CLASSER: Cross-lingual Annotation Projection enhancement through Script Similarity for Fine-grained Named Entity Recognition

Prachuryya Kaushik and **Ashish Anand**
Department of Computer Science and Engineering
Indian Institute of Technology Guwahati
Guwahati, Assam, India
{k.prachuryya, anand.ashish}@iitg.ac.in

Abstract

We introduce CLASSER, a cross-lingual annotation projection framework enhanced through script similarity, to create fine-grained named entity recognition (FgNER) datasets for low-resource languages. Manual annotation for named entity recognition (NER) is expensive, and distant supervision often produces noisy data that are often limited to high-resource languages. CLASSER employs a two-stage process: first projection of annotations from high-resource NER datasets to target language by using source-to-target parallel corpora and a projection tool built on a multilingual encoder, then refining them by leveraging datasets in script-similar languages. We apply this to five low-resource Indian languages: *Assamese*, *Marathi*, *Nepali*, *Sanskrit*, and *Bodo*, a vulnerable language. The resulting dataset comprises 1.8M sentences, 2.6M entity mentions and 24.7M tokens. Through rigorous analyses, the effectiveness of our method and the high quality of the resulting dataset are ascertained with F1 score improvements of 26% in Marathi and 46% in Sanskrit over the current state-of-the-art. We further extend our analyses to zero-shot and cross-lingual settings, systematically investigating the impact of script similarity and multilingualism on cross-lingual FgNER performance. The dataset is publicly available at huggingface.co/datasets/prachuryyaIITG/CLASSER.

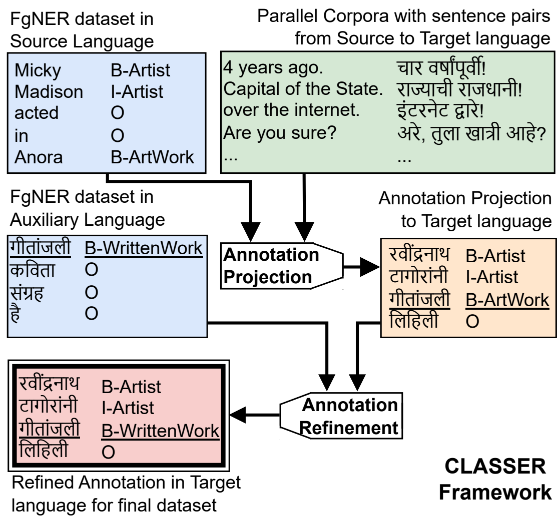

Figure 1: Illustration of CLASSER Framework

1 Introduction

Structured knowledge extraction from unstructured text underpins countless downstream applications, such as recommendation systems, knowledge-base construction, relation extraction, and beyond. Named Entity Recognition (NER), which identifies and classifies mentions of persons, locations, organizations, etc. has evolved from early rule-based systems (Rau, 1991) through the collective contributions in the 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 powerful neural architectures today (Zhang et al., 2019; Zhou et al., 2023). Yet, conventional coarse-grained categories in NER often fall short when applications demand more specific distinctions, e.g., "Scientist" from generic "Person" or "Clothing" from generic "Product" (Choi et al., 2018). The type and granularity of fine-grained entities differ depending on the domain and application requirements. Early efforts in fine-grained named entity recognition (FgNER) contributed hierarchical type systems (Sekine and Nobata, 2004), distant-supervision pipelines (Ling and Weld, 2012; Yosef et al., 2012), contextual embedding techniques (Gillick et al., 2014), and noise-aware neural architectures capable of predicting hundreds of labels (Murty et al., 2017). Although noise is reduced in FgNER resources by applying language-specific heuristics (Abhishek et al., 2019), expensive manual annotation yields higher reliability and improved annotation quality (Ding et al., 2021).

While coarse-grained NER for Indian languages

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, pages 1745–1760
December 20-24, 2025 ©2025 Association for Computational Linguistics

has seen considerable progress, fine-grained NER (FgNER) only began to emerge recently. The MultiCoNER2[1] shared task at SemEval-2023 introduced FgNER datasets for Hindi and Bengali via translated English annotations (Fetahu et al., 2023a), and the TAFSIL[2] initiative mined Wikidata and Wikipedia links to generate noisy FgNER data for six additional Indian languages (Kaushik et al., 2025). Despite these advances, comprehensive and high-quality fine-grained resources remain scarce for most low-resource Indian languages.

To address this gap, we present CLASSER: Cross-lingual Annotation Projection framework enhanced through Script Similarity for Fine-grained Named Entity Recognition. As illustrated in Figure 1, in Stage 1, we project English MultiCoNER2 annotations to the target language using BPCC parallel corpora (Gala et al., 2023) and a multilingual encoder-based annotation projection and word alignment tool (Dou and Neubig, 2021; García-Ferrero et al., 2022). In stage 2, we introduce a confidence-score-based cross-lingual refinement method that utilizes an auxiliary NER model fine-tuned on script-similar languages. For example, although Hindi (Indo-European language) and Bodo (Sino-Tibetan language) are typologically distinct, their shared Devanagari script allows us to significantly enhance the Bodo FgNER annotations using the available MultiCoNER2 FgNER dataset in Hindi. Figure 1 shows an example of annotation refinement from projected entity type *B-ArtWork* to *B-WrittenWork* based on the FgNER dataset in auxiliary language. We apply the CLASSER framework to generate FgNER dataset in five low-resource languages, including *Assamese (as)*, *Marathi (mr)*, *Nepali (ne)*, *Sanskrit (sa)*, along with a vulnerable language *Bodo (brx)* (UNESCO, 2017).

Our contributions can be summarized as follows:

1. Development of CLASSER framework for cross-lingual annotation projection with script-similarity-based refinement to create high-quality FgNER datasets.

2. Construction of a large-scale FgNER dataset comprising of 1.8M sentences, 2.6M entity mentions, and 24.7M tokens for five low-resource Indian languages: Assamese (as), Bodo (brx), Marathi (mr), Nepali (ne), and Sanskrit (sa).

3. Creation of a high-quality human-annotated test set consisting of 1000 sentences for each language with inter-annotator agreement (κ) above 0.86.

4. Our rigorous analyses establish the effectiveness of the proposed method and the good quality of the generated dataset, which has achieved 26% and 46% improvement in F1 scores over equal-sized TAFSIL (Kaushik et al., 2025) datasets in Marathi and Sanskrit, respectively.

5. 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) 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), BBN (93 types) (Weischedel and Brunstein, 2005), HYENA (505 types) (Yosef et al., 2012), FIGER (113 types) (Ling and Weld, 2012), and OntoNotes (88 types) (Gillick et al., 2014). 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) 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 dataset covering 66 fine-grained types.

Early Indian-language NER began with IJCNLP-2008 for Hindi, Bengali, Oriya, Telugu, and Urdu (Singh, 2008), and further enhanced by Gali et al. (2008), Saha et al. (2008), Gupta and Bhattacharyya (2010), Ekbal and Saha (2011), Bhagavatula et al. (2012), 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. Manually annotated corpora include NER in Bengali (Ekbal et al., 2008), Telugu (Reddy et al., 2018), Maithili (Priyadarshi and Saha, 2021), Hindi (Venkataramana et al., 2022), Assamese (Pathak et al., 2022), Marathi (Litake et al., 2022), Nepali (Niraula and Chapagain, 2022), Bishnupriya Manipuri (Jimmy et al., 2023), Bodo (Narzary et al., 2024) etc.

Yarowsky and Ngai (2001) pioneered projection via parallel corpora, but zero-shot transfer struggles with typological distance (Karthikeyan et al., 2020; Wu and Dredze, 2019), and cross-lingual encoders (Pires et al., 2019; Conneau et al., 2020) have not

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

[2]https://huggingface.co/datasets/prachuryyaIITG/TAFSIL

closed the gap yet (Ruder et al., 2021). Mhaske et al. (2023) used Samanantar corpora (Ramesh et al., 2022) and word alignment (Ruder et al., 2021; OCH and NEY, 2003) to project English NER to eleven Indian languages. More broadly, projection-based translation was used to generate NER resources across various languages (Mayhew et al., 2017; Ugawa et al., 2018; Jain et al., 2019; Yang et al., 2022; Liu et al., 2021; Lancheros et al., 2024), and utilization of script similarity boosted low-resource translation, Parts-of-Speech tagging, and NER (Aepli and Sennrich, 2022; Patil et al., 2022; Blaschke et al., 2023; Brahma et al., 2023).

MultiCoNER-1 (Malmasi et al., 2022) and MultiCoNER2 (Fetahu et al., 2023b) produced Hindi and Bengali datasets via translated English annotations, which were further improved by SemEval-2023 shared-task participating teams (Ma et al., 2023a,b; Tan et al., 2023; García-Ferrero et al., 2023). Recently, Kaushik et al. (2025) applied distant supervision to create noisy FgNER datasets for six Indian languages across four taxonomies. Despite these advances, high-quality multilingual FgNER resources for most Indian languages remain scarce.

3 Methodology

3.1 CLASSER Framework

The proposed framework CLASSER adopts a two-stage approach that incorporates three distinct categories of languages (Figure 1). The high-resource source language (src) provides both an annotated FgNER dataset and parallel corpora comprising sentence pairs aligned with the target language (tgt). The source language, however, differs notably from the target language in terms of grammatical structure and script. In contrast, an auxiliary language (aux) is characterized by the availability of an annotated FgNER dataset written in a script similar to that of the target language, but it lacks corresponding parallel sentence pairs with the target language. For a language l, the FgNER dataset $D_l = \{(s_l^{(k)}, y_l^{(k)})\}_{k=1}^{N}$ consists of N samples where each sentence s is paired with its corresponding annotation sequence y. For the source language src, we possess both FgNER dataset D_{src} and a parallel corpus having source-to-target language aligned-sentence pairs, $P = \{(s_{src}^{(k)}, s_{tgt}^{(k)})\}_{k=1}^{N}$. The auxiliary language provides only the FgNER dataset D_{aux}, sharing its script with tgt but offering no additional parallel data.

Algorithm 1 Algorithm for CLASSER Framework

Symbols:

- M: Pretrained multilingual encoder.
- N: Number of sentences per dataset.
- y: List containing FgNER annotations for each token of sentence s.
- $D_l = \{(s_l^{(k)}, y_l^{(k)})\}_{k=1}^{N}$: FgNER dataset of size N in language l.
- src, tgt, aux: Source, Target and Auxiliary languages respectively.
- $P = \{(s_{src}^{(k)}, s_{tgt}^{(k)})\}_{k=1}^{N}$: Parallel corpora.
- $\mathcal{A}(s_{src}, s_{tgt})$: Annotation alignment service produces a mapping map.
- $\hat{y}_{tgt}$: Initial annotation sequence for s_{tgt}.
- M_{aux}: Multilingual encoder fine-tuned on D_{aux}.
- $p_{aux}(j) = M_{aux}(s_{tgt})[j]$: The annotation probability distribution for jth token in s_{tgt}th sentence after refinement using M_{aux}
- $\ddot{y}_{tgt}$: Auxiliary annotation sequence s_{tgt}.
- $p_{aux}^{max}(j)$: Maximum probability of $\ddot{y}_{tgt}^{(j)}$ for jth token.
- τ: Confidence Threshold.
- Refinement function:

$$f\left(\hat{y}_{tgt}^{(j)}, \ddot{y}_{tgt}^{(j)}\right) = \begin{cases} \ddot{y}_{tgt}^{(j)}, & \text{if } p_{aux}^{max}(j) \geq \tau \\ & \text{and } \ddot{y}_{tgt}^{(j)} \neq \hat{y}_{tgt}^{(j)}, \\ \hat{y}_{tgt}^{(j)}, & \text{otherwise.} \end{cases}$$

- Final refined annotation: $\bar{y}_{tgt}^{(j)} = f(\hat{y}_{tgt}^{(j)}, \ddot{y}_{tgt}^{(j)})$
- $D_{tgt}^{ref} = \{(s_{tgt}^{(k)}, \bar{y}_{tgt}^{(k)})\}_{k=1}^{N}$: Final refined target FgNER dataset of size N.

Algorithm:

```
for each (s_src, s_tgt) ∈ P do
    Compute embeddings:
    E_src = M(s_src), E_tgt = M(s_tgt)
    Obtain alignment mapping:
    map = A(s_src, s_tgt) using E_src & E_tgt
    for each token j in s_tgt do
        Project annotation: ŷ_tgt^(j) = y_src^(map^{-1}(j))
    end for
    Obtain auxiliary predictions:
    ∀j in s_tgt: p_aux(j) = M_aux(s_tgt)[j]
    for each token j in s_tgt do
        Refine annotation: ȳ_tgt^(j) = f(ŷ_tgt^(j), ÿ_tgt^(j))
    end for
    Update (s_tgt, ȳ_tgt) to D_tgt^ref
end for
```

Stage 1 (Annotation Projection):

As per the Algorithm 1, we first fine-tune a pretrained multilingual encoder M on parallel corpora

P so that it learns to produce contextual embeddings $\mathbf{E}_{src} = M(s_{src})$ and $\mathbf{E}_{tgt} = M(s_{tgt})$ for source and target languages, respectively. For each parallel pair $(s_{src}, s_{tgt}) \in P$, a word alignment service $\mathcal{A}(s_{src}, s_{tgt})$ leverages $\mathbf{E}_{src}$ and $\mathbf{E}_{tgt}$ to yield a mapping map from source-sentence tokens to target-sentence tokens based on the FgNER dataset in the source language (D_{src}). We then transfer each source annotation sequence y_{src} to the target sentence by setting $\hat{y}_{tgt}^{(j)} = y_{src}^{(map^{-1}(j))}$ for every token j in the target sentence s_{tgt}. This produces an **initial annotation** sequence $\hat{y}_{tgt}$ for each target sentence, which may be noisy due to alignment errors, script mismatches, or syntactic divergences.

Stage 2 (Annotation Refinement):

To correct such projection noise, we exploit the observation that many named entities retain nearly identical orthographic forms when expressed in the same script, even across grammatically distinct languages. The semantics of entities are already captured by a pre-trained multilingual encoder during fine-tuning on a language written using the similar script. Although other factors, such as typological differences between languages play a role in various NLP tasks, we have found that leveraging shared script characteristics is highly effective for the FgNER task. We therefore introduce a refinement stage driven by an auxiliary language (aux) whose writing system matches that of the target language. A multilingual encoder model M_{aux} is fine-tuned on the auxiliary FgNER dataset D_{aux}. When applied to each target sentence s_{tgt}, for every token j, M_{aux} outputs a discrete probability distribution $p_{aux}(j) = M_{aux}(s_{tgt})[j]$ and an auxiliary annotation sequence $\ddot{y}_{tgt}$. The maximum probability of the auxiliary annotation $\ddot{y}_{tgt}^{(j)}$ for the jth token is denoted as $p_{aux}^{max}(j)$.

We then apply a refinement function:

$$f\left(\hat{y}_{tgt}^{(j)}, \ddot{y}_{tgt}^{(j)}\right) = \begin{cases} \ddot{y}_{tgt}^{(j)}, & \text{if } p_{aux}^{max}(j) \geq \tau \\ & \text{and } \ddot{y}_{tgt}^{(j)} \neq \hat{y}_{tgt}^{(j)}, \\ \hat{y}_{tgt}^{(j)}, & \text{otherwise.} \end{cases}$$

where τ is a confidence threshold. For every token j, the refinement function selectively replaces the initial projected label $\hat{y}_{tgt}^{(j)}$ with the auxiliary label $\ddot{y}_{tgt}^{(j)}$ only when the auxiliary annotation is both confident (i.e. $p_{aux}^{max}(j) \geq \tau$) and disagrees with the initially projected annotation (i.e. $\ddot{y}_{tgt}^{(j)} \neq \hat{y}_{tgt}^{(j)}$). The result is the **final refined annotation** $\bar{y}_{tgt}^{(j)}$ for each token j, i.e. $\bar{y}_{tgt}^{(j)} = f(\hat{y}_{tgt}^{(j)}, \ddot{y}_{tgt}^{(j)})$.

Finally, aggregation of all sentence–annotation pairs $\left(s_{tgt}^{(k)}, \bar{y}_{tgt}^{(k)}\right)$ produces the fully refined target-language dataset of size N:

$$D_{tgt}^{ref} = \left\{\left(s_{tgt}^{(k)}, \bar{y}_{tgt}^{(k)}\right)\right\}_{k=1}^{N}$$

3.2 Implementation of CLASSER framework

MultiCoNER2 (Fetahu et al., 2023a), a SemEval-2023 shared task (Fetahu et al., 2023b), provides 33 fine-grained entity types across 12 languages (including English, Hindi, Bengali). In our setup, the source language (src) is English; target languages (tgt) are Assamese (as), Bodo (brx), Marathi (mr), Nepali (ne), and Sanskrit (sa); auxiliary languages (aux) are Bengali (bn) and Hindi (hi). Assamese and Bengali use the Bengali-Assamese script, whereas Hindi, Bodo, Marathi, Nepali, and Sanskrit use Devanagari. We employ BPCC (Gala et al., 2023) as parallel corpora (P), augmenting the smaller Bodo data with Islam et al. (2018a,b). 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}). The Hindi and Bengali MultiCoNER2 datasets serve as D_{aux}, and we fine-tune IndicBERTv2 (Doddapaneni et al., 2022) as M and M_{aux}. IndicBERTv2 is the only encoder pre-trained on all src, tgt, and aux languages. Based on language-specific evaluations (Figure 4, 5), we set the confidence threshold τ=0.85. Following the Algorithm 1, the CLASSER framework is implemented with the mentioned details, and the dataset is created.

3.3 CLASSER Dataset

As shown in Table 1, the created CLASSER dataset consists of more than 157 thousand sentences, 222 thousand entity mentions, and 2.2 million tokens in each of the five low-resource Indian languages. After the creation of the dataset through the proposed method, 1000 sentences are randomly selected for human annotation. From the rest of the dataset, 10% is considered as the development set and the remaining as the training set.

3.4 Gold dataset

Two volunteer annotators per language, having a minimum education of an undergraduate degree, were chosen based on their mother tongue. For Sanskrit, the annotations were done by professional

Lng	Train set			Development set			Test set			
	Sent	Ent	Token	Sent	Ent	Token	Sent	Ent	Token	IAA(κ)
as	140,257	204,611	1,972,697	15,585	15,763	219,114	1000	1,407	14,270	0.901
brx	212,835	302,713	2,958,455	23,649	33,808	329,145	1000	1,423	14,082	0.875
mr	611,902	889,217	8,135,813	67,990	97,943	948,020	1000	1,443	13,996	0.887
ne	414,561	617,957	5,531,683	46,062	64,098	642,489	1000	1,436	14,142	0.882
sa	265,114	378,287	3,488,871	29,458	40,589	377,306	1000	1,412	12,925	0.861

Table 1: CLASSER dataset statistics. **Lng** means language, and **Sent**, **Ent** and **Token** means number of Sentences, Entities and Tokens respectively. **IAA** (κ) gives the inter-annotator agreement.

Lng	Dataset	Tokens	Ent	Tp
as	CLASSER	**2,206,081**	**221,781**	**33**
	Naamapadam[1]	<u>122,413</u>	5,045	3
	AsNER[2]	98,623	<u>34,963</u>	5
	WikiANN[3]	7,632	1,418	3
brx	CLASSER	**3,301,682**	<u>337,944</u>	**33**
	Bodo NER[4]	<u>2,797,101</u>	**641,604**	5
mr	CLASSER	**9,097,829**	**988,603**	**33**
	TAFSIL[5]†	3,628,450	174,861	**33**
	Naamapadam[1]	<u>6,086,136</u>	<u>529,000</u>	3
	MahaNER[6]	231,959	27,300	7
	WikiANN[3]	123,556	18,756	3
ne	CLASSER	**6,188,314**	**683,491**	**33**
	EverestNER[7]	<u>616,706</u>	<u>24,587</u>	5
	OurNepali[8]	16,225	11,183	4
	WikiANN[3]	14,535	2,326	3
sa	CLASSER	**3,879,102**	**420,288**	**33**
	TAFSIL[5]†	<u>479,185</u>	<u>23,372</u>	**33**
	WikiANN[3]	2,255	115	3

Table 2: Comparison of CLASSER dataset with some publicly available NER datasets in low-resource Indian languages: [1]Mhaske et al. (2023), [2]Pathak et al. (2022), [3]Rahimi et al. (2019), [4]Narzary et al. (2024), [5]Kaushik et al. (2025) (†: MultiCoNER2 taxonomy), [6]Litake et al. (2022), [7]Niraula and Chapagain (2022), and [8]Singh et al. (2019). Abbreviations: **Lng**: language, **Ent**: number of entity mentions, **Tp**: number of entity types.

Sanskrit teachers. Annotators were first briefed on entity types with examples, then instructed to perform the task on 1000 sentences in two stages: detecting relevant entity mentions and assigning types from a given list using the BRAT tool (Stenetorp et al., 2012). With two annotators per language, one annotator's work was treated as the gold standard, and inter-annotator agreement (IAA) was measured against it. The quality of these gold datasets can be ascertained based on a high Cohen's kappa coefficient (κ) (Deleger et al., 2012), which is above 0.86 for each language (Table 1).

3.5 Comparison with public dataset

As shown in Table 2, CLASSER is the largest NER dataset across all languages in terms of the tokens compared to the publicly available datasets. In fact, except for Bodo (brx), it is the largest dataset in terms of the number of entities as well. To the best of our knowledge, CLASSER is the only FgNER dataset created through the entity projection method for these five low-resource languages. In this paper, whenever comparative analysis is done, the **highest value** or the **best result** in the tables are shown in **bold** and the <u>second highest value</u> or the <u>second best result</u> are shown as <u>underlined</u>.

3.6 Entity type frequency distribution

As shown in Figure 2, a larger number of entity mentions are detected for the fine types of Location (e.g. HumanSettlement) and Person (e.g. Artist) because the HumanSettlement includes the mentions of cities, provinces and countries, and the Artist type includes the mentions of musicians, actors, directors, authors, etc. Whereas, very specific fine types such as AerospaceManufacturer, Drink, AnatomicalStructure, etc., are very scarce. Similar trends can be observed across all five languages (Figure 7 in Appendix).

4 Analysis & Results

4.1 Experimental Setup

The state-of-the-art approach for sequence labeling tasks involves fine-tuning pre-trained language models (PLM) with the NER datasets (Venkataramana et al., 2022; Litake et al., 2022; Malmasi et al., 2022; Mhaske et al., 2023; Fetahu et al., 2023a; Tulajiang et al., 2025; del Moral-González et al., 2025). Similarly, we have fine-tuned mBERT (bert-base-multilingual-cased) (Devlin et al., 2019), IndicBERTv2 (IndicBERTv2-MLM-Sam-TLM) (Doddapaneni et al., 2022),

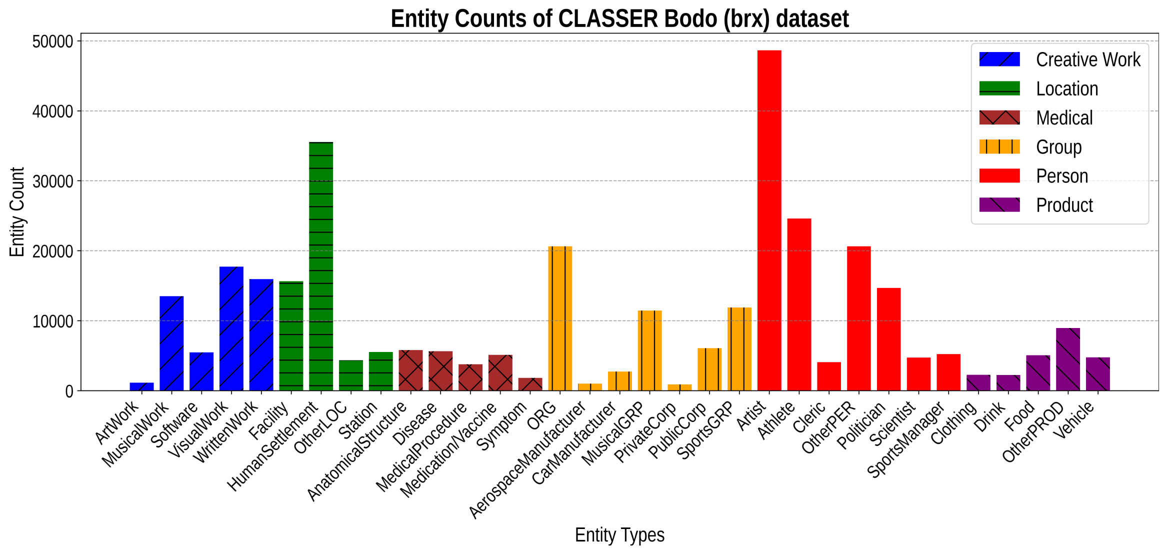

Figure 2: Entity counts of CLASSER Bodo (brx) dataset.

MuRIL (muril-large-cased) (Khanuja et al., 2021) and XLM-RoBERTa (XLM-RoBERTa-large) (Conneau et al., 2020) 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 following Golde et al. (2025); Ding et al. (2025).

To compare with the state-of-the-art noisy FgNER dataset, following Kaushik et al. (2025), we adopted the best-performing two variations of DECENT (Sierra-Múnera et al., 2023) to accommodate Indian languages by changing its base encoder from RoBERTa-large (Liu et al., 2019), to XLM-RoBERTa-large (Conneau et al., 2020), and IndicBERTv2-MLM-Sam-TLM (Doddapaneni et al., 2022). 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, epochs = 2, batch size = 16, negative oversampling rate = 31, and prediction threshold = 0.9.

4.2 Comparison with SOTA baseline

To the best of our knowledge, the only existing FgNER datasets in Indian languages are MultiCoNER2 (Fetahu et al., 2023a) for Hindi and Bengali, and TAFSIL (Kaushik et al., 2025) for Hindi, Marathi, Sanskrit, Tamil, Telugu, and Urdu. Accordingly, we used the Marathi and Sanskrit TAFSIL datasets in the MultiCoNER2 taxonomy and fine-tuned DECENT model variants as described in the previous section. As shown in Table 3, when tested on the TAFSIL test set, the models fine-tuned on CLASSER outperform models fine-tuned on TAFSIL by a large margin. With the subset of CLASSER train set having an equal number of entities as TAFSIL, the F1 scores improve by about 22% for Marathi and 38% for Sanskrit, respectively. Similarly, with the subset of CLASSER train set having an equal number of sentences as TAFSIL, the F1 scores improve by about 26% for Marathi and 40% for Sanskrit, and using the entire CLASSER dataset, gains rise to roughly 40% and 95%, respectively. These results demonstrate both the effectiveness of our method and the high quality of the generated CLASSER dataset.

4.3 Performance of PLMs on unseen languages

Marathi (mr) and Nepali (ne) are the only languages among the five languages on which all four PLMs (IndicBERTv2, mBERT, MuRIL, and XLM-RoBERTa) are pre-trained (Table 7 in Appendix). Therefore, the performance of all the fine-tuned models in Marathi and Nepali is superior (Table 4). Although mBERT was not pre-trained on Assamese (as) and Sanskrit (sa), it performs well after fine-tuning with CLASSER dataset on these languages. This is due to the script similarity between Assamese with Bengali (Bengali-Assamese script) and between Sanskrit with Hindi (Devanagari script). Similarly, although mBERT, MuRIL, and XLM-RoBERTa are not pre-trained in Bodo, the fine-tuned models could per-

DECENT model with XLM-RoBERTa-large base encoder

				Micro			Macro		
Lang	**Dataset**	**Sent**	**Ent**	**P**	**R**	**F1**($\uparrow\Delta\%$)	**P**	**R**	**F1**($\uparrow\Delta\%$)
mr	TAFSIL	126k	175k	49.36	83.73	62.10	50.89	81.93	62.28
	CLASSER	120k	175k	72.85	80.12	76.32(**23**)	73.45	79.13	75.61(**21**)
	CLASSER	126k	183k	76.12	80.97	78.28(**26**)	73.42	82.19	77.56(**25**)
	CLASSER	612k	989k	82.66	91.74	86.98(**40**)	84.40	91.78	87.94(**41**)
sa	TAFSIL	18k	23k	31.35	53.40	40.20	32.74	53.29	40.56
	CLASSER	16k	23k	52.54	59.03	55.04(**37**)	53.29	58.66	55.84(**38**)
	CLASSER	18k	25k	56.14	60.03	58.02(**44**)	57.25	60.99	59.06(**46**)
	CLASSER	265k	378k	77.21	81.69	79.38(**97**)	78.60	80.89	79.93(**97**)

DECENT model with IndicBERTv2-MLM-Sam-TLM base encoder

				Micro			Macro		
Lang	**Dataset**	**Sent**	**Ent**	**P**	**R**	**F1**($\uparrow\Delta\%$)	**P**	**R**	**F1**($\uparrow\Delta\%$)
mr	TAFSIL	126k	175k	48.57	82.81	61.23	49.33	81.49	61.46
	CLASSER	120k	175k	72.11	78.10	75.04(**23**)	72.62	76.03	74.34(**21**)
	CLASSER	126k	183k	76.65	80.12	77.29(**26**)	75.38	79.27	76.36(**24**)
	CLASSER	612k	989k	83.07	89.44	86.14(**41**)	81.75	89.37	85.43(**39**)
sa	TAFSIL	18k	23k	32.74	56.03	41.03	32.94	55.96	41.47
	CLASSER	16k	23k	56.29	60.66	57.82(**40**)	56.27	60.71	57.33(**38**)
	CLASSER	18k	25k	59.70	61.60	60.64(**48**)	59.34	61.50	60.36(**46**)
	CLASSER	265k	378k	79.52	78.94	79.01(**93**)	78.09	81.49	80.81(**95**)

Table 3: Performance of different DECENT models fine-tuned on TAFSIL and CLASSER train sets and tested on the TAFSIL test set. Abbreviations: **Sent**: Number of sentences, **Ent**: Number of entities.

form better due to their pre-training on the script-similar language Hindi. MuRIL, pre-trained exclusively on 16 Indian languages, outperforms other PLMs. These results emphasize the importance of language-specific pre-training and the effect of script-similarity in fine-tuning, which are discussed further in the following sections.

4.4 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. Zero-shot performance of mBERT model is quite poor across all the languages. A similar trend is observed in the case of XLM-RoBERTa on unseen languages during its pre-training. However, there is an improvement in the case of MuRIL because of its pre-training on 16 Indian languages. The impact of pre-training of an encoder is imminent through the zero-shot performance of IndicBERTv2. Since IndicBERTv2 is pre-trained on all five languages, its zero-shot performance is superior to other PLMs. Whereas, fine-tuning on Bodo (brx) significantly improved the performance of mBERT, XLM-RoBERTa, and MuRIL over their zero-shot performances. These emphasize that due to script similarity, PLMs can perform better after fine-tuning on an unseen language. But, without fine-tuning with language-specific datasets, the pre-trained knowledge cannot capture the intricacies of an unseen language.

4.5 Multilingualism

We have extended our analysis to evaluate the multilingual aspect of the FgNER task. We constructed a balanced **all5** set, comprising an equal number of samples from all five languages. As seen in Figure 3, there are significant improvements in every encoder model when fine-tuned with all the languages and tested on test sets of individual languages. In fact, for a vulnerable language like Bodo (brx), multilingual fine-tuning can be very beneficial. These results also suggest the necessity of language-specific pre-training and task-specific fine-tuning.

4.6 Ablation study

As already seen in different analyses, the script similarity of the language plays a major role in FgNER

	Micro P	Micro R	Micro F1	Macro P	Macro R	Macro F1
Assamese (as)						
IB	71.10	71.50	71.30	62.33	63.67	63.11
mB	66.09	68.30	67.18	64.12	61.48	63.13
MR	74.88	75.62	**75.25**	73.91	70.54	**72.44**
XL	71.35	73.63	<u>72.47</u>	69.50	69.42	<u>69.46</u>
Bodo (brx)						
IB	70.61	72.64	71.61	69.53	68.58	68.98
mB	68.75	71.03	69.87	68.65	68.23	68.59
MR	73.83	76.37	**75.08**	74.76	73.73	**74.08**
XL	71.60	73.77	<u>72.67</u>	73.28	71.66	<u>72.60</u>
Marathi (mr)						
IB	74.38	76.49	75.42	70.67	71.89	71.22
mB	74.83	76.28	75.55	69.81	71.57	70.82
MR	79.24	81.00	**80.11**	75.94	77.59	**76.83**
XL	78.58	78.85	<u>78.71</u>	73.55	75.36	<u>74.41</u>
Nepali (ne)						
IB	73.80	75.24	74.52	71.52	70.50	71.02
mB	74.33	75.52	74.92	73.40	72.40	72.91
MR	76.92	79.50	**78.19**	75.34	76.37	**75.88**
XL	74.93	78.80	<u>76.82</u>	73.32	75.95	<u>74.14</u>
Sanskrit (sa)						
IB	73.45	75.02	74.23	72.96	72.17	72.58
mB	70.26	72.25	71.24	71.41	69.24	70.35
MR	77.62	78.99	**78.30**	77.61	76.57	**77.04**
XL	75.41	77.50	<u>76.44</u>	74.79	74.17	<u>74.49</u>

Table 4: Performance of different models fine-tuned on CLASSER dataset. Abbreviations: IB: IndicBERTv2, mB: mBERT, MR: MuRIL, XL: XLM-RoBERTa.

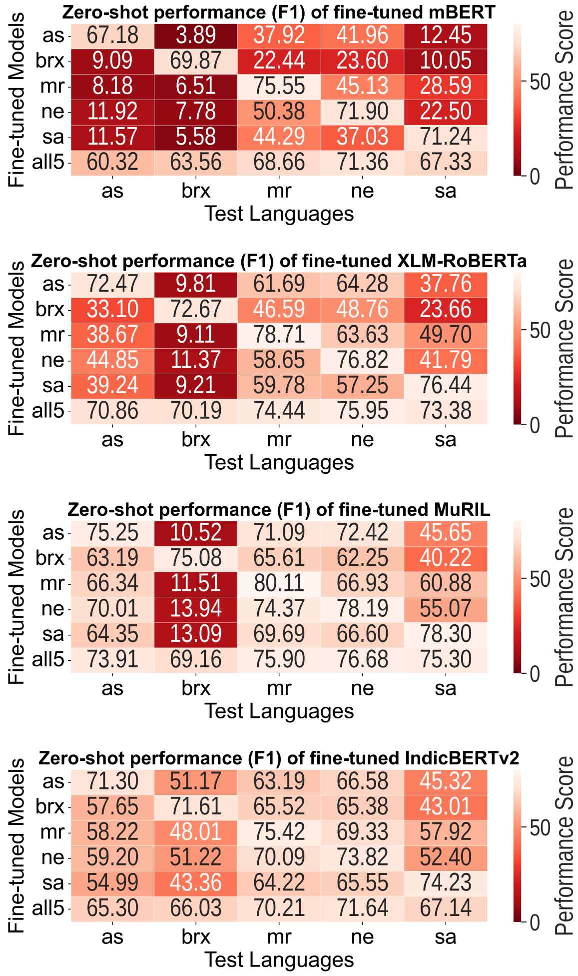

Figure 3: Zero-shot performance (micro F1) of fine-tuned mBERT, MuRIL, XLM-RoBERTa and IndicBERTv2 models on different test languages. The **all5** set includes samples from all five languages.

task. The intuition of cross-lingual refinement after annotation projection is based on this property. The performance of fine-tuned MuRIL models in terms of micro-F1 scores are shown in Table 5. For Assamese (as), refinement using Bengali (bn) is the most effective, since both of these languages use the Bengali-Assamese script. Similarly, for Bodo (brx), Marathi (mr), Nepali (ne) and Sanskrit (sa), refinement using Hindi (hi) is the most effective due to the shared Devanagari script. But refinement using both hi+bn gives the best result across all the languages. These improvements are highest for extremely low-resource languages Assamese (as), Bodo (brx), and Sanskrit (sa).

For the selection of the best confidence threshold τ in the cross-lingual refinement stage of CLASSER method, we conducted experiments with four empirically selected values 0.75, 0.80, 0.85, 0.90 for all the languages. In Figure 4, the performances of fine-tuned MuRIL on Assamese (as), Marathi (mr), and Nepali (ne) are shown. A similar trend is observed across other languages (Figure 5 in Appendix), and hence, based on the experiments, finally τ is set to be 0.85 for all the languages.

4.7 Error Analysis

FgNER is very crucial because an entity mention's type may vary significantly depending on the context within a sentence. Therefore, we have analyzed

	Stg 1	⊕ bn	⊕ hi	⊕ hi+bn
as	63.11	74.48(**18**)	67.43(**7**)	75.25(**19**)
brx	61.82	63.90(**3**)	73.84(**19**)	75.08(**21**)
mr	71.08	73.26(**3**)	78.85(**11**)	80.11(**13**)
ne	70.01	73.10(**4**)	77.05(**10**)	78.19(**12**)
sa	65.14	67.12(**3**)	76.87(**18**)	78.30(**20**)

Table 5: Ablation study: The impact of refinement using Bengali (⊕ bn) and Hindi (⊕ hi) on the initial annotation projection (Stg 1). The performance of fine-tuned MuRIL models in terms of micro-F1 scores are shown with **percentage improvements**.

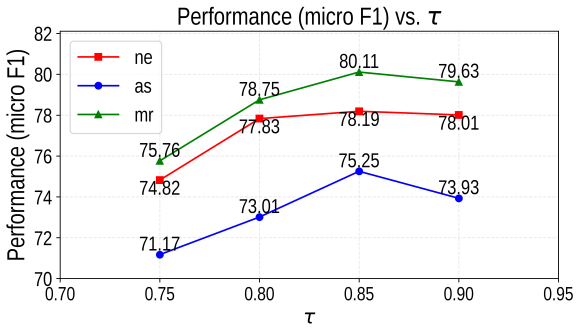

Figure 4: Impact of confidence score threshold (τ) on the performance of MuRIL fine-tuned on CLASSER dataset for Assamese (as), Marathi (mr), & Nepali (ne).

the errors on the test set in two different approaches. First, the details of entity errors in terms of the percentage of predicted entities are shown in Table 6. The common errors that occur include the boundary error (such as "Little Mermaid" is marked as *VisualWork* instead of "The Little Mermaid"), entity type mismatch error (e.g. "Sneezing" is categorized as *Disease* instead of *Symptom*) and spurious errors (such as "purple" is marked as an entity whereas the entity type *color* is not defined in MultiCoNER2 taxonomy). Entity boundary mismatch errors are highest in Assamese (as), entity type mismatch errors and spurious entity errors occur the most in Bodo (brx) and Sanskrit (sa), respectively.

Moreover, we have analyzed the often co-predicted fine-grained types. From Table 6, we have selected Bodo (brx) for this analysis as this language has the highest percentage of mismatch entity types. As shown in Figure 6 in Appendix, the fine types of *Artist*, *Athlete*, *Politician*, and *Scientist* are sometimes confused with *OtherPER*. Similarly, *WrittenWork* is sometimes confused with *VisualWork*. Apart from such closely related fine entity types, most of the other fine entity types are learned by the models without much confusion.

	as	brx	mr	ne	sa
BM	13.67	9.27	10.37	8.23	9.82
ET	13.75	14.40	11.44	12.35	11.32
SP	2.41	2.67	2.42	2.44	4.27

Table 6: Entity errors on test set in terms of the percentage of predicted entities for different languages fine-tuned on MuRIL. Abbreviations: BM: Boundary Mismatch, ET: Entity Type Mismatch, SP: Spurious Entity.

5 Conclusion & Future Works

We introduce CLASSER, a cross-lingual annotation-projection framework that leverages script similarity to create high-quality FgNER dataset. The generated CLASSER dataset comprises of 1.8M sentences, 2.6M entity mentions, and 24.7M tokens covering five low-resource languages (Assamese, Bodo, Marathi, Nepali, and Sanskrit). Extensive experiments confirm its quality, and zero-shot cross-lingual analyses reveal the importance of language-specific pre-training and task-specific fine-tuning. Given the availability of MultiCoNER2 FgNER resources in Bengali, Hindi, and Farsi, CLASSER can be readily extended to other Indian languages (e.g., Maithili, Konkani, Dogri, Bhojpuri, Chhattisgarhi, Bishnupriya Manipuri, Urdu, Kashmiri, Sindhi etc.). We expect the CLASSER framework, the generated dataset, and fine-tuned models will significantly advance FgNER across Indian languages and facilitate further developments in multilingual research.

Limitations

Despite encouraging initial results, this study has limitations that require further exploration. First, the proposed cross-lingual refinement stage relies primarily on resources in languages with similar scripts, for which its direct applicability and scalability to languages with significantly different syntactic structures remain an open question for future research. Second, the generated dataset may inherit biases or specificities from the source FgNER dataset. Third, the dataset's volume and quality depend on the availability of parallel corpora and the choice of annotation projection tools and multilingual encoders. Assessment of different combinations of these resources remains essential. Moreover, while a confidence score of 0.85 yielded the best results among the four tested empirical values (0.75, 0.80, 0.85, 0.90), a more systematic and theoretically grounded analysis is needed. Finally, evaluating Large Language Models (LLMs) on the FgNER task, both in a zero-shot setting and fine-tuned on the CLASSER dataset, remains underexplored.

Ethical considerations

The annotations were generated using the openly accessible MultiCoNER2 dataset[3] and BPCC parallel corpora[4] released under CC-BY-4.0[5] and CC0[6] 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. The dataset created in this work is available at huggingface.co/datasets/prachuryyaIITG/CLASSER under an MIT license[7].

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*.

Noëmi Aepli and Rico Sennrich. 2022. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

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.

[3]https://multiconer.github.io/dataset

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

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

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

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

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.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does manipulating tokenization aid cross-lingual transfer? a study on POS tagging for non-standardized languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 40–54, Dubrovnik, Croatia. Association for Computational Linguistics.

Maharaj Brahma, Kaushal Maurya, and Maunendra Desarkar. 2023. Selectnoise: Unsupervised noise injection to enable zero-shot machine translation for extremely low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1615–1629.

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.

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

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

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*.

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.

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, and 1 others. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144.

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*.

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)*.

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.

Zhuojun Ding, Wei Wei, and Chenghao Fan. 2025. Selecting and merging: Towards adaptable and scalable named entity recognition with large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9869–9886, Vienna, Austria. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2022. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. *arXiv preprint arXiv:2212.05409*.

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)*.

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*.

Asif Ekbal and Sriparna Saha. 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.

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

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

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.

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*.

Jonas Golde, Patrick Haller, Max Ploner, Fabio Barth, Nicolaas Jedema, and Alan Akbik. 2025. Familiarity: Better evaluation of zero-shot named entity recognition by quantifying label shifts in synthetic training data. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 820–834, Albuquerque, New Mexico. 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.

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

Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. 2019. Entity projection via machine translation for cross-lingual ner. 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 1083–1092.

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.

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*.

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.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, and 1 others. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Brayan Stiven Lancheros, Gloria Corpas Pastor, and Ruslan Mitkov. 2024. Data augmentation and transfer learning for cross-lingual named entity recognition in the biomedical domain. *Language Resources and Evaluation*, pages 1–20.

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.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. Mulda: A multilingual data augmentation framework for low-resource cross-lingual ner. 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 5834–5846.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

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.

Stephen Mayhew, Chen-Tse Tsai, and Dan Roth. 2017. Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2536–2545.

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.

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

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.

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.

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.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, Dublin, Ireland. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Ankur Priyadarshi and Sujan Kumar Saha. 2021. The first named entity recognizer in maithili: Resource creation and system development. *Journal of Intelligent & Fuzzy Systems*, 41(1):1083–1095.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for ner. *arXiv preprint arXiv:1902.00193*.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, and 1 others. 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.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and 1 others. 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. 2008. Word clustering and word selection based feature reduction for maxent based hindi ner. In *proceedings of ACL-08: HLT*, pages 488–495.

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*.

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).

Alejandro Sierra-Múnera, Jan Westphal, and Ralf Krestel. 2023. Efficient ultrafine typing of named entities. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 205–214. IEEE.

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.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

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*.

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*.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250.

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

Rudra Murthy Venkataramana, Pallab Bhattacharjee, Rahul Sharnagat, Jyotsana Khatri, Diptesh Kanojia, and Pushpak Bhattacharyya. 2022. Hiner: A large hindi named entity recognition dataset. In *International Conference on Language Resources and Evaluation*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and 1 others. 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.

Jian Yang, Shaohan Huang, Shuming Ma, Yuwei Yin, Li Dong, Dongdong Zhang, Hongcheng Guo, Zhoujun Li, and Furu Wei. 2022. Crop: Zero-shot cross-lingual named entity recognition with multilingual labeled sequence translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 486–496.

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*.

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.

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.

A Appendix

A.1 Selection of τ value

Similar to Assamese (as), Marathi (mr), and Nepali (ne) as shown in Figure 4, the trend is observed for Bodo (brx) and Sanskrit (sa) as well (Figure 5). Hence, based on the experiments, finally τ is set to be 0.85 for all the languages.

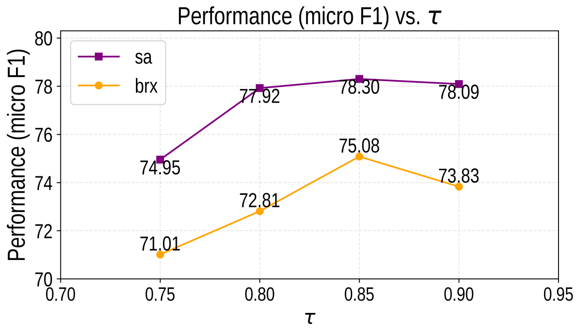

Figure 5: Impact of confidence score threshold (τ) on the performance of MuRIL fine-tuned on CLASSER dataset for Bodo (brx) & Sanskrit (sa).

A.2 Error Analysis

We examined the fine-grained types that are frequently co-predicted. We focused this analysis on the test set of Bodo (brx), as shown in Table 6, since it exhibits the highest percentage of mismatch entity types. As illustrated in Figure 6, fine types like *Artist*, *Athlete*, *Politician*, and *Scientist* are often confused with *OtherPER*, while *WrittenWork* is occasionally mistaken for *VisualWork*. Beyond these closely related categories, most other fine-grained entity types are accurately learned by the models with minimal confusion.

A.3 Pre-trained Language Models details

The details of encoder models used in this work, i.e. bert-base-multilingual-cased[8] (Devlin et al., 2019), IndicBERTv2-MLM-Sam-TLM[9] (Doddapaneni et al., 2022), muril-large-cased[10] (Khanuja et al., 2021) and XLM-RoBERTa-large[11] (Conneau et al., 2020) are shown in Table 7.

[8]https://huggingface.co/google-bert/bert-base-multilingual-cased

[9]https://huggingface.co/ai4bharat/IndicBERTv2-MLM-Sam-TLM

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

[11]https://huggingface.co/FacebookAI/xlm-roberta-large

Model	Parameters	No. of Languages	Indian languages covered
bert-base-multilingual-cased (Devlin et al., 2019)	110M	104	Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Panjabi, Tamil, Telugu
IndicBERTv2-MLM-Sam-TLM (Doddapaneni et al., 2022)	278M	26	Assamese, Bengali, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Marathi, Manipuri, Muna, Nepali, Oriya, Panjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu, Urdu
muril-large-cased (Khanuja et al., 2021)	340M	17	Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Sindhi, Tamil, Telugu, Urdu
XLM-RoBERTa-large (Conneau et al., 2020)	355M	100	Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Oriya, Panjabi, Tamil, Telugu, Urdu

Table 7: Details of multilingual encoder models used: size, languages pretrained on, Indian languages covered

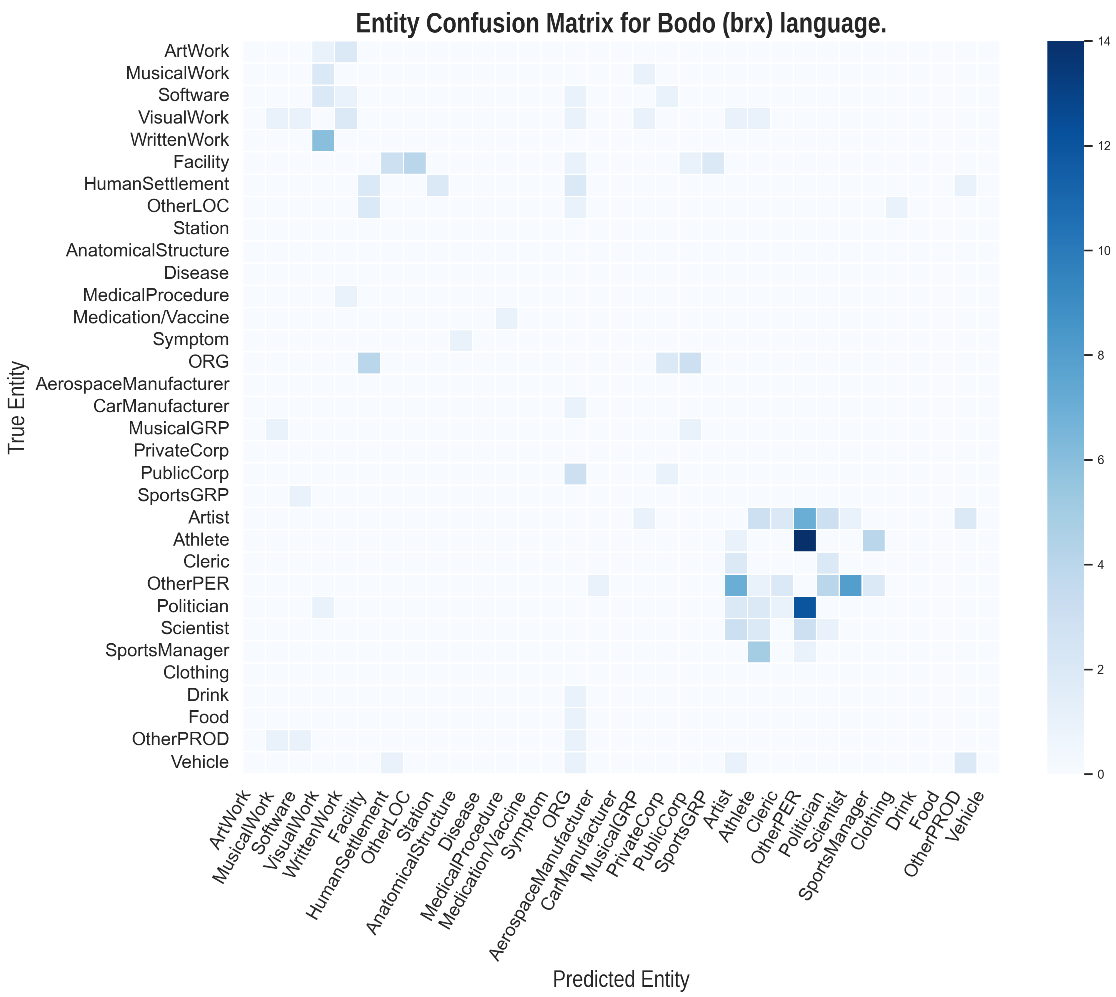

Figure 6: Entity Confusion matrix of Bodo (brx) language.

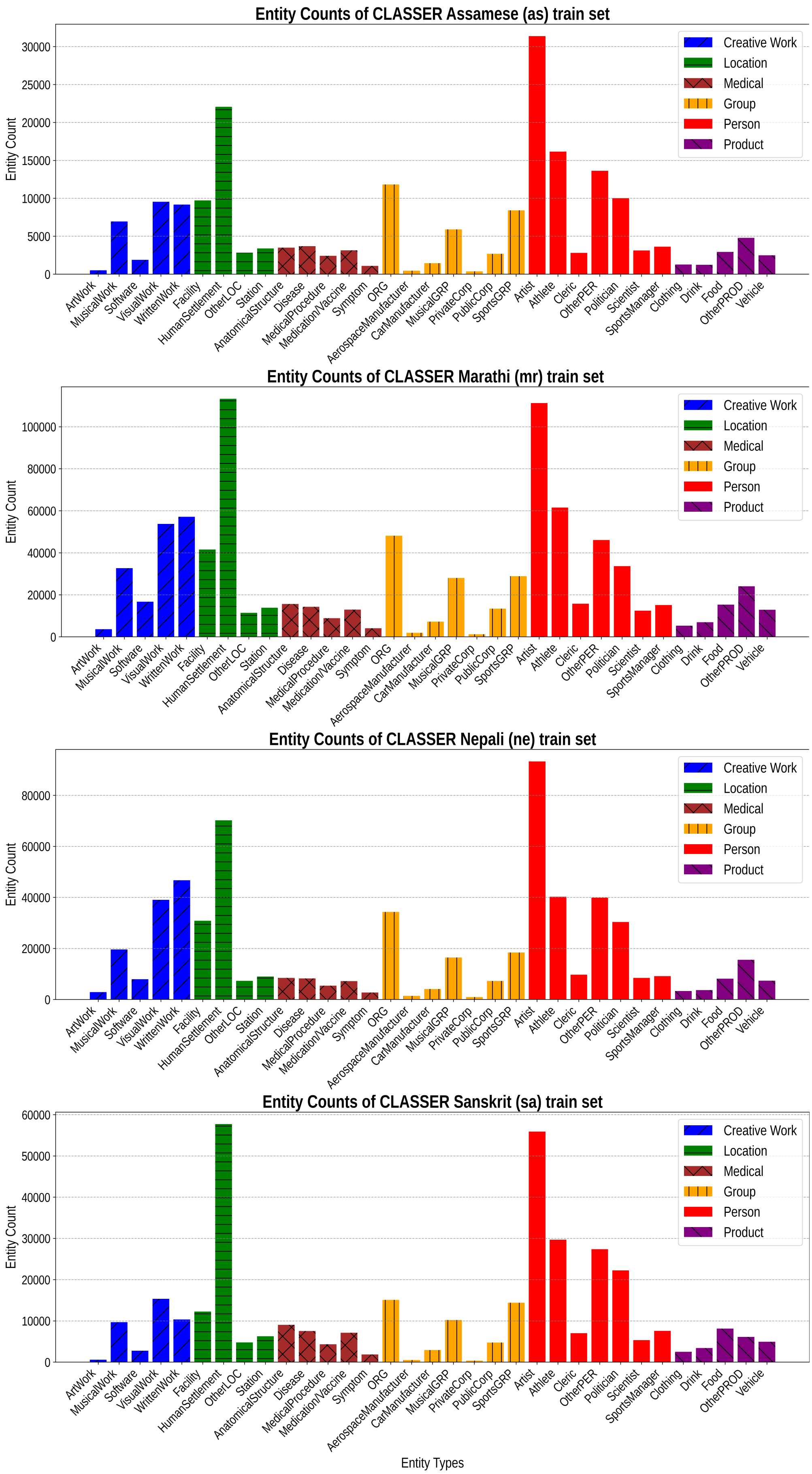

Figure 7: Entity counts of CLASSER train sets of Assamese (as), Marathi (mr), Nepali (ne), and Sanskrit (sa).