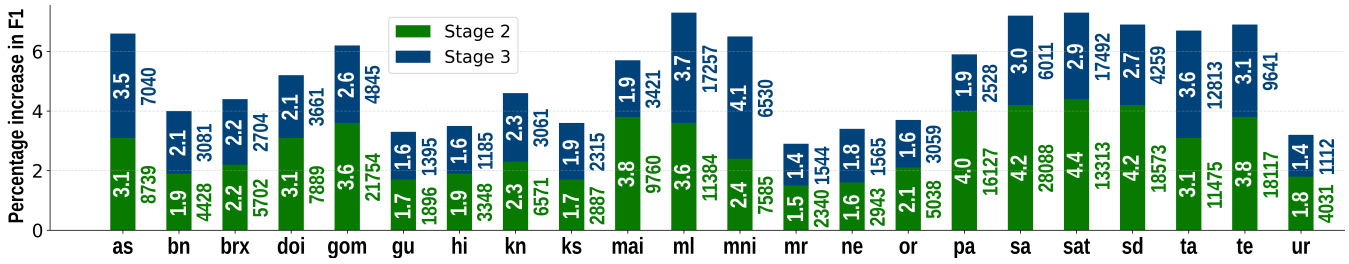

Figure 3: Impact of Stage 2 and Stage 3 on percentage increase in F1 scores. Values inside bars indicate percentage increase of F1-scores after Stage 2 and Stage 3. Values outside indicate the number of sentences removed at each stage.

lar tests are performed in fine-tuned IndicBERTv2 due to its pre-training on all 22 scheduled Indian languages.

Multilingualism & Script Similarity

Our analysis extends to evaluating multilingualism, script similarity, and the influence of language families. We constructed balanced datasets comprising multiple languages, ensuring an equal number of samples per language. The **All22 language** set includes all 22 scheduled Indian languages, while the **Indo-European language** set consists of Indo-European languages: Assamese, Bengali, Dogri, Konkani, Gujarati, Hindi, Kashmiri, Maithili, Marathi, Nepali, Odia, Punjabi, Sanskrit, Sindhi, and Urdu. Similarly, the **Dravidian language** set includes Kannada, Malayalam, Tamil, and Telugu.

Based on script similarity, we categorized languages as follows: Assamese and Bengali under **Bengali-Assamese script**, Bodo, Dogri, Konkani, Hindi, Maithili, Marathi, Nepali, and Sanskrit under **Devanagari script**, and Kashmiri, Sindhi, and Urdu under **Perso-Arabic script**. As expected, the best results for each language occur when it is fine-tuned only on that language (Figure 2). Performance also improves when languages share the same script or belong to the same language family, while it typically degrades when the language differs in script or linguistic lineage.

However, two intriguing observations emerge. Fine-tuning IndicBERTv2 only on Perso-Arabic-script languages yields better zero-shot Santali performance than the All22 model, despite the latter being trained with Santali samples. Second, Manipuri, a Sino-Tibetan language, attains superior zero-shot results when the model is fine-tuned on Dravidian languages rather than on the All22 model.

Error Analysis

FgNER is very crucial, as depending on the context, an entity's mention type may vary significantly. Therefore, we have analyzed the errors in two different approaches. First, the details of entity errors in terms of the percentage of predicted entities are shown in Table 3. The common errors that occur include the Boundary Mention error (such as 'Good Dinosaur' is marked as *Film* instead of 'The Good Dinosaur'), Entity Type error (e.g. 'Fainting' is categorized as *Disease* instead of *Symptom*) and Spurious Entity errors (such as 'red' is marked as an entity although the entity type *Color* is not defined in FewNERD taxonomy).

Moreover, we have analyzed the often co-predicted fine-grained types. From Table 3, we have selected Santali for this analysis, as this language has the highest percentage of mismatched entity types. As shown in the extended version, the fine types *actor*, *artist/author*, and *politician* are confused with *person-other*. Similarly, *location-GPE* is sometimes confused with *location-other*. Apart from such closely related entity types, most of the other fine entity types are learned by the models without much confusion.

Ablation Study

The impact of Stage 2 and Stage 3 are discussed through the Figure 3. The removal of erroneous sentences due to entity-anchors mismatch in Stage 2 improves the F1 scores by 2.8% on average. Similarly, the effect of the removal of sentences having a mismatch of total entity counts in Stage 3 is imminent, with an average improvement of F1 scores by 2.4%. Both stages improve the recall and eventually improve the F1 scores.

Conclusion

We introduced the entity-anchored machine translation (EaMaTa) framework to create SampurNER, the first FgNER dataset for 22 scheduled Indian languages comprising over 153k annotated sentences and 354k entities. Our analyses confirm the effectiveness of the EaMaTa framework and the high quality of the SampurNER dataset. Extensive experiments, including cross-lingual zero-shot analyses, underscore the importance of language-specific pre-training and task-specific fine-tuning of PLMs. Fine-tuning IndicBERTv2 on SampurNER yields strong FgNER performance, and the All22 model exhibits robust cross-lingual capabilities, making it well-suited for broad multilingual applications. We believe that EaMaTa, SampurNER, and the associated fine-tuned models will advance NER across Indian languages and support wider multilingual research.

However, translation-based dataset creation risks introducing errors and is limited by the source dataset's size and diversity. Although the FewNERD dataset spans multiple domains, the SampurNER dataset still lacks certain linguistic, geographical, and cultural nuances specific to the Indian subcontinent. In the future, we aim to develop culturally enriched datasets spanning all 22 languages. Furthermore, we plan to systematically evaluate the FgNER capabilities of diverse large language models across these languages.

References

Abhishek, A.; Taneja, S. B.; Malik, G.; Anand, A.; and Awekar, A. 2019. Fine-grained Entity Recognition with Reduced False Negatives and Large Type Coverage. In *AKBC*, 1–17.

Al-Rfou, R.; Kulkarni, V.; Perozzi, B.; and Skiena, S. 2015. Polyglot-NER: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, 586–594.

Bing Translator. 2025. Microsoft. https://www.bing.com/translator/. Accessed: 2025-01-06.

Brahma, M.; Maurya, K.; and Desarkar, M. 2023. SelectNoise: Unsupervised Noise Injection to Enable Zero-Shot Machine Translation for Extremely Low-resource Languages. In *Findings of the ACL: EMNLP 2023*, 1615–1629.

Choi, E.; Levy, O.; Choi, Y.; and Zettlemoyer, L. 2018. Ultra-Fine Entity Typing. In *Proceedings of the ACL*, 87–96.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the ACL*, 8440–8451.

Del Corro, L.; Abujabal, A.; Gemulla, R.; and Weikum, G. 2015. Finet: Context-aware fine-grained named entity typing. In *Proceedings of the EMNLP*, 868–878.

Devi, S. L.; Rao, P. R.; Malarkodi, C.; and Ram, R. V. S. 2014. Indian language NER annotated FIRE 2014 corpus (FIRE 2014 NER corpus). *Named-Entity Recognition Indian Languages FIRE*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the NAACL-HLT*, 4171–4186.

Ding, N.; Xu, G.; Chen, Y.; Wang, X.; Han, X.; Xie, P.; Zheng, H.; and Liu, Z. 2021. Few-NERD: A Few-shot Named Entity Recognition Dataset. In *Proceedings of the ACL-IJCNLP*, 3198–3213.

Doddapaneni, S.; Aralikatte, R.; Ramesh, G.; Goyal, S.; Khapra, M. M.; Kunchukuttan, A.; and Kumar, P. 2023. Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages. In *Proceedings of the ACL*, 12402–12426.

Doddington, G. R.; Mitchell, A.; Przybocki, M. A.; Ramshaw, L. A.; Strassel, S. M.; and Weischedel, R. M. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Proceedings of the LREC*, volume 2, 837–840.

Ekbal, A.; and Saha, S. 2011. A multiobjective simulated annealing approach for classifier ensemble: Named entity recognition in Indian languages as case studies. *Expert Systems with Applications*, 38(12): 14760–14772.

Fetahu, B.; Chen, Z.; Kar, S.; Rokhlenko, O.; and Malmasi, S. 2023a. MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition. In *Findings of the ACL: EMNLP 2023*, 2027–2051.

Fetahu, B.; Kar, S.; Chen, Z.; Rokhlenko, O.; and Malmasi, S. 2023b. SemEval-2023 Task 2: Fine-grained Multilingual Named Entity Recognition (MultiCoNER 2). In *Proceedings of the SemEval-2023*, 2247–2265.

Gala, J.; Chitale, P. A.; Raghavan, A. K.; Gumma, V.; Doddapaneni, S.; M, A. K.; Nawale, J. A.; Sujatha, A.; Puduppully, R.; Raghavan, V.; Kumar, P.; Khapra, M. M.; Dabre, R.; and Kunchukuttan, A. 2023. IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages. *Transactions on Machine Learning Research*.

Gali, K.; Surana, H.; Vaidya, A.; Shishtla, P. M.; and Sharma, D. M. 2008. Aggregating machine learning and rule based heuristics for named entity recognition. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 25–32.

Gillick, D.; Lazic, N.; Ganchev, K.; Kirchner, J.; and Huynh, D. 2014. Context-dependent fine-grained entity type tagging. *arXiv preprint arXiv:1412.1820*.

Google Translate. 2025. Google. https://cloud.google.com/translate. Accessed: 2025-01-06.

Gupta, S.; and Bhattacharyya, P. 2010. Think globally, apply locally: using distributional characteristics for Hindi named entity identification. In *Proceedings of the 2010 Named Entities Workshop*, 116–125.

Huang, L.; Liu, H.; Gao, Q.; Yu, J.; Liu, G.; and Chen, X. 2025. Adversity-aware Few-shot Named Entity Recognition via Augmentation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 24132–24140.

IndicTrans2. 2023. AI4Bharat. https://ai4bharat.iitm.ac.in/areas/model/NMT/IndicTrans2. Accessed: 2025-01-06.

Kaushik, P.; and Anand, A. 2025. CLASSER: Cross-lingual Annotation Projection enhancement through Script Similarity for Fine-grained Named Entity Recognition. In *Proceedings of the IJCNLP-AACL*.

Kaushik, P.; Mishra, S.; and Anand, A. 2025. TAFSIL: Taxonomy Adaptable Fine-grained Entity Recognition through Distant Supervision for Indian Languages. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3753–3763.

Licht, D.; Gao, C.; Lam, J.; Guzman, F.; Diab, M.; and Koehn, P. 2022. Consistent Human Evaluation of Machine Translation across Language Pairs. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, 309–321.

Ling, X.; and Weld, D. S. 2012. Fine-grained entity recognition. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 94–100.

Litake, O.; Sabane, M. R.; Patil, P. S.; Ranade, A. A.; and Joshi, R. 2022. L3cube-mahaner: A marathi named entity recognition dataset and bert models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, 29–34.

Liu, L.; Ding, B.; Bing, L.; Joty, S.; Si, L.; and Miao, C. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the ACL-IJCNLP*, 5834–5846.

Liu, Y. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Ma, J.-Y.; Gu, J.-C.; Qi, J.; Ling, Z.; Liu, Q.; and Zhao, X. 2023a. USTC-NELSLIP at SemEval-2023 Task 2: Statistical Construction and Dual Adaptation of Gazetteer for Multilingual Complex NER. In *Proceedings of the SemEval-2023*, 651—-659.

Ma, L.; Lu, K.; Che, T.; Huang, H.; Gao, W.; and Li, X. 2023b. PAI at SemEval-2023 Task 2: A Universal System for Named Entity Recognition with External Entity Information. In *Proceedings of the SemEval-2023*, 744–750.

Malmasi, S.; Fang, A.; Fetahu, B.; Kar, S.; and Rokhlenko, O. 2022. MultiCoNER: A Large-scale Multilingual Dataset for Complex Named Entity Recognition. In *Proceedings of the COLING*, 3798–3809.

Mayhew, S.; Tsai, C.-T.; and Roth, D. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the EMNLP*, 2536–2545.

Mhaske, A.; Kedia, H.; Doddapaneni, S.; Khapra, M. M.; Kumar, P.; Murthy, R.; and Kunchukuttan, A. 2023. Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages. In *Proceedings of the ACL*, 10441–10456.

Murty, S.; Verga, P.; Vilnis, L.; and McCallum, A. 2017. Finer grained entity typing with typenet. *arXiv preprint arXiv:1711.05795*.

Narzary, S.; Brahma, A.; Nandi, S.; and Som, B. 2024. Deep Learning based Named Entity Recognition for the Bodo Language. *Procedia Computer Science*, 235: 2405–2421.

Niraula, N.; and Chapagain, J. 2022. Named entity recognition for Nepali: data sets and algorithms. In *The International FLAIRS Conference Proceedings*, volume 35, 1–6.

Pan, X.; Zhang, B.; May, J.; Nothman, J.; Knight, K.; and Ji, H. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the ACL*, 1946–1958.

Pathak, D.; Nandi, S.; and Sarmah, P. 2022. AsNER-Annotated Dataset and Baseline for Assamese Named Entity recognition. In *Proceedings of the LREC*, 6571–6577.

Priyadarshi, A.; and Saha, S. K. 2021. The first named entity recognizer in Maithili: Resource creation and system development. *Journal of Intelligent & Fuzzy Systems*, 41(1): 1083–1095.

Rau, L. F. 1991. Extracting company names from text. In *Proceedings the Seventh IEEE Conference on Artificial Intelligence Application*, 29–30. IEEE Computer Society.

Reddy, A. J.; Adusumilli, M.; Gorla, S. K.; Neti, L. B. M.; and Malapati, A. 2018. Named entity recognition for telugu using lstm-crf. In *WILDRE4–4th Workshop on Indian Language Data: Resources and Evaluation*, volume 6, 1–5.

Ruder, S.; Constant, N.; Botha, J.; Siddhant, A.; Firat, O.; Fu, J.; Liu, P.; Hu, J.; Garrette, D.; Neubig, G.; et al. 2021. XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation. In *Proceedings of the EMNLP*, 10215–10245.

Sekine, S.; Sudo, K.; and Nobata, C. 2002. Extended Named Entity Hierarchy. In *Proceedings of the LREC*, 1818–1824.

Sierra-Múnera, A.; Westphal, J.; and Krestel, R. 2023. Efficient Ultrafine Typing of Named Entities. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 205–214. IEEE.

Singh, A. K. 2008. Named entity recognition for south and south east Asian languages: taking stock. In *Proceedings of the IJCNLP-08 workshop on named entity recognition for South and South East Asian languages*, 5–16.

Tan, Z.; Huang, S.; Jia, Z.; Cai, J.; Li, Y.; Lu, W.; Zhuang, Y.; Tu, K.; Xie, P.; and Huang, F. 2023. DAMO-NLP at SemEval-2023 Task 2: A Unified Retrieval-augmented System for Multilingual Named Entity Recognition. In *Proceedings of the SemEval 2023*, 2014—2028.

Tulajiang, P.; Sun, Y.; Zhang, Y.; Le, Y.; Xiao, K.; and Lin, H. 2025. A Bilingual Legal NER Dataset and Semantics-Aware Cross-Lingual Label Transfer Method for Low-Resource Languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*

Ugawa, A.; Tamura, A.; Ninomiya, T.; Takamura, H.; and Okumura, M. 2018. Neural machine translation incorporating named entity. In *Proceedings of the COLING*, 3240–3250.

UNESCO, A. 2017. UNESCO Atlas of the World's Languages in Danger.

Vavekanand, R.; Das, B.; and Kumar, T. 2025. DAugSindhi: a data augmentation approach for enhancing Sindhi language text classification. *Discover Data*, 3(1): 22.

Venkataramana, R. M.; Bhattacharjee, P.; Sharnagat, R.; Khatri, J.; Kanojia, D.; and Bhattacharyya, P. 2022. HiNER: A large Hindi Named Entity Recognition Dataset. In *Proceedings of the LREC*, 4467–4476.

Weischedel, R.; and Brunstein, A. 2005. BBN pronoun coreference and entity type corpus. *Linguistic Data Consortium, Philadelphia*, 112.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the EMNLP: system demonstrations*, 38–45.

Yang, J.; Huang, S.; Ma, S.; Yin, Y.; Dong, L.; Zhang, D.; Guo, H.; Li, Z.; and Wei, F. 2022. CROP: Zero-shot Cross-lingual Named Entity Recognition with Multilingual Labeled Sequence Translation. In *Findings of the ACL: EMNLP 2022*, 486–496.

Yosef, M. A.; Bauer, S.; Hoffart, J.; Spaniol, M.; and Weikum, G. 2012. Hyena: Hierarchical type classification for entity names. In *Proceedings of the COLING: Posters*, 1361–1370.