FiNERVINER: Fine-grained Named Entity Recognition for Vulnerable languages of India's North Eastern Region

Prachuryya Kaushik, Ashish Anand
Indian Institute of Technology Guwahati
Guwahati, Assam, India
{k.prachuryya, anand.ashish}@iitg.ac.in

Abstract

Named entity recognition (NER), particularly fine-grained NER (FgNER), extracts domain-specific entity information for Natural Language Processing (NLP) applications such as knowledge base construction and relation extraction. While manual annotation for creating relevant data is expensive, distant supervision often produces noisy data. Moreover, resources for coarse-grained and fine-grained NER in Indian languages, particularly in the vulnerable languages of India's North Eastern Region, remain scarce. This work aims at creating such a resource for three vulnerable languages: *Bodo/Boro (brx)*, *Manipuri/Meitei (mni)*, and *Mizo/Lushai (lus)*, which are regarded as official languages in three Indian states and spoken by more than six million people across five countries in South and Southeast Asia. We use annotations projection from high-resource FgNER datasets using source-to-target parallel corpora and a projection tool built on a multilingual encoder. The dataset comprises over 198k sentences, 282k entities, and 2.8M tokens in each low-resource language. Our thorough analyses validate the dataset's high quality. We further explore zero-shot and cross-lingual settings, examining the impact of script similarity and multilingualism in cross-lingual FgNER performance. The dataset, expert detector models, the agentic tool, and the interactive web application are available as open-source resources at: https://hf.co/collections/prachuryyaIITG/finerviner.

Keywords: Named Entity Recognition, Fine-grained Named Entity Recognition, Annotation Projection, Indian languages, Low resource languages, Vulnerable languages, Multilingual, Information Extraction

1. Introduction

Structured knowledge extraction from unstructured text underpins many downstream applications, including recommendation systems, knowledge-base construction, and relation extraction. A relevant foundational task is Named Entity Recognition (NER). NER identifies and classifies mentions of entities such as persons, locations, and organizations. The development of NER has progressed from early rule-based systems Rau (1991), through the collaborative contributions of dedicated events (Grishman and Sundheim, 1996; Chinchor et al., 1998; Satoshi, 2000; Tjong Kim Sang, 2002; Doddington et al., 2004; Santos et al., 2006), to the powerful neural architectures today (Zhang et al., 2019; Yan et al., 2023; Zhou et al., 2023; Xue et al., 2024; Zhang et al., 2024; Huang et al., 2025). Conventional coarse-grained NER categories often prove insufficient for applications demanding greater specificity. For instance, a generic "Product" tag may fail to capture more detailed information such as "Vehicle" or "Food" (Sekine and Nobata, 2004). This requirement for finer distinctions has driven the emergence of Fine-grained Named Entity Recognition (FgNER), where the type and granularity of entities are tailored to specific domains and application requirements (Ling and Weld, 2012; Choi et al., 2018).

Despite considerable progress in coarse-grained NER for Indian languages (Murthy et al., 2022; Pathak et al., 2022; Litake et al., 2022; Mhaske et al., 2023), the development of FgNER resources for these languages remains in its early stages. Recent initiatives, such as the MultiCoNER2 shared task at SemEval-2023 (Fetahu et al., 2023), have introduced FgNER datasets for Hindi and Bengali through translated English annotations (Fetahu et al., 2023). Similarly, the TAFSIL initiative generated noisy FgNER data for six other Indian languages by mining Wikipedia links and Wikidata (Kaushik et al., 2025). Despite these important advances, comprehensive and high-quality FgNER resources remain scarce for most low-resource Indian languages.

To address this gap, we create FiNERVINER: Fine-grained Named Entity Recognition dataset for Vulnerable languages of India's North Eastern Region. We have projected the FgNER annotations of the MultiCoNER2 English dataset to the target languages, utilizing the parallel corpora and a multilingual encoder-based annotation projection and word alignment tool (Dou and Neubig, 2021; García-Ferrero et al., 2022). The vulnerable languages (UNESCO, 2017) considered are: *Bodo/Boro (brx)*, *Manipuri/Meitei (mni)*, and *Mizo/Lushai (lus)*, which are regarded as one of the official languages in the Indian states of Assam, Manipur, and Mizoram, respectively. These Sino-Tibetan languages are spoken by more than six million people across five countries in South and Southeast Asia. To the best of our knowledge, this is the first FgNER dataset

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7655–7667
11–16 May 2026.

created for these three vulnerable languages. Our contributions can be summarized as follows:

1. Construction of a large-scale FgNER dataset FiNERVINER comprising over a total of 700k sentences, 963k entity mentions, and 11M tokens for three low-resource and vulnerable Indian languages: Bodo/Boro (brx), Manipuri/Meitei (mni), and Mizo/Lushai (lus).

2. Creation of high-quality human-annotated test sets consisting of 1000 sentences for each language with inter-annotator agreement (κ) above 0.81.

3. Our rigorous analyses establish the good quality of the generated dataset.

4. Zero-shot and cross-lingual analysis to examine the influence of multilingualism and script similarities on cross-lingual FgNER performance.

2. Related Works

Sekine et al. (2002) pioneered fine-grained entity classification with a hierarchy of 150 types, and Ling and Weld (2012) advanced fine-grained named entity recognition (FgNER) by defining 113 types in a two-level hierarchy. FgNER datasets have been introduced in varied entity type count: e.g. ACE has 52 types (Doddington et al., 2004), BBN has 93 (Weischedel and Brunstein, 2005), HYENA has 505 (Yosef et al., 2012), and OntoNotes has 88 (Gillick et al., 2014). Meanwhile, large-scale resources like WikiSense (Chang et al., 2009), FINET (Del Corro et al., 2015),TypeNet (Murty et al., 2017), and UFET (Choi et al., 2018) offer thousands of categories. Abhishek et al. (2019) improved quality with language-specific heuristics and refined selection, whereas Ding et al. (2021) provides a large, manually annotated dataset covering 66 fine-grained types.

Yarowsky and Ngai (2001) pioneered entity projection using parallel corpora and word alignment, yet zero-shot transfer remains limited for linguistically distant languages (Wu and Dredze, 2019), especially when structural and word order differences are significant (Karthikeyan et al., 2020). Annotation projection has been widely used for dependency parsing (Rai and Chatterji, 2022), relation extraction (Nag et al., 2023), and NER (Prabhakar et al., 2022; Tulajiang et al., 2025). To address the need for larger, more diverse datasets in Indian languages, Naamapadam (Mhaske et al., 2023) was developed for 11 Indian languages via annotation projection from English NER data, utilizing Samanantar parallel corpora (Ramesh et al., 2022) and word alignment techniques (Ruder et al., 2021; Och and Ney, 2003).

An early step in developing NER resources for Indian languages was the IJCNLP-2008 dataset (Singh, 2008), which provided data for Hindi, Bengali, Oriya, Telugu, and Urdu and became a key resource in early Hindi NER research (Gali et al., 2008; Saha et al., 2008a,b; Gupta and Bhattacharyya, 2010; Bhagavatula et al., 2012). The FIRE-2014 dataset (Devi et al., 2014) expanded coverage to Tamil and Malayalam using sources like Wikipedia, blogs, and forums. Polyglot NER further contributed by covering over a hundred languages (Al-Rfou et al., 2015). WikiANN spans 282 languages, including 16 Indian languages (Pan et al., 2017). Subsequent manual annotation efforts produced datasets such as HiNER (Murthy et al., 2022), AsNER (Pathak et al., 2022), MahaNER (Litake et al., 2022), EverestNER (Niraula and Chapagain, 2022), OurNepali (Singh et al., 2019), NER in Bengali (Ekbal et al., 2008), NER in Telugu (Reddy et al., 2018), Maithili (Mundotiya et al., 2023), Bishnupriya Manipuri (Laishram et al., 2020; Jimmy et al., 2023), Bodo (Narzary et al., 2024), etc.

MultiCoNER-1 (Malmasi et al., 2022) and MultiCoNER-2 (Fetahu et al., 2023) created datasets in Hindi and Bengali, by translating English coarse-grained and fine-grained NER datasets, respectively. In the SemEval-2023 shared task, multiple teams (Ma et al., 2023a,b; Tan et al., 2023; García-Ferrero et al., 2023) further enhanced FgNER resources in Hindi and Bengali. Some of the most recent works in FgNER include datasets created for several Indian languages, including a couple of vulnerable languages (Kaushik et al., 2025; Kaushik and Anand, 2025, 2026a,b). Despite these advancements, resources in vulnerable Indian languages remain scarce, underscoring the need for more multilingual FgNER datasets.

3. About the languages

Bodo (ISO 639-3 code: brx) is an official language of India and in the Indian state of Assam. It is primarily spoken in the Bodoland Territorial Region, Bangladesh, and Nepal. Bodo, written using the Devanagari script, is recognized as a vulnerable language (UNESCO, 2017) as there are around 1.5 million native speakers in India (Census of India, 2011). Mizo, also known as Lushai (ISO 639-3 code: lus), is an official language in the Indian state Mizoram, although it is yet to be included as a scheduled language of India (The Constitution of India, 1950). Mizo is written using the Latin script and spoken in some regions of Myanmar and Bangladesh as well. Manipuri (ISO 639-3 code: mni), also known as Meitei (or Meiteilon), is an official language in the Indian state of Manipur. This scheduled Indian language is written using Meitei script, and spoken by more than 1.7 million people (Census of India, 2011) across India, Bangladesh, and Myanmar.

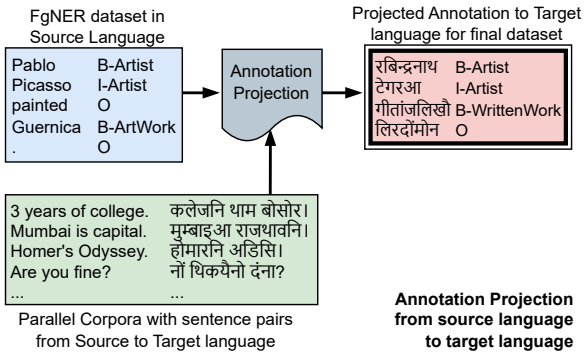

Figure 1: Annotation projection from FgNER dataset in source language to target language, utilizing parallel corpora, along with annotation projection and word alignment tool.

4. Annotation Projection

As illustrated in the Figure 1, the annotation projection method is a three-component process that leverages a high-resource source language (src) to project annotations to the target language (tgt). For the source language src, the FgNER dataset $D_{src} = \{(s_{src}^{(k)}, y_{src}^{(k)})\}_{k=1}^{N}$ consist of N samples where each sentence s_{src} is paired with its corresponding annotation sequence y_{src}. The source language provides both an annotated FgNER dataset D_{src} and a parallel corpus $P = \{(s_{src}^{(k)}, s_{tgt}^{(k)})\}_{k=1}^{\overline{N}}$ of size $\overline{N}$ that comprises sentence pairs aligned with the target language (tgt).

The projection process is executed for each parallel pair $(s_{src}, s_{tgt}) \in P$ and involves the following steps:

1. **Embedding Computation**: A pretrained multilingual encoder M is fine-tuned on the parallel corpus P to produce contextual embeddings $\mathbf{E}_{src} = M(s_{src})$ and $\mathbf{E}_{tgt} = M(s_{tgt})$ for the source and target sentences, respectively.

2. **Alignment Mapping**: A word alignment service, $\mathcal{A}(s_{src}, s_{tgt})$, leverages these embeddings to generate a mapping, map, from source-sentence tokens to target-sentence tokens:

$map = \mathcal{A}(s_{src}, s_{tgt})$ using $\mathbf{E}_{src}$ & $\mathbf{E}_{tgt}$.

3. **Annotation Projection**: The source annotation sequence y_{src} is projected to the target sentence by setting the annotation for each token j in s_{tgt} as $\hat{y}_{tgt}^{(j)} = y_{src}^{(map^{-1}(j))}$. This operation is performed iteratively for all tokens in the target sentence, yielding a complete annotation sequence $\hat{y}_{tgt}$.

4. **Dataset creation for target language**: The newly created pair $(s_{tgt}, \hat{y}_{tgt})$ is included to create the final FgNER dataset in the target language, D_{tgt}.

Table 1: FiNERVINER gold dataset. **IAA** (κ) gives the inter-annotator agreement.

Language	Bodo	Manipuri	Mizo
IAA (κ)	0.875	0.811	0.821

4.1. Implementation details

We use the MultiCoNER2 (Fetahu et al., 2023) dataset, which provides 33 fine-grained entity types across 12 languages. For our setup, English served as the source language (src), while the target languages (tgt) are Bodo (brx), Manipuri (mni), and Mizo (lus). To obtain the necessary parallel corpora (P), we used the BPCC corpora (Gala et al., 2023) for Manipuri and Bodo, which is augmented with additional resources from Islam (2018); Islam et al. (2018). The parallel corpus used for Mizo is the Mizo–English parallel corpus (Haulai and Hussain, 2023). Following García-Ferrero et al. (2022), we adopted AWESoME align (Dou and Neubig, 2021) as $\mathcal{A}$, fine-tuned on the English MultiCoNER2 dataset (D_{src}). Finally, we used different multilingual encoders (M) for our target languages. For Bodo and Manipuri, we fine-tuned IndicBERTv2 (Doddapaneni et al., 2023), as it is the only encoder pre-trained on all three languages (English, Bodo, and Manipuri). Since the Mizo language is written using the Latin script (similar to English), the MizBERT (Lalramhluna et al., 2024) utilized the BERT-base (Devlin et al., 2019) tokenizer on English and then pre-trained on a large-scale Mizo corpus (Lalramhluna et al., 2024). Therefore, we have used MizBERT as multilingual encoders (M) for our target language Mizo. Through the implementation details as mentioned in this section, the FiNERVINER dataset is generated, which is discussed in the next sections.

5. FiNERVINER dataset

As shown in Table 2, the created FiNERVINER dataset consists of more than 198 thousand sentences, 282 thousand entity mentions, and 2.8 million tokens in each of the three low-resource vulnerable Indian languages. We randomly pick 1000 sentences from the dataset for manual annotation to create the test set (Table 1). From the rest of the dataset, 10% is considered as the development set and the remaining as the training set.

5.1. Gold dataset

Volunteer annotators, who have a minimum education of an undergraduate degree, are chosen based on their mother tongue. As shown in the Table 1, we compute inter-annotator agreement (IAA) on the 1000 sentences annotated by two annotators

Table 2: FiNERVINER dataset statistics. **Sents**, **Ents** and **Tokens** means number of Sentences, Entities and Tokens respectively.

Language	Train Set			Development Set			Test Set		
	Sents	Ents	Tokens	Sents	Ents	Tokens	Sents	Ents	Tokens
Bodo (brx)	212,835	302,713	2,958,455	23,649	33,808	329,145	1,000	1,423	14,082
Manipuri (mni)	177,224	252,767	2,515,386	19,692	28,143	279,681	1,000	1,384	14,330
Mizo (lus)	239,813	302,713	4,422,373	26,646	38,330	484,212	1,000	1,426	18,765

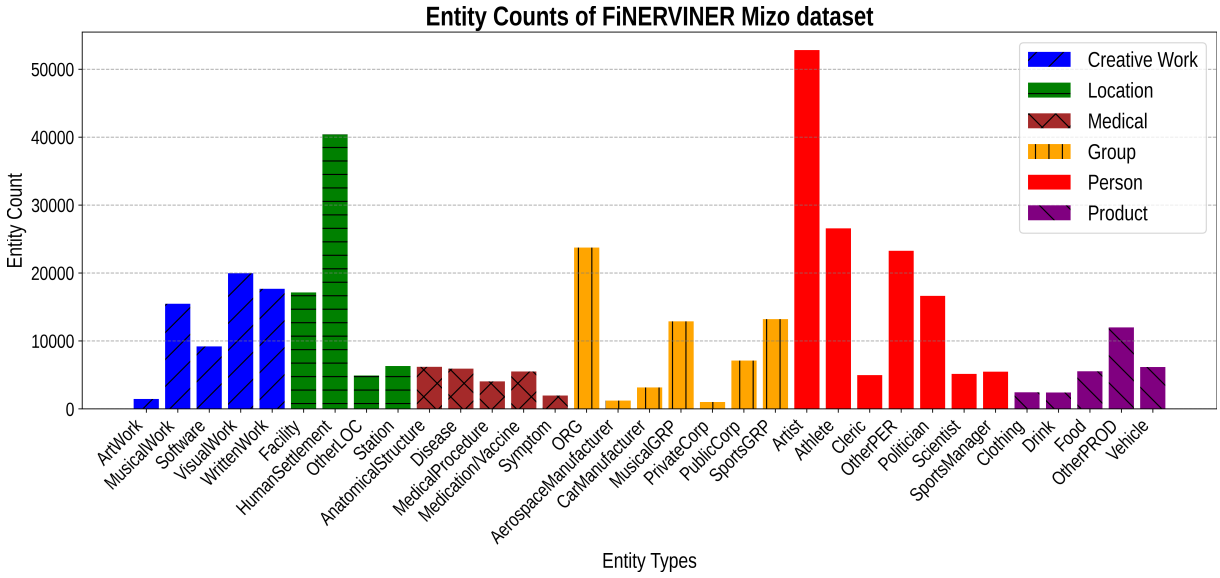

Figure 2: Entity type frequency distribution of the FiNERVINER Mizo dataset.

for each language. The good quality of these gold datasets can be ascertained based on the Cohen's kappa coefficient (κ) (Deleger et al., 2012), which is above 0.81 for each language.

5.2. Entity type frequency distribution

As shown in Figure 2, more number of entity mentions are detected for the fine types of Person (e.g. Artist) and Location (e.g. HumanSettlement). Since Artist type includes the mentions of musicians, actors, directors, authors, etc, the count is the highest among all the fine entity types. In contrast, very specific fine types such as AerospaceManufacturer, Drink, AnatomicalStructure, etc., are very scarce. Similar trends can be observed across all three languages.

6. Experimental Setup

The state-of-the-art approach for sequence labeling tasks involves fine-tuning pre-trained language models (PLM) with the NER datasets (Murthy et al., 2022; Marreddy et al., 2022; Malmasi et al., 2022; Mhaske et al., 2023; Fetahu et al., 2023; Tulajiang et al., 2025; del Moral-González et al., 2025; Maurya et al., 2026). Similarly, we have fine-tuned mBERT (bert-base-multilingual-cased) (Devlin et al., 2019), IndicBERTv2 (IndicBERTv2-MLM-Sam-TLM) (Doddapaneni et al., 2023), MuRIL (muril-large-cased) (Khanuja et al., 2021), XLM-RoBERTa (XLM-RoBERTa-large) (Conneau et al., 2020), and MizBERT (Lalramhluna et al., 2024) for fine-grained NER using the Hugging Face Transformers library (Wolf et al., 2020). The models were trained for six epochs with a batch size of 64, utilizing AdamW optimization (learning rate: 5e-5, weight decay: 0.01). Training was performed on an NVIDIA A100 GPU, with evaluation based on SeqEval metrics, and the best performance determined by the F1-score .

7. Results & Analysis

The following subsections cover the analyses of PLMs' performances fine-tuned on the FiNERVINER dataset, cross-lingual zero-shot evaluation, the impact of multilingualism and script similarity, and entity-specific error analyses.

7.1. Performance of PLMs fine-tuned on FiNERVINER dataset

Multilingual encoder model IndicBERTv2 is the only PLM pre-trained on Bodo (brx) and Manipuri (mni). Table 3 shows the performance of IndicBERTv2

Table 3: Performance of different models fine-tuned on FiNERVINER dataset. The best model performance values (in terms of F1 scores) for every language are in **bold**, and the second-best values are underlined.

Language	Model	Micro			Macro		
		P	R	F1	P	R	F1
Bodo (brx)	mBERT	59.04	61.20	60.17	59.33	62.11	60.61
	XLM-RoBERTa	61.02	62.17	61.82	60.04	61.73	60.89
	MuRIL	60.96	63.31	61.98	60.22	63.08	61.46
	IndicBERTv2	63.85	67.57	**65.66**	62.37	65.57	**64.02**
	MizBERT	41.31	38.96	40.10	39.81	37.61	38.68
Manipuri (mni)	mBERT	56.63	57.14	56.89	56.40	56.77	56.59
	XLM-RoBERTa	57.41	58.72	58.06	56.93	57.77	57.34
	MuRIL	58.32	60.04	59.17	57.16	59.28	58.11
	IndicBERTv2	60.49	64.42	**62.39**	61.39	63.34	**62.02**
	MizBERT	39.37	37.07	38.19	38.76	35.97	37.32
Mizo (lus)	mBERT	79.73	80.99	80.36	79.53	79.43	79.25
	XLM-RoBERTa	80.33	81.83	**81.07**	81.30	80.92	**80.89**
	MuRIL	78.51	81.36	79.91	77.19	77.60	77.02
	IndicBERTv2	76.66	78.62	77.63	77.84	76.81	76.61
	MizBERT	79.51	81.11	80.30	78.29	80.28	79.04

fine-tuned on FiNERVINER is superior. Similarly, the MizBERT's performance is superior when fine-tuned on Mizo (lus) as it is pre-trained on the same language. Although the multilingual PLMs mBERT, XLM-RoBERTa and MuRIL were not pre-trained on any of these three language, their performance is good after fine-tuned with the high-quality FiNERVINER dataset. In fact, an interesting result is observed in the case of the Mizo language. The Mizo language is written using the Latin script (similar to English, German, Spanish, French, Italian, Portuguese, etc.), and Bodo and Manipuri are written using Devanagari and Meitei scripts, respectively. Although the Mizo language was unseen during pre-training of mBERT, XLM-RoBERTa, MuRIL, and IndicBERTv2, the models could perform significantly better than other languages in the FgNER task due to the massive pre-training on the languages written using the Latin script. Also, since the MizBERT utilized BERT-base tokenizer on English and then pre-trained on only the Mizo language, its performance on Mizo after fine-tuning is superior but quite inferior after fine-tuning on the languages written using other scripts than the Latin script (Table 3). For further analysis on these observations, we have conducted cross-lingual zero-shot analysis, which is described in the following sections.

7.2. Cross-lingual zero-shot analysis

We have performed cross-lingual zero-shot analysis for every single language pair. As shown in Figure 3, the models are fine-tuned on datasets of respective languages and tested on the test set of other languages. As expected, the best performance of a model is found when it is fine-tuned on a particular language and tested on the same language. A model's performance is quite inferior on an unseen language. For example, when mBERT, XLM-RoBERTa, and MuRIL are fine-tuned only on Bodo (brx) or Mizo (lus), their zero-shot performance is very poor on the Manipuri (mni) test set. Similarly, when these models are fine-tuned on Manipuri (mni), their zero-shot performance on Bodo (brx) is quite inferior. But, interestingly, due to the pre-training on languages written using the Latin script, the zero-shot performance on Mizo (lus) is comparatively better. Moreover, if the PLM is pre-trained on a language, then its zero-shot performance improves. This is observed in the case of IndicBERTv2. Since IndicBERTv2 is pre-trained on 26 languages, including Manipuri, Bodo, and English, its zero-shot performance is better than the other three multilingual PLMs.

We have extended the zero-shot analysis to Bengali (bn), English (en), and Hindi (hi) languages, utilizing the publicly available MultiCoNER2 dataset (Fetahu et al., 2023). Since all these multilingual PLMs are pre-trained on these three high-resource languages, their zero-shot performances are superior compared to the zero-shot performances of the low-resource vulnerable languages.

7.3. The impact of Script Similarity

We further investigate the impact of script similarity for the FgNER task. As already discussed previously, the influence of the Latin script is imminent on the Mizo language. Similar observations are shown in Fig. 3 in the case of English (en) and Hindi (hi). As Bodo (brx) and Hindi (hi) are written using the Devanagari script, the zero-shot performance of Hindi improves on models fine-tuned on Bodo due

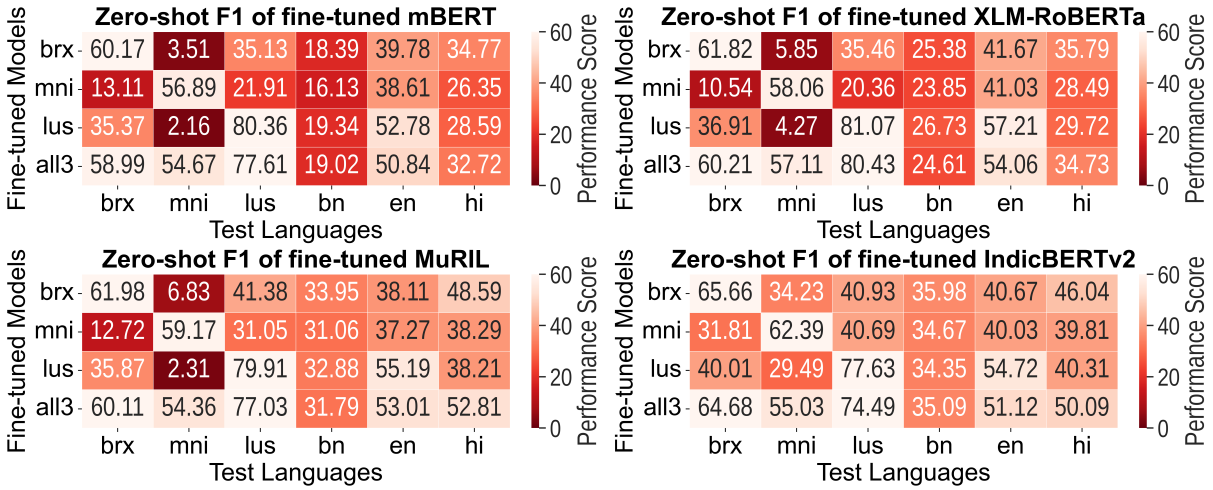

Figure 3: Zero-shot performance (micro F1) of mBERT, XLM-RoBERTa, MuRIL, and IndicBERT models fine-tuned on FiNERVINER dataset and tested on different languages.

to the script similarity. Similarly, we observe that the zero-shot performance of English is superior in the case of Mizo compared to other fine-tuned models due to the Latin script similarity. Named entities are nouns that mostly do not change spellings based on the language, if the languages are written using the same script. Therefore, in script-similar languages, in spite of grammatical and linguistic divergences, the detection of the entities is facilitated by the unchanged spelling. This is the reason why the zero-shot performance on languages written with a similar script improves. This is further established by the zero-shot performance of Bengali (bn), which is nearly consistent in any individual PLMs across all the fine-tuned variations. The effect of fine-tuned models is negligible as Bengali is written using a completely different script (Bengali-Assamese script), and its zero-shot performances are completely governed by the inclusion of Bengali during the pre-training of such PLMs.

7.4. Multilingualism

Our analysis extends to evaluating multilingualism. We constructed a balanced set (*all3*) comprising all three languages Bodo, Manipuri, and Mizo, ensuring an equal number of samples per language. As seen in the Figure 3, there is significant improvement over performances in every encoder model when fine-tuned with all the languages and tested on individual languages. Since the *all3* train set has samples of Bodo and Mizo, the zero-shot performance of Hindi and English improved due to the script similarity as discussed in the previous section. These results suggest the necessity of language-specific pre-training and task-specific fine-tuning.

Table 4: Entity errors in terms of the percentage of predicted entities for different languages fine-tuned on IndicBERTv2

Entity Error Type	brx	mni	lus
Boundary mismatch	9.27	11.46	4.43
Type mismatch	14.41	15.23	11.75
Spurious entity	2.67	2.21	0.84

7.5. Entity type specific performance

Table 5 shows the entity type-specific F1 scores of the best fine-tuned model for each language. Performance generally improved for entity types with more training samples (e.g. *HumanSettlement*). Interestingly, the low performance for the well-sampled *OtherPER* entity type is likely due to confusion with finer-grained entity types of *Person*, such as *Artist*, *Athlete*, and *Politician*, as detailed in the 7.6 Error Analysis section.

7.6. Error Analysis

Fine-grained named entity recognition is crucial, as entity types may vary by context. Therefore, we have analyzed the errors in two different approaches. Table 4 shows the details of entity errors in terms of the percentage of predicted entities. The common errors that occur include the boundary error (such as “Who” is marked as *MusicalGRP* instead of “The Who”), entity type mismatch error (e.g. “Nissan Cherry” is categorized as *Food* instead of *Vehicle*) and spurious errors (such a “blue” is marked as an entity whereas the entity type *color* is not defined in MultiCoNER2 taxonomy). Entity boundary mismatch errors and entity type mismatch errors are highest in Manipuri (mni), whereas spurious entity errors occur the most in Bodo (brx).

Table 5: Entity-type specific performance of the best models fine-tuned on the FiNERVINER dataset.

	Bodo (brx)			Manipuri (mni)			Mizo (lus)		
Entity Type	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
AerospaceManufacturer	69.61	51.74	58.23	69.23	56.25	62.07	94.12	84.21	88.89
AnatomicalStructure	80.77	75.01	77.78	75.00	78.95	76.92	95.83	98.12	97.23
ArtWork	60.71	36.42	44.74	65.02	35.14	45.61	90.91	66.67	76.93
Artist	72.40	80.51	76.24	71.90	80.32	75.88	82.89	85.16	84.01
Athlete	74.60	77.99	76.26	75.18	78.28	76.70	76.54	81.58	78.98
CarManufacturer	56.72	67.86	61.79	44.23	67.65	53.49	88.89	94.12	92.39
Cleric	65.52	51.82	57.87	52.94	43.90	48.02	61.54	53.33	57.14
Clothing	61.36	61.36	61.36	59.46	52.38	55.70	98.89	94.44	97.14
Disease	61.26	60.18	60.71	66.67	64.10	65.36	83.33	75.00	78.95
Drink	63.04	53.70	58.02	65.51	50.02	56.76	98.00	98.00	98.00
Facility	69.96	73.24	71.56	70.52	75.31	72.84	87.04	85.45	86.24
Food	67.74	61.32	64.37	52.54	57.41	54.87	84.62	95.65	89.80
HumanSettlement	86.25	86.91	86.58	83.77	85.25	84.25	91.20	95.80	93.44
MedicalProcedure	68.37	69.79	69.07	64.44	70.73	67.44	95.00	98.04	97.34
Medication/Vaccine	76.82	84.67	80.56	65.22	75.00	69.77	94.35	95.90	95.12
MusicalGRP	61.45	70.34	65.60	63.64	71.80	67.47	82.50	89.19	85.71
MusicalWork	73.11	70.78	71.93	69.18	69.59	69.36	79.17	77.55	78.35
ORG	63.73	66.39	65.03	64.85	64.58	64.35	73.86	87.50	80.03
OtherLOC	55.00	48.53	51.56	46.90	50.75	48.75	58.51	60.44	59.46
OtherPER	46.83	32.74	38.54	36.00	30.08	32.83	42.85	35.29	38.71
OtherPROD	49.62	48.53	49.07	51.14	45.03	47.97	93.18	85.17	89.13
Politician	52.61	50.68	51.63	60.13	53.80	56.79	63.16	69.23	66.06
PrivateCorp	46.66	50.72	48.53	47.19	48.80	47.98	84.62	73.33	78.57
PublicCorp	45.86	52.75	49.48	49.27	53.97	51.55	80.77	72.41	76.36
Scientist	49.10	53.37	51.14	46.90	50.75	48.75	58.51	60.44	59.46
Software	61.86	67.42	64.52	58.82	66.67	62.50	83.33	96.15	89.29
SportsGRP	86.49	86.49	86.49	88.28	86.92	87.60	90.00	90.00	90.00
SportsManager	75.44	58.90	66.15	77.27	62.97	69.39	84.62	68.75	75.86
Station	79.76	79.76	79.76	72.31	82.46	77.05	94.74	90.00	92.31
Symptom	58.33	57.14	57.73	71.05	60.04	65.26	93.75	93.75	93.75
Vehicle	54.42	54.42	54.42	45.61	60.47	52.05	81.82	85.71	83.72
VisualWork	70.08	72.25	71.15	68.57	67.80	68.19	78.33	78.33	78.33
WrittenWork	77.29	75.32	76.29	75.72	80.87	78.22	83.93	88.68	86.24

We have further analyzed the often co-predicted fine-grained types. From Table 4, we have selected Manipuri (mni) for this analysis as this language has the highest percentage of mismatch entity types. As shown in Figure 4, *Symptom* is sometimes confused with *Disease*. Similarly, the fine types *Artist*, *Athlete*, *Politician* are sometimes confused with *OtherPER*, due to which the F1-score of *OtherPER* is the lowest among all the fine entity types, as also discussed in the 7.5 Entity type specific performance section. Apart from such closely related fine entity types, most of the other fine entity types are learned by the models without much confusion.

8. Conclusion

We have generated FiNERVINER dataset by projecting English MultiCoNER2 annotations to three low-resource target languages, utilizing the parallel corpora and a multilingual encoder-based annotation projection and word alignment tool. The dataset comprises over 198k sentences, 282k entities, and 2.8M tokens in each vulnerable language: Bodo, Manipuri, and Mizo. As the first FgNER dataset for these languages created via an annotation projection method, extensive experiments validate its quality, and cross-lingual zero-shot analyses underscore the need for language-specific pre-training and task-specific fine-tuning.

9. Limitations

Despite the encouraging results of this study, it is important to acknowledge certain limitations that require further investigation. It is essential to recognize that the characteristics of the generated dataset are directly dependent on the characteristics of the source FgNER dataset. Any inherent biases or specificities within the source dataset have the potential to influence the generated dataset.

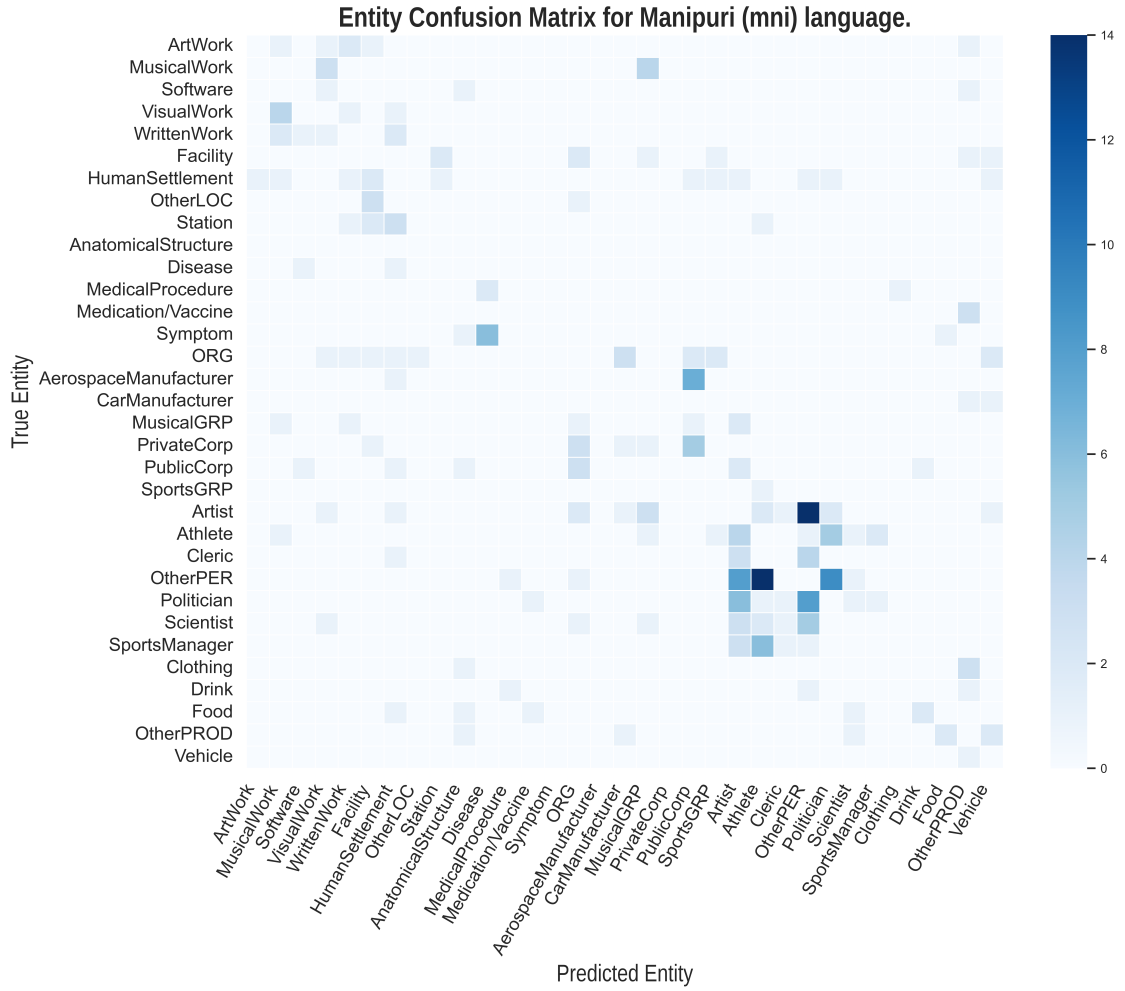

Figure 4: Entity type confusion matrix of Manipuri (mni) language

Second, the volume and quality of the generated dataset are linked to the availability of parallel corpora and the specific selection of the annotation projection tool and multilingual encoder models. Therefore, further investigation of the impact of different combinations of these tools and models remains a crucial direction for future research. Finally, a comprehensive analysis of the performance of Large Language Models (LLMs) on the FgNER task, both in a zero-shot setting and after fine-tuning with the FiNERVINER dataset, represents a crucial avenue for future work.

10. Ethics Statement

The annotations were generated using the openly accessible MultiCoNER2[1] dataset and BPCC[2] parallel corpora released under CC-BY-4.0[3] and CC0[4] licenses. In addition to collecting data from multiple domains, BPCC emphasizes geographically and culturally relevant information about India sourced from official Government of India websites. We did not modify these datasets to correct for potential biases and use them as-is. We have cited all the sources of resources, tools, packages, and models used in this work. The test-set annotations were provided pro bono by volunteers passionate about creating a fine-grained named entity recognition dataset for Indian languages. The annotators were clearly introduced to the task and assisted appropriately during the annotation process. These contributors received no financial compensation and were informed in advance that their annotations would be released publicly. Importantly, none of the submitted annotations include any personal or identifying information.

[1] https://multiconer.github.io/

[2] https://huggingface.co/datasets/ai4bharat/BPCC

[3] https://creativecommons.org/licenses/by/4.0/

[4] https://creativecommons.org/public-domain/cc0/

The FiNERVINER dataset, expert detector models, the agentic tool, and the interactive web application are available as open-source resources at: https://hf.co/collections/prachuryyaIITG/finerviner[5] under MIT license[6].

11. Bibliographical References

Mahathi Bhagavatula, GSK Santosh, and Vasudeva Varma. 2012. Language independent named entity identification using wikipedia. In *Proceedings of the First Workshop on Multilingual Modeling*, pages 11–17.

Statement-1 Census of India. 2011. Statement 1: Abstract of Language Data.

Nancy Chinchor, Patricia Robinson, and Elizabeth Brown. 1998. Hub-4 IE-NE Task Definition Version 4.8.

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 *ACL*.

Rodrigo del Moral-González, Helena Gómez-Adorno, and Orlando Ramos-Flores. 2025. Comparative analysis of generative llms for labeling entities in clinical notes. *Genomics & Informatics*, 23(1):1–8.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, Imre Solti, et al. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. SemEval-2023 task 2: Fine-grained multilingual named entity recognition (MultiCoNER 2). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2247–2265, Toronto, Canada. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Karthik Gali, Harshit Surana, Ashwini Vaidya, Praneeth M Shishtla, and Dipti Misra Sharma. 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*.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. Model and data transfer for cross-lingual sequence labelling in zero-resource settings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Shalini Gupta and Pushpak Bhattacharyya. 2010. Think globally, apply locally: using distributional characteristics for hindi named entity identification. In *Proceedings of the 2010 Named Entities Workshop*, pages 116–125.

Li Huang, Haowen Liu, Qiang Gao, Jiajing Yu, Guisong Liu, and Xueqin Chen. 2025. Adversity-

[5] https://hf.co/collections/prachuryyaIITG/finerviner

[6] https://opensource.org/license/MIT

aware few-shot named entity recognition via augmentation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24132–24140.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

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*.

Robert Lalramhluna, Sandeep Dash, and Dr Partha Pakray. 2024. Mizbert: a mizo bert model. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(7):1–14.

Mounika Marreddy, Subba Reddy Oota, Lakshmi Sireesha Vakada, Venkata Charan Chinni, and Radhika Mamidi. 2022. Am i a resource-poor language? data sets, embeddings, models and analysis for four different nlp tasks in telugu language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1):1–34.

Ajanta Maurya, V. Vijaya Saradhi, and Ashish Anand. 2026. Halo-gpt:hindi active learning with oracle gpt-3.5. In *Proceedings of the 17th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '25, page 113–123, New York, NY, USA. Association for Computing Machinery.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics-Association for Computational Linguistics (Print)*, 29(1):19–51.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245.

Sujan Kumar Saha, Pabitra Mitra, and Sudeshna Sarkar. 2008a. Word clustering and word selection based feature reduction for maxent based hindi ner. In *proceedings of ACL-08: HLT*, pages 488–495.

Sujan Kumar Saha, Sudeshna Sarkar, and Pabitra Mitra. 2008b. A hybrid feature set based maximum entropy hindi named entity recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. 2006. Harem: An advanced ner evaluation contest for portuguese. In *quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006)*.

SEKINE Satoshi. 2000. Irex: Ir and ie evaluation-based project in japanese. In *Proceedings of the Language Resource and Evaluation Conference, 2000*.

GOI The Constitution of India. 1950. *The Constitution of India*. Government of India.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Paerhati Tulajiang, Yuanyuan Sun, Yuanyu Zhang, Yingying Le, Kelaiti Xiao, and Hongfei Lin. 2025. A bilingual legal ner dataset and semantics-aware cross-lingual label transfer method for low-resource languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Atlas UNESCO. 2017. Unesco atlas of the world's languages in danger.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844.

Xiaojun Xue, Chunxia Zhang, Tianxiang Xu, and Zhendong Niu. 2024. Robust few-shot named entity recognition with boundary discrimination and correlation purification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19341–19349.

Yibo Yan, Peng Zhu, Dawei Cheng, Fangzhou Yang, and Yifeng Luo. 2023. Adversarial multi-task learning for efficient chinese named entity recognition. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(7):1–19.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Yu Zhang, Yunyi Zhang, Yanzhen Shen, Yu Deng, Lucian Popa, Larisa Shwartz, ChengXiang Zhai, and Jiawei Han. 2024. Seed-guided fine-grained entity typing in science and engineering domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19606–19614.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of ACL 2019*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition.

12. Language Resource References

Abhishek Abhishek, Sanya Bathla Taneja, Garima Malik, Ashish Anand, and Amit Awekar. 2019. Fine-grained entity recognition with reduced false negatives and large type coverage. In *AKBC*.

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-ner: Massive multilingual named entity recognition. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 586–594. SIAM.

Joseph Z Chang, Richard Tzong-Han Tsai, and Jason S Chang. 2009. Wikisense: Supersense tagging of wikipedia named entities based wordnet. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, pages 72–81.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. *arXiv preprint arXiv:1807.04905*.

Luciano Del Corro, Abdalghani Abujabal, Rainer Gemulla, and Gerhard Weikum. 2015. Finet: Context-aware fine-grained named entity typing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 868–878.

Sobha Lalitha Devi, Pattabhi RK Rao, CS Malarkodi, and R Vijay Sundar Ram. 2014. Indian language ner annotated fire 2014 corpus (fire 2014 ner corpus). *Named-Entity Recognition Indian Languages FIRE*.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-nerd: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213.

Asif Ekbal, Rejwanul Haque, and Sivaji Bandyopadhyay. 2008. Named entity recognition in bengali: A conditional random field approach. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. 2023. Multiconer v2: a large multilingual dataset for fine-grained and noisy named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2027–2051.

Iker García-Ferrero, Jon Ander Campos, Oscar Sainz, Ander Salaberria, and Dan Roth. 2023. Ixa/cogcomp at semeval-2023 task 2: Context-enriched multilingual named entity recognition using knowledge bases. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Dan Gillick, Nevena Lazic, Kuzman Ganchev, Jesse Kirchner, and David Huynh. 2014. Context-dependent fine-grained entity type tagging. *arXiv preprint arXiv:1412.1820*.

Thangkhanhau Haulai and Jamal Hussain. 2023. Construction of mizo: English parallel corpus for machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(8):1–12.

Saiful Islam. 2018. English to bodo statistical machine translation system using multi-domain parallel corpora. In *Proceedings of the 15th International Conference on Natural Language Processing*, pages 75–81.

Saiful Islam, Abhijit Paul, Bipul Shyam Purkayastha, and Ismail Hussain. 2018. Construction of english-bodo parallel text corpus for statistical machine translation. *International Journal on Natural Language Computing (IJNLC) Vol*, 7.

Laishram Jimmy, Kishorjit Nongmeikappam, and Sudip Kumar Naskar. 2023. Bilstm-crf manipuri ner with character-level word representation. *Arabian journal for science and engineering*, 48(2):1715–1734.

Prachuryya Kaushik and Ashish Anand. 2025. CLASSER: Cross-lingual annotation projection enhancement through script similarity for fine-grained named entity recognition. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. Main conference paper.

Prachuryya Kaushik and Ashish Anand. 2026a. AWED-FiNER: Agents, web applications, and expert detectors for fine-grained named entity recognition across 36 languages for 6.6 billion speakers.

Prachuryya Kaushik and Ashish Anand. 2026b. SampurNER: Fine-grained named entity recognition dataset for 22 indian languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40.

Prachuryya Kaushik, Shivansh Mishra, and Ashish Anand. 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*, pages 3753–3763.

Jimmy Laishram, Kishorjit Nongmeikapam, and Sudip Kumar Naskar. 2020. Deep neural model for manipuri multiword named entity recognition with unsupervised cluster feature. In *Proceedings of the 17th international conference on natural language processing (ICON)*, pages 420–429.

Xiao Ling and Daniel S Weld. 2012. Fine-grained entity recognition. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. 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*, pages 29–34.

Jun-Yu Ma, Jia-Chen Gu, Jiajun Qi, Zhenhua Ling, Quan Liu, and Xiaoyi Zhao. 2023a. Ustc-nelslip at semeval-2023 task 2: Statistical construction and dual adaptation of gazetteer for multilingual complex ner. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Long Ma, Kai Lu, Tianbo Che, Hailong Huang, Weiguo Gao, and Xuan Li. 2023b. Pai at semeval-2023 task 2: A universal system for named entity recognition with external entity information. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 744–750.

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

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. Naamapadam: A large-scale named entity annotated data for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456, Toronto, Canada. Association for Computational Linguistics.

Rajesh Mundotiya, Shantanu Kumar, Ajeet Kumar, Umesh Chaudhary, Supriya Chauhan, Swasti Mishra, Praveen Gatla, and Anil Kumar Singh. 2023. Development of a dataset and a deep learning baseline named entity recognizer for three low resource languages: Bhojpuri, maithili, and magahi. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(1):1–20.

Rudra Murthy, Pallab Bhattacharjee, Rahul Sharnagat, Jyotsana Khatri, Diptesh Kanojia, and Pushpak Bhattacharyya. 2022. Hiner: A large hindi named entity recognition dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4467–4476.

Shikhar Murty, Patrick Verga, Luke Vilnis, and Andrew McCallum. 2017. Finer grained entity typing with typenet. *arXiv preprint arXiv:1711.05795*.

Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2023. Transfer learning for low-resource multilingual relation classification. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–24.

Sanjib Narzary, Anjali Brahma, Sukumar Nandi, and Bidisha Som. 2024. Deep learning based named entity recognition for the bodo language. *Procedia Computer Science*, 235:2405–2421.

Nobal Niraula and Jeevan Chapagain. 2022. Named entity recognition for nepali: data sets and algorithms. In *The International FLAIRS Conference Proceedings*, volume 35.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. 2022. Asner-annotated dataset and baseline for assamese named entity recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6571–6577.

Akshara Prabhakar, Gouri Sankar Majumder, and Ashish Anand. 2022. Cl-neril: A cross-lingual model for ner in indian languages (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 13031–13032.

Pooja Rai and Sanjay Chatterji. 2022. Annotation projection-based dependency parser development for nepali. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–19.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

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

Aniketh Janardhan Reddy, Monica Adusumilli, Sai Kiranmai Gorla, Lalita Bhanu Murthy Neti, and Aruna Malapati. 2018. Named entity recognition for telugu using lstm-crf. In *WILDRE4–4th Workshop on Indian Language Data: Resources and Evaluation*, volume 6.

Satoshi Sekine and Chikashi Nobata. 2004. Definition, dictionaries and tagger for extended named entity hierarchy. In *LREC*, pages 1977–1980. Lisbon, Portugal.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended named entity hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Anil Kumar Singh. 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*.

Oyesh Mann Singh, Ankur Padia, and Anupam Joshi. 2019. Named entity recognition for nepali language. In *2019 IEEE 5th international conference on collaboration and internet computing (cic)*, pages 184–190. IEEE.

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, and Fei Huang. 2023. Damo-nlp at semeval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Ralph Weischedel and Ada Brunstein. 2005. Bbn pronoun coreference and entity type corpus. *Linguistic Data Consortium, Philadelphia*, 112.

Mohamed Amir Yosef, Sandro Bauer, Johannes Hoffart, Marc Spaniol, and Gerhard Weikum. 2012. Hyena: Hierarchical type classification for entity names. In *Proceedings of COLING 2012: Posters*, pages 1361–1370.