Naamapadam: A Large-Scale Named Entity Annotated Data for Indic Languages

Arnav Mhaske[1,2*] **Harshit Kedia**[1,2*]
Sumanth Doddapaneni[1,2] **Mitesh M. Khapra**[1,2] **Pratyush Kumar**[1,2,3]
Rudra Murthy V[2,4†] **Anoop Kunchukuttan**[1,2,3†]

[1]Indian Institute of Technology Madras [2]AI4Bharat
[3]Microsoft India [4]IBM Research India

Abstract

We present, *Naamapadam*, the largest publicly available Named Entity Recognition (NER) dataset for the 11 major Indian languages from two language families. The dataset contains more than 400k sentences annotated with a total of at least 100k entities from three standard entity categories (Person, Location, and, Organization) for 9 out of the 11 languages. The training dataset has been automatically created from the Samanantar parallel corpus by projecting automatically tagged entities from an English sentence to the corresponding Indian language translation. We also create manually annotated testsets for 9 languages. We demonstrate the utility of the obtained dataset on the Naamapadam-test dataset. We also release *IndicNER*, a multilingual IndicBERT model fine-tuned on Naamapadam training set. IndicNER achieves an F1 score of more than 80 for 7 out of 9 test languages. The dataset and models are available under open-source licences at `https://ai4bharat.iitm.ac.in/naamapadam`.

1 Introduction

Named Entity Recognition (NER) is a fundamental task in natural language processing (NLP) and is an important component for many downstream tasks like information extraction, machine translation, entity linking, co-reference resolution, etc. The most common entities of interest are `person`, `location`, and, `organization` names, which are the focus of this work and most work in NLP. Given high-quality NER data, it is possible to train good-quality NER systems with existing technologies (Devlin et al., 2019). For many high-resource languages, publicly available annotated NER datasets (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003a; Pradhan et al., 2013; Benikova et al., 2014) as well as high-quality taggers (Wang et al., 2021; Li et al., 2020) are available.

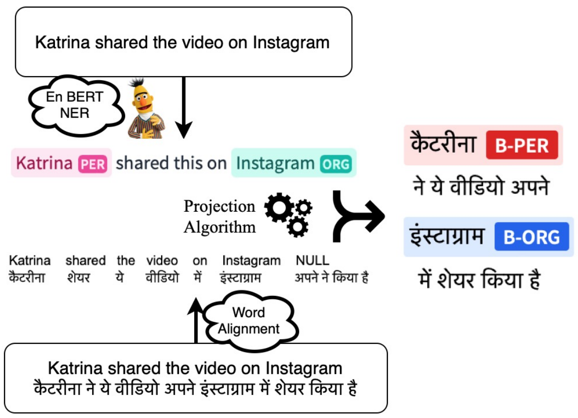

Figure 1: Illustration of Named Entity projection. We perform (i) NER with fine-tuned English BERT model, followed by (ii) word alignment between parallel sentence pair and (iii) projection of the English entities onto Indic sentence.

However, most Indic languages do not have sufficient labeled NER data to build good-quality NER models. All existing NER corpora for Indic languages have been manually curated (Lalitha Devi et al., 2014; Murthy et al., 2018; Pathak et al., 2022; Murthy et al., 2022; Malmasi et al., 2022; Litake et al., 2022). Given the number of languages, the expenses and the logistical challenges, these datasets are limited on various dimensions *viz.* corpus size, language coverage, and broad domain representation. In recent years, zero-shot cross-lingual transfer from pre-trained models, fine-tuned on task-specific training data in English has been proposed as a way to support various language understanding tasks for low-resource languages (Hu et al., 2020). However, this approach is more suitable for semantic tasks and the cross-lingual transfer does not work as well for syntactic tasks like NER when transferring across distant languages like English and Indian languages (Wu and Dredze, 2019; Karthikeyan et al., 2020; Ruder et al., 2021). Hence, there is a need for in-language NER training

* Equal contribution
† Project leads; correspondence to rmurthyv@in.ibm.com, ankunchu@microsoft.com

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pages 10441–10456
July 9-14, 2023 ©2023 Association for Computational Linguistics

	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te
Naamapadam	5.0K	1.6M	769.3K	2.2M	658K	1.0M	735.0K	190.0K	880.2K	745.2K	751.1K
WikiANN	218	12K	264	7.3K	220	13K	7.3K	265	211	19.7K	2.4K
FIRE-2014	-	6.1K	-	3.5K	-	4.2K	-	-	-	3.2K	-
CFILT	-	-	-	262.1K	-	-	4.8K	-	-	-	-
MultiCoNER	-	9.9K	-	10.5K	-	-	-	-	-	-	-
MahaNER	-	-	-	-	-	-	16K	-	-	-	-
AsNER$^{\phi}$	6K	-	-	-	-	-	-	-	-	-	-

Table 1: Comparison of Indian language Named Entity training dataset statistics (total number of named entities), For all datasets, the statistics include only LOC, PER and ORG named entities. ϕ - While the dataset contains a total of 29K entities, most of the examples are gazetteer entries without sentence context.

data for Indic languages.

In recent years, the paradigm of mining datasets from publicly available data sources has been successfully applied to various NLP tasks for Indic languages like machine translation (Ramesh et al., 2022), machine transliteration (Madhani et al., 2022) and many natural language understanding and generation tasks (Kakwani et al., 2020; Kumar et al., 2022). These approaches have led to the creation of large-scale datasets and models with broad coverage of Indic languages in a short amount of time. Taking inspiration from these successes, we also explore the automatic creation of NER datasets by utilizing publicly available parallel corpora for Indian languages and high-quality English-named entity taggers. In this work, we undertake the task of building large-scale NER datasets and models for all major Indic languages.

The following are the contributions of our work:

• We build Naamapadam[1], the largest publicly available NER dataset for Indic languages for 11 languages from **2** language families. Naamapadam contains **5.7M** sentences and **9.4M** entities across these languages from three categories: `PERSON`, `NAME`, and `LOCATION`. This is significantly larger than other publicly available NER corpora for Indian languages in terms of the number of named entities and language coverage. Table 1 compares Naamapadam with other NER datasets.

• We create the Naamapadam test set, containing human-annotated test sets for 9 languages on general domain corpora, that can help in benchmarking NER models for Indic languages. Existing testsets are limited to fewer languages or are domain-specific.

• We also train a multilingual NER model, IndicNER, supporting 11 Indic languages. Our models achieve more than 80% F1 score on most languages on the Naamapadam test set. To the best of our knowledge, no publicly available NER models exist for Indian languages.

• We create the NER training corpora by projecting annotations from English sentences to their Indic language translations in parallel corpora. We show that the projection approach is better than approaches based on zero-shot transfer or teacher-student training. This allows for the inexpensive creation of data, at scale, while maintaining high quality. Hence, we recommend the use of a projection approach compared to these approaches when a reasonable amount of parallel corpora is available. This is a valid assumption for many mid-resource languages which today lack good NER models.

2 Related Work

We discuss the state of NER datasets for Indian languages and common methods used to improve NER for low-resource languages.

2.1 NER data for Indian languages

Very limited NER corpora are available for Indian languages. They are mostly small in size and do not cover all major Indian languages. The FIRE-2014 dataset (Lalitha Devi et al., 2014) is available for 4 languages. It was created by collecting sentences/documents from Wikipedia, blogs, and, online discussion forums. The WikiAnn dataset (Pan et al., 2017) is available for around 16 Indian languages - but these are all Wikipedia article titles that are not representative of natural language sentences and the annotations are very noisy. Moreover, the examples are Wikipedia article titles and are not representative of natural language sentences. Murthy et al. (2022) contributed the largest human-annotated dataset for Hindi (CFILT-Hindi) in terms of volume and diversity with over 100k sentences, all annotated by a single expert individual over a span of several years. There are

[1] Naamapadam means named entity in Sanskrit

a few small datasets for Indian languages: CFILT-Marathi (Murthy et al., 2018), MahaNER (Litake et al., 2022), AsNER (Pathak et al., 2022) and MultiCoNER (Malmasi et al., 2022). In contrast, Naamapadam has greater language coverage and is much larger compared to other datasets. It is also representative of general domain text.

2.2 Annotation Projection

Named entity corpora can be created for low-resource languages by projecting named entity annotations from sentences in the source language (high-resource) to the corresponding words in the translated sentence in the target language (low-resource). Yarowsky et al. (2001) first demonstrated how annotations can be projected using word alignments given parallel corpora between two languages. In addition to word alignments, projection can also be based on matching tokens via translation and entity dictionaries as well as transliteration (Zhang et al., 2016; Jain et al., 2019). Agerri et al. (2018) extended this approach to multiple languages by utilizing multi-way parallel corpora to project named entity labels from multiple source languages to the target language. When parallel corpus is not available, but good quality MT systems are available, annotated corpora in one language can be translated to another language followed by annotation projection (Jain et al., 2019; Shah et al., 2010). Bilingual dictionaries or bilingual embeddings have been used as a cheap alternative for translation in low-resource scenarios (Mayhew et al., 2017; Xie et al., 2018). The WikiAnn project creates '*silver standard*' NER corpora using a weakly supervised approach leveraging knowledge bases and cross-lingual entity links to project English entity tags to other languages (Pan et al., 2017). Given the availability of sufficient parallel corpora for major Indian languages (Ramesh et al., 2022), we use the annotation projection approach for building Indian language NER datasets.

2.3 Zero-shot Cross-lingual Transfer

This is a popular method for low-resource languages that relies on shared multilingual representations to help low-resource languages by transferring information from high-resource language NER models. Particularly, NER models finetuned on pre-trained language models like mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020) for high resource languages are used to tag low-resource language sentences (zero-shot NER). Pires et al. (2019) demonstrate that multilingual models perform well for zero-shot NER transfer on related languages. However, zero-shot performance is limited for distant languages Wu and Dredze (2019), particularly when there are structural/word order differences between the two languages (Karthikeyan et al., 2020). Unlike many other NLP tasks, zero-shot cross-lingual NER has seen only limited benefit from recent advances in cross-lingual representation learning (Ruder et al., 2021). To overcome this limitation, a knowledge distillation approach has been proposed to create synthetic in-language training data (Wu et al., 2020). Here, the source language *teacher* NER model is used to create distillation data in the target language via zero-shot cross-lingual transfer, which is used to train a target language model. We find that the projection-based approaches outperform zero-shot transfer and knowledge distillation approaches.

3 Mining NER Corpora

Following Yarowsky and Ngai (2001a,b), our method for building NER corpora is based on projecting NER tags from the English side of an English-Indic language parallel corpora to the corresponding Indic language words. For our work, we use the Samanantar parallel corpus (Ramesh et al., 2022) which is the largest publicly available parallel corpora between English and 11 Indic languages. Figure 1 illustrates our workflow for extracting named entity annotated Indic sentences from an English-Indic parallel sentence pair. It involves the following stages: (a) tagging the English sentence with a high-accuracy English NER model (Sec 3.1), (b) aligning English and Indic language words in the parallel sentence pair (Sec 3.2), (c) projecting NER tags from the English sentence to Indic words using the word alignments (Sec 3.3). These stages are further described in this section.

3.1 Tagging the English side

We tag the named entities on the English side of the parallel corpus using a publicly available, high-quality off-the-shelf English NER tagger. We evaluated various English taggers on the CoNLL dataset (see Table 9 for comparison). Since the parallel corpora contain a significant number of Indian named entities, we also performed a manual analysis to understand the taggers' performance on these entities. Based on these comparisons, we used the *BERT-*

base-NER[2] model for tagging the English portion of the Samanantar parallel corpus. We ignore the *MISC* tags predicted by the BERT-base-NER and focus on `PERSON`, `LOCATION`, and `ORGANIZATION` tags only. `MISC` is a very open-ended category and we found that it was not easy to reliably align `MISC` tagged entities from English to Indian languages.

3.2 Word Alignment

For every sentence pair in the parallel corpus, we align English words to the corresponding Indic language words. We explore two approaches for learning word alignments: (a) GIZA++ (Och and Ney, 2003) with default settings, (b) Awesome-align (Dou and Neubig, 2021) finetuned on parallel corpora with Translation Language Modeling and Self-training objectives. We use *softmax* to normalize the alignment scores in our experiments.

3.3 Projecting Named Entities

The next step is the projection of named entity labels from English to the Indic language side of the parallel corpus using English-Indic language word alignment information. We want the entity projection algorithm to ensure the following: (1) adjacent entities of the same type should not be merged into one single entity, and (2) small errors in word alignment should not cause drastic changes in the final NER projection. To ensure these, we project the entities as a whole (*i.e.,* the entire English entity phrase and not word by word) by identifying the minimal span of Indic words that encompass all the aligned Indic words. Word alignment errors could lead to incorrect named entity projection as shown in Figure 2. In this case, alignments in one direction are erroneous leading to wrong projection. We rely on the intersection of alignments in both directions to reduce alignment errors and thus ensure improved projection as illustrated in Figure 3. We show some examples of projections from Awesome-align in Appendix C.

In Figure 2 and 3, we use black arrows to indicate the alignment from Hindi to English direction and blue arrows to indicate the alignment from English to Hindi. The alignment from Hindi to English is correct. On the contrary, the alignment in English to Hindi direction suffers due to the presence of additional Hindi words. The word 'Soren' gets aligned to additional Hindi words 'photo' and 'PTI' which are not part of `PERSON` named entity (Figure 2). In order to minimize such errors, we take advantage of bidirectional alignments. We take the intersection of alignments in both directions, which improves the precision of alignments and hence improves projection accuracy (Figure 3). We will include the description in the revised version. Figure C is described in detail in Appendix C.

3.4 Sentence Filtering

After NER projection, we apply the following filters to the tagged Indic sentences.

Sentences without Named Entities. Many English sentences in the Samanantar corpus are not annotated with any entities. We retain only a small fraction of such sentences ($\approx 1\%$) for training the NER model so the model is exposed to sentences without any NER tags as well.

Sentences with low-quality alignments. We observe that most of the errors in the Indic-tagged corpus arise from word alignment errors. Hence, we compute a word alignment quality score for each sentence pair. This score is the product of the probability of each aligned word pair (as provided by the forward alignment model in the case of GIZA++ and the alignment model by awesome align) normalized by the number of words in the sentence. We retain the top 30-40% sentences to create the final NER-tagged data for Indic languages (See Table 11 for filtered data statistics).

3.5 Qualitative Analysis

To quantify the quality of the labeled data obtained, we select a small sample of 50 sentences[3] and obtain manual annotation for the 9 languages namely *Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil,* and, *Telugu*. We also project the named entities on this small set of 50 sentences using the projection approach discussed earlier. Since the ground truths are known, the F1 scores can be calculated. Table 2 presents the F1 scores on the manually labeled set using various projection approaches. We observe that both GIZA++ and *Awesome-align* word alignment approaches obtain similar performance. On average, *Awesome-align* provides the best F1 scores, hence, moving forward, we consider the datasets from the *Awesome-align* approach unless specified otherwise.

[2] https://huggingface.co/dslim/bert-base-NER

[3] this set is later expanded by annotating more sentences to form the test set

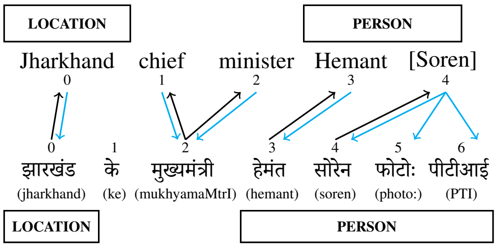

Figure 2: Error in Word Alignments could lead to incorrect entity projection

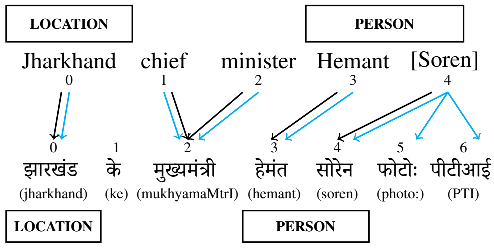

Figure 3: Understanding how the intersection helps reduce errors

Language	bn	gu	hi	kn	ml	mr	ta	te	Average
Awesome-align	82.11	69.77	**90.32**	70.22	**69.83**	**76.51**	**70.09**	77.70	**75.82**
GIZA++	**86.55**	**72.91**	85.22	**71.56**	64.38	76.21	56.82	**79.09**	74.09

Table 2: F1 scores from Different Projection Methods

Lang.	Sentence Count			Train			Dev			Test		
	Train	Dev	Test	Org	Loc	Per	Org	Loc	Per	Org	Loc	Per
bn	961.7K	4.9K	607	340.7K	560.9K	725.2K	1.7K	2.8K	3.7K	207	331	457
gu	472.8K	2.4K	1.1K	205.7K	238.1K	321.7K	1.1K	1.2K	1.6K	419	645	673
hi	985.8K	13.5K	867	686.4K	731.2K	767.0K	9.7K	10.2K	10.5K	521	613	788
kn	471.8K	2.4K	1.0K	167.5K	177.0K	310.5K	882	919	1.6K	291	397	614
ml	716.7K	3.6K	974	234.5K	308.2K	501.2K	1.2K	1.6K	2.6K	309	482	714
mr	455.2K	2.3K	1.1K	164.9K	224.0K	342.3K	868	1.2K	1.8K	391	569	696
pa	463.5K	2.3K	993	235.0K	289.8K	351.1K	1.1K	1.5K	1.7K	408	496	553
ta	497.9K	2.8K	758	177.7K	281.2K	282.2K	1.0K	1.5K	1.6K	300	388	481
te	507.7K	2.7K	847	194.1K	205.9K	347.8K	1.0K	1.0K	2.0K	263	482	608
as	10.3K	52	51	2.0K	1.8K	1.2K	18	5	3	11	7	6
or	196.8K	993	994	45.6K	59.4K	84.6K	225	268	386	229	266	431

Table 3: Statistics for the Naamapadam dataset. The testsets for `as` and `or` are silver standard. Work on the creation of larger, manually annotated testsets is in progress for these languages.

3.6 Dataset Statistics

Table 3 shows the statistics of the final Naamapadam dataset. We create train, dev, and, test splits. Testsets are then manually annotated as described later in Section 4. Most languages have training datasets of more than $100K$ sentences and $500K$ entities each. Some languages like Hindi have more than $1M$ sentences in the training set. Compared to other datasets (See Table 1), the Naamapadam has a significantly higher number of entities. Even though the dataset is slightly noisy due to alignment errors, we hope that the large dataset size can compensate for the noise as has been seen in many NLP tasks (Bansal et al., 2022).

We have manually annotated testsets of around 500-1000 sentences for most languages. The `Assamese` and `Oriya` testsets are silver-standard (the named entity projections have not been verified yet). Work on the creation of larger, manually annotated testsets for these languages is in progress.

4 Testset Creation

We have created Naamapadam-test: manually annotated test set for Indian language NER evaluation. The Naamapadam-test comprises 500-1000 annotated sentences per language for 9 languages namely Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, and, Telugu. The annotators were provided sentences with named entity annotations obtained using the methodology described in Section 3. The annotators had to verify if the projected NER annotations were correct and rectify the annotations if incorrect. They were asked to follow the CoNLL 2003 annotation guidelines (Tjong Kim Sang and De Meulder, 2003b). The human annotations were contributed by volunteers who are native language speakers.

Language	F1-Score	Token-level Cohen's Kappa	
		All Tokens	Entity Tokens
Bengali	78.51	0.8506	0.7033
Gujarati	74.45	0.7965	0.6169
Hindi	93.60	0.9536	0.8996
Kannada	89.20	0.9217	0.8452
Malayalam	84.28	0.9006	0.8156
Marathi	88.20	0.9037	0.8047
Punjabi	60.01	0.6948	0.4605
Tamil	64.29	0.7176	0.5209
Telugu	78.40	0.8888	0.7850

Table 4: IAA scores for 9 languages in the Naamapadam testset, We report F1-Score and token-level Cohen's Kappa for IAA. (Deleger et al., 2012). Token-level Cohen's Kappa* refers to the configuration where we consider tokens that are part of at least one mention according to at least one annotator (Ringland et al., 2019).

4.1 Inter-Annotator Agreement

We compute the inter-annotator agreement on a sample of two annotators for each language using Cohen's kappa coefficient (Cohen, 1960). The scores are shown in Table 4. They are all above 69% signifying good-quality annotations.

5 Experimental Setup

We analyze the performance of models trained on the Naamapadam-train dataset with alternative approaches for low-resource NER and to models trained on publicly available datasets. To this end, we investigate the following research questions:

RQ1: Are models trained on data obtained from projection approach (Naamapadam-train) better than zero-shot and teacher-student models?

RQ2: How does the model trained on publicly-available dataset fare against the model trained on Naamapadam-train data? We evaluate it on the Naamapadam-test set.

5.1 Train Dataset

In order to demonstrate the usefulness of our Naamapadam-train dataset, we fine-tune the mBERT model (Devlin et al., 2019) on the Naamapadam-train data and test on Naamapadam-test set. We additionally fine-tune the mBERT model on the train split of publicly available datasets and test on Naamapadam-test set. We consider the following datasets in our experiments

- **WikiANN**: We use the train split of the data released by Rahimi et al. (2019). Due to the languages covered, this is the most widely used dataset. However, we observe the tagged data to be highly erroneous and does not contain complete sentences, but just titles. Appendix A discusses the issues with the WikiNER dataset.

- **FIRE-2014**: The FIRE-2014 dataset (Lalitha Devi et al., 2014) contains named entity annotated dataset for Hindi, Bengali, Malayalam, Tamil, and, English languages. We train language-specific models on the train splits of these datasets and evaluate the performance on our test set.

- **MultiCoNER**: We use the Hindi and Bengali named entity annotated data from Malmasi et al. (2022).[4]

- **CFILT**: We use the CFILT-HiNER dataset created for Named Entity Recognition in Hindi language (Murthy et al., 2022). The dataset was from various government information web pages, and newspaper articles. The sentences were manually annotated. We also use the CFILT-Marathi dataset created for Named Entity Recognition in Marathi (Murthy et al., 2018).

- **MahaNER**: We use the Marathi named entity annotated data from Litake et al. (2022).

For a fair comparison with models trained on our dataset, we include only `PERSON`, `LOCATION`, and, `ORGANIZATION` entities. The rest of the named entities if present (FIRE 2014, CFILT Marathi, MultiCoNER) are considered non-named entities.

5.2 NER Fine-tuning

Recently, sequence labeling via fine-tuning of pretrained language models has become the norm (Devlin et al., 2019; Conneau et al., 2020; Kakwani et al., 2020). We fine-tune the pre-trained mBERT model (Devlin et al., 2019) and report the results in our experiments. The input to the model is a sequence of sub-word tokens that pass through the Transformer encoder layers. The output from the transformer is an encoder representation for each token in the sequence. We take the encoder representation of the first sub-word (in case the word gets split into multiple sub-words) and is passed through the output layer. The output layer is a linear layer followed by the softmax function. The

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

model is trained using cross-entropy loss. We use the Dhamecha et al. (2021) toolkit for fine-tuning our models.

5.3 Baseline Comparison

Our proposed approach can be seen as a cross-lingual approach since the training data is created by projection from English to Indic sentences. Hence, we compare the performance of our model with zero-shot learning (Pires et al., 2019) and teacher-student learning (Wu et al., 2020). We describe the baseline approaches in detail below:

5.3.1 Zero-shot NER

To perform Zero-shot transfer, we consider the mBERT model fine-tuned for NER task in English. We use the publicly available fine-tuned NER model[5] which is trained for NER in 10 high-resource languages (English, Arabic, Chinese, and some European languages). We directly test the performance of this model on Naamapadam large-test dataset (Bn, Gu, Hi, Kn, Ml, Mr, Pa, Ta, Te) and Naamapadam small-test datasets (As, Or) respectively.

Langs.	ZS	Teacher Student	Mined Data	
			GIZA++	Awesome Align
bn	64.83	63.07	79.35	81.02
gu	61.31	39.98	76.00	80.59
hi	71.77	73.62	80.44	82.69
kn	65.37	43.96	74.01	80.33
ml	70.47	70.97	72.35	81.49
mr	71.94	61.09	74.34	81.37
pa	58.03	44.90	70.91	71.51
ta	61.21	42.72	72.50	73.36
te	66.55	48.33	78.26	82.49
as	25.40	13.03	13.04	45.37
or	1.71	0.22	18.77	25.01
Mean	56.96	46.36	64.54	**71.38**

Table 5: F1 scores of various models on Naamapadam Test Set. Zero-Shot (ZS) and Teacher-Student models are described in Section 5.3.1 and 5.3.2 respectively.

5.3.2 Teacher-Student Learning

We use the publicly available fine-tuned NER model[5] to create synthetic named entity annotated training data for the Indic languages. We annotate the Indic language portion of the Samanantar corpus using the above NER model. This synthetic labeled data is used to fine-tune for NER task in each of the languages respectively.

Wu et al. (2020) trained the student model to mimic the probability distribution of the entity labels by the teacher model. In our approach, we follow the *Hard Labels* approach where we round the probability distribution of the entity labels into a one-hot labeling vector to guide the learning of the student model.

5.4 Implementation Details

We use the Huggingface library[6] (Wolf et al., 2020) to train our NER models. We use NVIDIA A100 Tensor Core GPU to run all the experiments. We use `bert-base-multilingual-cased` (169.05M) as the base pre-trained model in all our experiments. We tune hyper-parameters based on F1-Score on the validation set. We use the following range of values for selecting the best hyper-parameter.

- Batch Size: 8, 16, 32
- Learning Rate: 1e-3, 1e-4, 1e-5, 1e-6, 3e-3, 3e-4, 3e-5, 3e-6, 5e-3, 5e-4, 5e-5, 5e-6

Once we obtain the best hyper-parameter, we fine-tune the model for 2 epochs with 5 different random seeds. We report the mean and standard deviation of the 5 runs.

Languages	Monolingual	Multilingual
bn	81.02 ± 0.40	80.74 ± 0.43
gu	80.59 ± 0.57	81.10 ± 0.39
hi	82.69 ± 0.45	82.93 ± 0.47
kn	80.33 ± 0.60	81.07 ± 0.55
ml	81.49 ± 0.15	81.13 ± 0.43
mr	81.37 ± 0.29	81.13 ± 0.47
pa	71.51 ± 0.59	71.81 ± 0.46
ta	73.36 ± 0.56	74.11 ± 0.46
te	82.49 ± 0.60	82.20 ± 0.31
as	45.37 ± 2.66	60.19 ± 4.80
or	25.01 ± 1.22	25.91 ± 0.40
Average	71.38 ± 0.69	72.94 ± 0.40

Table 6: Comparison of Monolingual *vs.* Multilingual Fine-tuning (F1 score). We report mean and standard deviation from 5 runs

6 Results

We now present the results from our experiments.

6.1 RQ1

We now answer the question if *the models trained using data from projection approach are better than*

[5]Davlan/bert-base-multilingual-cased-ner-hrl

[6]`https://github.com/huggingface/transformers/tree/main/examples/pytorch/token-classification`

	PER	LOC	ORG	Overall
bn	77.63	84.29	73.25	80.06
gu	81.14	88.65	67.63	80.83
hi	82.31	89.37	74.03	83.27
kn	78.16	87.29	73.12	81.28
ml	84.49	87.85	61.49	81.67
mr	83.70	88.66	66.33	81.88
pa	76.26	77.95	55.68	72.08
ta	76.01	83.09	58.73	74.48
te	84.38	84.77	70.92	81.90
as	75.00	54.55	57.14	62.50
or	41.78	21.40	13.39	26.42

Table 7: Entity-wise F1 score from the best multilingual model on Naamapadam Test Set.

cross-lingual zero-shot and teacher-student models?

Table 5 reports the results from our experiments. Apart from Hindi, Malayalam, and, Marathi we observe relatively poor results for other Indic languages in the Zero-Shot setting. Zero-shot techniques perform quite well in high-resource languages like Hindi, scoring a respectable 75.96%. However, for Assamese and Oriya languages the results are very poor. The Teacher-Student approach in comparison with the zero-shot approach gives very poor results.

We observe that the models trained using the Naamapadam-train dataset give the best F1 scores across languages. In general, we observe better performance from data obtained using Awesome-align (Dou and Neubig, 2021) compared to GIZA++ (Och and Ney, 2003). Moving forward, we choose the data obtained using Awesome-align (Dou and Neubig, 2021) in all our experiments.

IndicNER: Multilingual Fine-tuning

Multilingual fine-tuning on a downstream task has been shown to outperform language-specific fine-tuning in low-resource scenarios (Dhamecha et al., 2021). We also fine-tune a multilingual model on the combined data of all languages in Naamapadam-train. We refer to this model as IndicNER. Table 6 reports the results from our experiments. We observe that the multilingual model on average performs better than the monolingual models.

It can also be seen that for extremely low-resource languages like Assamese, the multilingual model performs a lot better than the others with a jump in F1 score from 45.37 to 60.19.

6.2 RQ2

In this section, we answer the question *if the models trained on the Naamapadam-train data fare better against models trained on other publicly available labeled datasets when tested on Naamapadam-test set?*

Table 8 reports the results of our experiments. We observe that the model fine-tuned on Naamapadam-train data outperforms all other models by a significant margin indicating the utility of our labeled data. Only the models trained using CFILT-HiNER (Murthy et al., 2022) and MahaNER (Litake et al., 2022) obtain reasonable F1 on Hindi and Marathi. This underlines the importance of large, high-quality data and shows that projection methods can help to create such data at scale.

6.3 Error Analysis

We observe that boundary error is our model's most common error type. The model sometimes identifies named entities partially. For example, in the case of Organization entities, our model only tags *A B C* as an organization entity when the correct entity phrase is, say, *A B C Limited*. Similarly, for Location entities, our model only tags *A B* as location entity when the correct entity phrase is *A B Hospital*. This could be attributed to some of the boundary errors present in our training data due to alignment errors.

7 Conclusion

We take a major step towards creating publicly available, open datasets and open-source models for named entity recognition in Indic languages. We introduce Naamapadam, the largest entity recognition corpora for 11 Indic languages containing more than $100K$ training sentences per language, and covering 11 of the 22 languages listed in the Indian constitution. Naamapadam also includes manually labelled test set for 9 Indic languages. We also build IndicNER, an mBERT based multilingual named entity recognition model for 11 Indic languages. We also provide baseline results on our test set along with a qualitative analysis of the model performance. The datasets and models will be available publicly under open-source licenses. We hope the dataset will spur innovations in entity recognition and its downstream applications in the Indian NLP space.

Language	Naamapadam	FIRE-2014	WikiANN	MultiCoNER	CFILT	MahaNER
bn	81.02 ± 0.40	35.68 ± 3.96	51.67 ± 1.24	26.12 ± 1.96	-	-
gu	80.59 ± 0.57	-	0.11 ± 0.12	-	-	-
hi	82.69 ± 0.45	47.23 ± 0.92	59.84 ± 1.25	41.85 ± 2.34	75.71 ± 0.67	-
kn	80.33 ± 0.60	-	2.73 ± 1.47	-	-	-
ml	81.49 ± 0.15	58.51 ± 1.13	62.59 ± 0.32	-	-	-
mr	81.37 ± 0.29	-	62.37 ± 1.12	-	58.41 ± 0.62	71.45 ± 1.44
pa	71.51 ± 0.59	-	0.7 ± 0.37	-	-	-
ta	73.36 ± 0.56	44.89 ± 0.94	49.15 ± 1.17	-	-	-
te	82.49 ± 0.60	-	49.28 ± 2.17	-	-	-

Table 8: Comparison of models trained on different datasets and evaluated on Naamapadam-test set (F1 score).

Acknowledgements

We would like to thank the Ministry of Electronics and Information Technology[7] of the Government of India for their generous grant through the Digital India Bhashini project[8]. We also thank the Centre for Development of Advanced Computing[9] for providing compute time on the Param Siddhi Supercomputer. We also thank Nilekani Philanthropies for their generous grant towards building datasets, models, tools and resources for Indic languages. We also thank Microsoft for their grant to support research on Indic languages. Most importantly, we would like to thank Archana Mhaske, Anil Mhaske, Sangeeta Rajagopal, Vindhya DS, Nitin Kedia, Yash Madhani, Kabir Ahuja, Shallu Rani, Armin Virk and Gowtham Ramesh for volunteering to annotate the testset.

Limitations

This work applies to languages that have a modest amount of data in parallel with English and are represented in pre-trained language models. These are typically high to mid-resource languages. Very low-resource languages might not have enough parallel corpora to extract sufficient NER training data. With limited parallel data and/or limited representation in pre-trained LMs, it will be difficult to get high-quality word alignments for projection. We use span-based annotation projection to alleviate word alignment errors to some extent.

Ethics Statement

The annotations are collected on a publicly available dataset and will be released publicly for future use. Some of these datasets originate from webcrawls and we do not make any explicit attempt to identify any biases in these datasets and use them as-is. All the datasets used have been cited. All the datasets created as part of this work will be released under a CC-0 license[10] and all the code and models will be release under an MIT license.[11]

The annotations in the testset were mostly contributed by volunteers interested in contributing to building a benchmark NER dataset. The volunteers were not made any payment and worked *pro bono*. Some annotators were paid for their services. These language experts were paid a competitive monthly salary to help with the task. The salary was determined based on the skill set and experience of the expert and adhered to the norms of the government of our country. The annotators were made aware that the annotations would be made publicly available. The annotations contains no personal information.

References

Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, and German Rigau. 2018. Building named entity recognition taggers via parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, and Orhan Firat. 2022. Data scaling laws in nmt: The effect of noise and architecture. In *International Conference on Machine Learning*, pages 1466–1482. PMLR.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D named entity annotation for German: Guidelines and dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).

[7] https://www.meity.gov.in/
[8] https://www.bhashini.gov.in/
[9] https://www.cdac.in/index.aspx?id=pune
[10] https://creativecommons.org/publicdomain/zero/1.0
[11] https://opensource.org/licenses/MIT

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, and Imre Solti. 2012. Building gold standard corpora for medical natural language processing tasks. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2012:144–153.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tejas Dhamecha, Rudra Murthy, Samarth Bharadwaj, Karthik Sankaranarayanan, and Pushpak Bhattacharyya. 2021. Role of Language Relatedness in Multilingual Fine-tuning of Language Models: A Case Study in Indo-Aryan Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8584–8595, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2022. Indicxtreme: A multi-task benchmark for evaluating indic languages.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

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, Hong Kong, China. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

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, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Amogh Mishra, Mitesh M. Khapra, and Pratyush Kumar. 2022. Indicnlg benchmark: Multilingual datasets for diverse nlg tasks in indic languages. In *EMNLP*.

Sobha Lalitha Devi, Pattabhi RK Rao, Malarkodi C.S, and R Vijay Sundar Ram. 2014. Indian Language NER Annotated FIRE 2014 Corpus (FIRE 2014 NER Corpus). In *In Named-Entity Recognition Indian Languages FIRE 2014 Evaluation Track*.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

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, Marseille, France. European Language Resources Association.

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Ruchi Khapra, Vivek Seshadri, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Aksharantar: Towards building open transliteration tools for the next billion users.

Shervin Malmasi, Besnik Fetahu, Anjie Fang, Sudipta Kar, and Oleg Rokhlenko. 2022. SemEval-2022 task 11: Multiconer multilingual complex named entity recognition. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Online. Association for Computational Linguistics.

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.

Rudra Murthy, Pallab Bhattacharjee, Rahul Sharnagat, Jyotsana Khatri, Diptesh Kanojia, and Pushpak Bhattacharyya. 2022. Hiner: A large hindi named entity recognition dataset.

Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2018. Judicious selection of training data in assisting language for multilingual neural NER. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–406, Melbourne, Australia. Association for Computational Linguistics.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 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, Marseille, France. European Language Resources Association.

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, Florence, Italy. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R. Curran. 2019. NNE: A dataset for nested named entity recognition in English newswire. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5176–5181, Florence, Italy. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xavier Schmitt, Sylvain Kubler, Jérémy Robert, Mike Papadakis, and Yves LeTraon. 2019. A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 338–343. IEEE.

Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. 2010. Synergy: a named entity recognition system for resource-scarce languages such as swahili using online machine translation. In *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, pages 21–26.

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)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003a. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003b. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021.

Automated concatenation of embeddings for structured prediction. 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 2643–2660, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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, Online. Association for Computational Linguistics.

Qianhui Wu, Zijia Lin, Börje Karlsson, Jian-Guang Lou, and Biqing Huang. 2020. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6505–6514, Online. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

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

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

David Yarowsky and Grace Ngai. 2001a. 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*.

David Yarowsky and Grace Ngai. 2001b. Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection Across Aligned Corpora. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Dongxu Zhang, Boliang Zhang, Xiaoman Pan, Xiaocheng Feng, Heng Ji, and Weiran Xu. 2016. Bitext name tagging for cross-lingual entity annotation projection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 461–470, Osaka, Japan. The COLING 2016 Organizing Committee.

A Issues with WikiAnn

On manual inspection, the sentences in the Wiki dataset had a lot of issues. The “sentences" were mostly just phrases and titles where more often than not, the entire thing would be considered a named entity. Such a skewed dataset can heavily influence the quality of a model trained on it. A few examples depicting the above issues are shown below.

- दमन और दीव
  B-LOC I-LOC I-LOC
- लोकमान्य तिलक टर्मिनस रेलवे स्टेशन
  B-ORG I-ORG I-ORG I-ORG I-ORG
- लाल बहादुर शास्त्री स्टेडियम
  B-ORG I-ORG I-ORG I-ORG
- सवाई मान सिंह स्टेडियम
  B-LOC I-LOC I-LOC I-LOC

B Comparison English NER taggers

We compared many English NER taggers. The results are shown in Table 9.

Model	Reference	F1
Spacy NER	Schmitt et al. (2019)	91.60
LUKE	Yamada et al. (2020)	94.30
BERT-base-NER	Devlin et al. (2019)	91.30

Table 9: F1 scores for various off-the-shelf English models on CoNLL-2003 testset

C Examples of alignments generated by Awesome-align

We now present a few examples from our projection method. Figure 4 presents examples of correct alignments and hence correct projections of NER tags. As can be seen, the alignment is fairly sparse and the model aligns only those words in which it

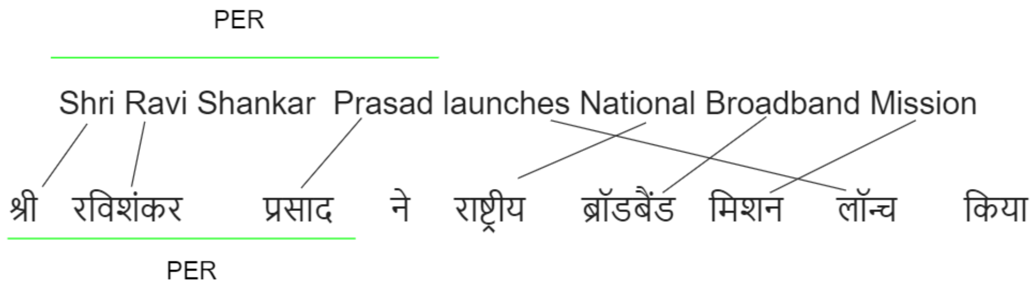

Figure 4: Correct Projection and Alignment using Awesome Align

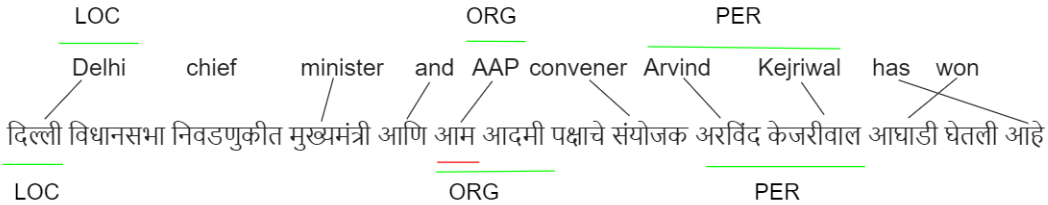

Figure 5: Incorrect Projection and Alignment using Awesome Align

is extremely confident. In this sentence, both words “Ravi" and “Shankar" had to be aligned to “Ravishankar" in Hindi, but only “Ravi" was aligned. But due to our range projection, the entire entity “Shri Ravi Shankar Prasad" was projected successfully with the tag PERSON.

Figure 5 shows an example of incorrect word alignment using the awesome align method for word alignment. In this sentence, “AAP" which is the abbreviated name of a political party is mapped only to “Aam" in Marathi instead of the entire phrase “Aam Aadmi pakshanche". This causes the projected entity to be only partially tagged with the entity type Organization.

D Comparison with Other Pre-Trained Language Models

Table 10 reports the performance of various pre-trained models fine-tuned on Naamapadam-train set in a multilingual fashion. We observe both MuRIL (Khanuja et al., 2021) and IndicBERT (Doddapaneni et al., 2022) outperform mBERT model.

Langs.	mBERT	MuRIL	IndicBERT
bn	80.74 ± 0.43	80.97 ± 0.28	81.02 ± 0.53
gu	81.10 ± 0.39	80.08 ± 0.27	80.34 ± 0.20
hi	82.93 ± 0.47	82.25 ± 0.28	82.40 ± 0.11
kn	81.07 ± 0.55	80.38 ± 0.38	80.74 ± 0.43
ml	81.13 ± 0.43	80.53 ± 0.44	80.45 ± 0.44
mr	81.13 ± 0.47	80.16 ± 0.28	80.52 ± 0.29
pa	71.81 ± 0.46	72.01 ± 0.26	71.66 ± 0.25
ta	74.11 ± 0.46	74.90 ± 3.87	74.85 ± 2.74
te	82.20 ± 0.31	81.83 ± 0.29	82.33 ± 0.50
as	60.19 ± 4.80	66.03 ± 3.30	66.65 ± 3.73
or	25.91 ± 0.40	39.29 ± 0.60	39.44 ± 0.65
Average	72.94 ± 0.40	74.40 ± 0.34	74.58 ± 0.55

Table 10: Comparison of various pre-trained models fine-tuned on Naamapadam-train in a multilingual fashion.

	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te
Un-filtered	141K	8.6M	3M	10M	4M	5.9M	3.6M	998K	2.9M	5.2M	4.9M
Filtered	10K	966K	475K	999K	474K	720K	457K	197K	465K	500K	510K

Table 11: Filtering Statistics (Number of Sentences)

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
LimitationSection discusses the limitation of our work

☑ A2. Did you discuss any potential risks of your work?
Ethics Section discusses the potential risks of our work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3 talks about creating the dataset

☑ B1. Did you cite the creators of artifacts you used?
Yes, Section 3 and Section 5.1 cites the creators of artifacts used

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Ethics Section describes the license

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Ethics Section

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Some of these datasets originate from web crawls and we do not make any explicit attempt to identify any biases in these datasets and use them as-is. The huge volume of data and lack of tools available to anonymize/ remove biases in the languages we are dealing with make it difficult to anonymize identities or remove offensive content

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We have listed the languages used. It is discussed in section 3.6

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3.6

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ **Did you run computational experiments?**

Section 5.4 provides details about the computation experiments performed

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 5.4

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 6

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5.4

D ☑ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

Section 4

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We follow conll 2003 annotation guidelines and have placed reference to the same in section

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Ethics Section talks about the same

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Ethics Section

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics section describes the same

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The annotators are native speakers from the Indian subcontinent. We mention the same in Ethics section