GUIT-AsTourNE: A Dataset of Assamese Named Entities in the Tourism Domain

Bhargab Choudhury*, **Vaskar Deka**, **Shikhar Kumar Sarma**
Department of Information Technology
Gauhati University
Guwahati-781014 Assam India
bhargabchoudhury24@gmail.com
{vaskardeka, sks}@gauhati.ac.in

Abstract

Named Entity Recognition is a fundamental task in Natural Language Processing that involves classifying text into predefined classes such as person, location, organisation etc. Annotated data for the Named Entity Recognition task is lacking for Indian languages, including Assamese, whereas English and European languages have plenty of data. In this paper, we presented a manually annotated Assamese Named Entity dataset on the tourism domain. The dataset contains 7166 sentences and 94604 tokens. The resulting dataset contains 9151 named entities tagged into eight Named Entity classes: location, organisation, person, entertainment, facilities, year, date and miscellaneous. Also, we trained and evaluated transformer-based language models like mBERT, XLM-RoBERTa, IndicBERT, and MuRIL on our dataset. The XLM-RoBERTa model outperforms all others with an F1 score of 78.51%.

1 Introduction

Named Entity Recognition (NER) is a Natural Language Processing (NLP) task used to detect and classify tokens into some predefined classes. The term Named Entity (NE) was introduced in the sixth Message Understanding Conference (MUC) (Grishman and Sundheim, 1996). Phrases containing the names of people, places, and organisations are known as NE (Sang and De Meulder, 2003). More generally, NE is a real-world object that can be denoted as a proper noun, but it is not limited to this. NER plays an important role in many NLP applications such as text understanding (Zhang et al., 2019), information retrieval (Guo et al., 2009), question answering (Mollá et al., 2006), machine translation (Babych and Hartley, 2003), relation extraction (RE), knowledge graph construction (Kejriwal, 2022) etc. The recognition of NE can be attained through four methods: rule-based, unsupervised learning, feature-based supervised learning, and deep learning-based approaches (Li et al., 2020). Deep learning (DL) has gained a lot of attention recently because of its success in a variety of fields. A significant number of studies have used DL to improve NER over the last few years, progressively raising the bar for performance. In order to train a supervised deep learning-based NER system, a substantial quantity of annotated data is essential. The quantity and quality of data determine how well DL based models perform. In the context of NER datasets and tools, Assamese is regarded as a low-resource language. In contrast to languages such as English or European languages, there is a notable lack of publicly accessible NER datasets for Assamese.

The official language of Assam, a northeastern state of India, is Assamese (অসমীয়া, *asomiya*). Assamese is spoken by the native inhabitants of the state. The language is known for its highly inflected forms and the utilisation of pronouns and noun plural markers in both honorific and non-honorific constructions.

There are some difficulties in creating the Assamese NE dataset. The following are a few challenges.

No Capitalisation: Unlike English language Assamese does not follow capitalisation, a feature that would have been useful for completing the NER task. Example: ৰাম গুৱাহাটীলৈ গৈছে (*Ram Guwahatiloi goise*, Ram has gone to Guwahati). In this sentence, there is no distinguish between proper nouns or the beginning of the sentence, maintaining a uniform script

* Corresponding author

throughout.

NE Ambiguity: In Assamese, proper nouns can be confusing as the same word might fall under more than one POS categories. Example: The word আকাশ (*Akash*) can be the name of a person, or it refers to the sky.

Language Complexity: Assamese is a morphologically complex, inflectional language. This means that words can take different forms depending on their grammatical role in a phrase. Example: ঘৰ (*ghor*), meaning "house", can be inflected to ঘৰৰ (*ghoror*), meaning "of the house", and ঘৰত (*ghorot*), meaning "in the house".

Free Word Order: Assamese language with a flexible word order presents a greater challenge for the NER problem as precise word order patterns cannot be implemented in combination with computational techniques. Example: The sentences মই মাছ খাওঁ (*moi maas khaon*, I eat fish) and মাছ মই খাওঁ (*maas moi khaon*, I eat fish) have different arrangements of words; however, their core meanings remain the same.

In this paper, we present an Assamese NE dataset, namely GUIT-AsTourNE, which consists of 94604 tokens classified into eight NE classes. This is the first Assamese NE dataset in the tourism domain. Also, we present the results of different transformed-based models trained on the GUIT-AsTourNE dataset. The followings are the summary of our contribution:

- We gather textual information in Assamese on the tourism domain. The text data is annotated into eight NE classes.
- Then we perform the blind validation by two validators. We evaluate the agreement between annotator and validators.
- We resolve the conflicts through the intervention of a linguist.
- We train and evaluate transformer-based models such as mBERT, XLM-RoBERTa, IndicBERT, and MuRIL on our dataset.
- We release[1] our data and the best-performing model.

[1] https://github.com/nlp30/GUIT-AsTourNE

2 Related Work

Research and development for most of the NLP tasks for the Assamese language are still in their early stages compared to languages with abundant linguistic resources. Significant studies have been conducted in Word embedding (Pathak et al., 2024), POS tagging (Saharia et al., 2009; Pathak et al., 2022b, 2023; Baishya and Baruah, 2024), UPoS tagging (Talukdar et al., 2024; Talukdar and Sarma, 2023), and WordNet (Sarma et al., 2010; Sarmah et al., 2019; Phukon et al., 2021). Also, a few NER works on the Assamese language have been documented(Sharma et al., 2012; Talukdar et al., 2014; Sharma et al., 2014; Mahanta et al., 2016; Sharma et al., 2016; Talukdar et al., 2018). WikiAnn(Pan et al., 2017) is the first publicly available dataset on Assamese language and 282 global languages. The AsNER(Pathak et al., 2022a) dataset, available only in the Assamese language, contains 34K entities. However, around 29K entities are without sentence context(Mhaske et al., 2022). The Naamapadam(Mhaske et al., 2022) dataset, which covers 11 Indian languages, including Assamese, contains 5K entities. Table 1 lists the statistics of publicly available Assamese NER datasets.

3 Corpus Acquisition and Pre-processing

In this section, we outline the process of obtaining and preparing the corpus. We explain the source from which the corpus was developed, and then we describe the preprocessing techniques used to clean and prepare the raw data for the annotation process.

3.1 Source of Corpus

The first step towards annotated data is to collect text on the tourism domain. Using a crawler, we extract text from Wikipedia on the tourism domain. The laboratory-developed GUIT tourism corpus is an additional source. Table 2 displays the statistics for the corpus.

3.2 Preprocessing

Preprocessing is an important step in generating high-quality data. Other language terminology, extraneous characters, gaps, typos, etc. are all present in the data. Therefore, in

Dataset	#Sentence	#Tokens	#NE
WikiAnn	300	1516	329
AsNER	24040	98623	34963
Naamapadam	10369	112048	5045

Table 1: Statistics from the current datasets.

Source	#Sentence	#Tokens
Wikipedia	3693	54246
GUIT	3473	40358
Total	**7166**	**94604**

Table 2: Statistic of the two sources.

order to obtain real vocabulary, data cleaning is essential.

Removing Noisy Characters: White spaces are used in place of punctuation markers such as quotation marks, periods, ellipses, and special characters. Unwanted noisy characters, extra spaces and the HTML tag are eliminated.

Language Normalisation: The text might contain elements in other languages. These words are translated into the Assamese. The translation of some words is not available; those words are transliterated.

4 Annotation Process

In this section, we describe how the dataset is created. We discuss the background of NE classes, the NE classes that were considered and the annotation methodology. We evaluate the Inter-Annotator agreement (IAA) to measure the consistency between the annotator and validators. Finally, we resolve the annotation conflicts with the help of a linguistic expert.

4.1 NE Classes

Selecting the NE classes is the first step towards creating the NER dataset. NE classes specify the categories into which various text elements can be classified. The first NE classes defined on MUC 6[2] are organisation, person, location, date, time, money, and percent. In 2000, artefact NE class was introduced as part of the IREX project (Sekine and Isahara, 2000), a Japanese language evaluation effort. In the CoNLL-2003 shared task: language-independent Named Entity Recognition (Sang and De Meulder, 2003) defined four types of classes: persons, organisations, locations, and miscellaneous. In the ACE[3] programme, seven NE classes were defined: person, organisation, location, facility, weapon, vehicle, and geo-political. The dataset AnCora[4] (Taulé et al., 2008) consists of two corpora, one in Catalan and the other in Spanish, categorised tokens into six NE classes. The multilingual dataset OntoNotes 5.0 (Weischedel et al., 2011) contains 18 NE classes. The NoSta-D (Benikova et al., 2014) entity annotation guideline defines four primary classes: person, organisation, location, and other. Five entity classes were defined in the Rich ERE (Song et al., 2015) guidelines. The RuNNE Shared Task (Artemova et al., 2022) in Russian was concerned with nested NE, and the dataset it utilises NEREL contains 29 NE classes. In WojoodNER-2023 (Jarrar et al., 2023), the first Arabic NER Shared Task, 21 NE classes were defined. The NE classes developed at the AU-KBC Research Centre[5] (Rao et al., 2015) are hierarchical classes with three major classes: name, time, and numerical expressions. This NE classification is standardised by the Ministry of Communications and Information Technology, Government of India. It is used for Cross-Lingual Information Access (CLIA) and Indian Language - Indian Language Machine Translation (IL-IL MT) consortium projects. Named entities include people, organisations, locations, facilities, cuisines, locomotives, artefacts, entertainment, organisms, plants, and diseases. Distance, money, quantity, and count are the four different types of numerical expressions. Time expressions include year, month, date, day, period, and special day. In FIRE 2018 (HB et al.,

[2] https://cs.nyu.edu/~grishman/muc6.html

[3] https://www.ldc.upenn.edu/collaborations/past-projects/ace

[4] http://clic.ub.edu/corpus/en/ancora

[5] https://au-kbc.org/

2018), the Information Extractor for Conversational Systems in Indian Languages (IECSIL) track introduced a taxonomy of nine entity types for Hindi, Tamil, Malayalam, Telugu, and Kannada. The entity types are date, event, location, name, number, occupation, organisation, other, and things. In the outlook of tourism domain, Zahra et al., Hidayatullah et al., and Fudholi et al. classified text into three NE classes: natural, heritage, and purpose. The NE classes: nature, place, city, region, and negative for tourism domain were defined by Saputro et al.. A summary of the NE classes is shown in Table 3.

Based on our analysis and the current NE classes, we have identified the following NE classes as relevant for tourism text: location, organisation, person, entertainment, facilities, year, date, and miscellaneous. Seven of these classes are a subset of the NE classes developed by the AU-KBC Research Centre. In addition, we have considered the miscellaneous class to tag tokens that are NE but do not fit into any of the defined NE classes. We have briefed about the NE classes along with examples transliterated from Assamese to English.

LOCATION (LOC): Villages, towns, cities, road, provinces, countries, bridges, ports, dams, hills, mountains, water bodies, valleys, gardens, beaches, national parks, landscapes, parks, clubs, monuments, religious places, museum etc. Examples: মাজুলী (*Majuli*), কমলাবাৰী সত্ৰ (*Kamalabari Satra*), দীঘলীপুখুৰী (*Dighalipukhuri*).

ORGANISATION (ORG): Government, government agencies, public organisations, companies, non-profit organisations, trust, educational institute etc. Examples: তিৱা স্বায়ত্বশাসিত পৰিষদ (*Tiwa Swayatwashasit Parishad* , Tiwa Autonomous Council), অসম ক্ষুদ্ৰ উদ্যোগ উন্নয়ন নিগম (*Asom Khudra Udyog Unnayan Nigam*, Assam Small Industries Development Corporation), কটন বিশ্ববিদ্যালয় (*Cotton Bishwavidyalaya*, Cotton University).

PERSON (PER): First name, middle name, last name, historical figure, fictional character etc. Examples: লাচিত বৰফুকন (*Lachit Borphukan*), শংকৰদেৱ (*Sankardev*), পদ্মনাথ গোহাঞি বৰুৱা (*Padmanath Gohain Baruah*).

ENTERTAINMENT (ENT): Cultural festival, dance, music, drama, traditional performances, exhibitions, sporting event, boat race, religious ceremonies and festival etc. Examples: সত্ৰীয়া নৃত্য (*Sattriya Nritya*), অৰণ্যত গধুলি (*Aranyat Godhuli*), অম্বুবাচী মেলা (*Ambubachi Mela*).

FACILITIES (FAC): Hotel, restaurant, guest house, hospital, police station, bus terminal or station, railway station, airport etc. Examples: অৰণ্য অতিথিশালা (*Aranya Atithishala*, Aranya Guesthouse), কহুৱা ৰিজৰ্ট (*Kahuwa Resort*), লীলাবাৰী বিমানবন্দৰ (*Lilabari Bimanbandar*, Lilabari Airport).

YEAR (YEAR): Expressions that represent year. Examples: ১৯৯০ (1909), ১৯২১-১৯২২ (1921-1922).

DATE (DATE): Expressions that represent date. Examples: ১ এপ্ৰিল (1 April), ২৪/১/১৯৯০ (24/1/1990).

MISCELLANEOUS (MISC): This category is used to tag entities like political ideologies, book names, nationalities, products, languages etc., that do not fit neatly into other classes. Examples: ভাৰতীয় (*Bharatiya*, Indian), আহোম (*Ahom*), কালিকা পুৰাণ (*Kalika Puran*).

Corpus/Paper	Year	#Class
MUC 6	1995	7
IREX	2000	8
CoNLL-2003	2003	4
ACE	2000-2008	7
AnCora	2008	6
OntoNotes	2008	18
NoSta-D	2014	4
Rich ERE	2015	5
AU-KBC	2015	21
Saputro et al.	2016	5
FIRE	2018	9
NEREL	2021	29
Zahra et al.	2022	3
Hidayatullah et al.	2022	3
WojoodNER	2023	21
Fudholi et al.	2023	3

Table 3: Summary of NE Class.

4.2 Annotation Methodology

We selected one annotator for annotation. The annotator is a native speaker with a Bachelor's Degree in Assamese. We use the IOB2 tagging format, where the I tag denotes the inside of a NE chunk (excluding the beginning), the B tag marks the beginning of a NE chunk, and the O tag is used when a word is outside of any NE. Annotation guidelines were prepared and explained to the annotator. After tagging the initial 100 sentences, a linguist reviewed the tags to identify any problems or inconsistencies in the guidelines. This feedback was then used to enhance the guidelines. Following these guidelines, the annotator carried out the annotation. After completing the annotation, two validators were engaged to cross-check the annotations. The two validators independently perform the validation. In cases where the validators disagreed on an annotation, they added a new annotation.

4.3 Inter Annotator Agreement

Inter Annotator Agreement (IAA) score assess how consistently different annotators label the same text for named entities. Cohen's Kappa (κ) and F1 Score are commonly used metrics for calculating IAA. But, Cohen's Kappa is not an appropriate metric for NER (Hripcsak and Rothschild, 2005; Grouin et al., 2011). In NER, a considerable amount of the data may be classified as O (not NE). This can inflate the κ score, indicating a high level of agreement, which is not actual agreement. So, we calculate the macro-averaged F1 score as an alternative to Cohen's Kappa. The arithmetic mean of the F1 score of all the NE classes is calculated to get an overall measure of agreement. However, we calculate Cohen's Kappa for tokens that have atleast one annotation. Table 4 displays the F1 score and Cohen's Kappa between the Annotator and Validators, revealing substantial agreement among them.

	***F*1**	$\boldsymbol{\kappa^a}$	$\boldsymbol{\kappa^b}$
Annotator vs Validator-1	0.94	0.89	0.95
Annotator vs Validator-2	0.89	0.81	0.91
Average	0.92	0.85	0.93

Table 4: Calculated F1 score and Cohen's Kappa values between annotator and validators. [a] on annotated tokens, [b] on all tokens

4.4 Conflict Resolution

Only one annotator annotated the data, so it's important to ensure that the dataset's quality is not compromised. Despite a substantial agreement between the annotator and the validator, we identified conflicts in 2737 tokens. Resolving these conflicts is essential to ensure the reliability of the NER system. Table 5 shows the agreement and disagreement between the annotator and the validators. Out of 94604 tokens, the annotator and validators agreed on 91867 tokens, which is approximately 97%.The validators did not agree on 603 tokens with the annotator, but both the validators assigned the same NE tag. For these 603 tokens, we use the NE tag assigned by the validators. For the remaining 2134 tokens, where either one of the validators or both did not agree with the annotator, we seek the opinion of a linguistic expert. Two such cases are explained in Table 6.

4.5 Dataset Statistics and Format

The annotated dataset is prepared in column format; the first column represents the words, and the second column represents corresponding NE tag. A blank line separates two sentences in the dataset. A total of 9151 ($\approx$ 9.67%) tokens were reported as NE. Table 7 list the frequency distribution of the various classes.

5 Experiments

In this section, we discuss the fine-tuning of various transformer-based models like mBERT, XLM-RoBERTa, IndicBERT and MuRIL on our dataset. We plot the confusion matrix of the best-performing model and also evaluate the model performance using the *nervaluate*[6] package.

5.1 Model

mBERT: mBERT (Multilingual BERT) (Devlin et al., 2019) is a pre-trained language model designed to comprehend and analyse text in multiple languages. It is a variation of the popular BERT model that has been trained on an extensive dataset containing 104 languages including Assamese. mBERT can be fine-tuned using labelled data from

[6]https://pypi.org/project/nervaluate/

Validator-1	Validator-2	#Tokens	Remarks
Agree	Agree	91867	–
Disagree	Disagree	603	Both the validators assign same NE Tag
Agree	Disagree	1611	–
Disagree	Agree	452	–
Disagree	Disagree	71	Both the validators assign different NE Tag

Table 5: Statistics of validators agreement and disagreement.

Sentence	Conflict & Resolution
মই তাত এটা **অসমীয়া** পৰিয়াল লগ পাইছিলোঁ । *moi tat eta* ***asomiya*** *poriyal log paisilu* I met an **Assamese** family there.	**Conflict:** In this case, the disagreement arises for the word অসমীয়া (*asomiya*). One annotator tagged it as B-LOC, while a validator classified it as O, and another validator identified it as B-MISC. **Resolution:** The word অসমীয়া (*asomiya*, Assamese) is derived from the word অসম (*Asom*, Assam) (a location), which undergoes a morphological transformation to convey a different meaning, such as Assamese language or people. In this context, it refers to the Assamese people, and the linguist categorised it as B-MISC.
এইক্ষেত্ৰত **কেৰালাও** এখন উল্লেখযোগ্য ঠাই । *eikhetrat* ***Keralao*** *ekhon ullekhjogya thaai* **Kerala** is also an important place in this regard.	**Conflict:** In this sentence, the conflict arises for the word কেৰালাও (Keralao). The annotator categorised it as an O, while one validator tagged it as B-MISC and another as B-LOC. **Resolution:** The root word for কেৰালাও (*Keralao*) is কেৰালা (*Kerala*), which denotes a LOCATION NE. The suffix ও (*o*) is added to কেৰালা (*Kerala*). In this context, the addition of the suffix does not alter the NE category of the word. Consequently, the linguist classified it as B-LOC.

Table 6: Examples of Sentences depicting conflict and resolution for final tagging.

any language within its multilingual training corpus.

XLM-RoBERTa: XLM-RoBERTa(Conneau et al., 2020) is an enhanced iteration of XLM that builds upon RoBERTa architecture. It is pre-training on 2.5TB of data in 100 languages. XLM-RoBERTa inherits the cross-lingual capabilities of XLM while benefiting from the improved representation learning of RoBERTa.

IndicBERT: IndicBERT(Kakwani et al., 2020) is a multilingual language model specifically designed for processing 12 major Indian languages including Assamese. It makes use of the more effective ALBERT (Lan et al., 2019) architecture, which is also a variation of BERT model.

MuRIL: Another important model in the multilingual landscape is MuRIL (Multilingual Representations for Indian Languages)(Khanuja et al., 2021), which was created especially for processing 16 Indian languages and English. It makes use of a transformer-based architecture that is comparable to but distinct from BERT.

5.2 Implementation Details

We split our dataset into training (70%), development (15%), and testing (15%) sets, as shown in Table 8. When splitting the data, we ensure a balanced stratified distribution of tags across all sets, as presented in Table 9.

We use *bert-base-multilingual-cased* variation for mBERT, *xlm-roberta-base* for XLM-

NE Tag	Frequency	%Frequency
LOC	5164	56.43
ORG	382	4.17
PER	1941	21.21
ENT	238	2.60
FAC	159	1.74
YEAR	454	4.96
DATE	88	0.96
MISC	725	7.92
Total	**9151**	-

Table 7: Frequency distribution of the different classes

	#Sentences	#Tokens
Train	4988	66184
Dev	1073	14215
Test	1105	14205

Table 8: Count of sentences and tokens in the train, dev and test splits for the GUIT-AsTourNE dataset.

NE	Train	Dev	Test
LOC	3615	774	775
ORG	268	57	57
PER	1358	293	290
ENT	166	36	36
FAC	113	23	23
YEAR	318	68	68
DATE	62	13	13
MISC	509	108	108

Table 9: Count of NE classes for train, dev and test splits for the GUIT-AsTourNE dataset.

RoBERTa and *muril-base-cased* for MuRIL. To train NER model, we use the Huggingface Trainer API. We employed Weighted Cross Entropy Loss function during the training phase, which is particularly effective for dealing with imbalanced datasets by assigning more significance to underrepresented classes. This is achieved by integrating class weights into the loss function, ensuring more balanced learning and improving the model's ability to generalise across all classes. Additionally, we used AdamW as an optimiser with a linear learning rate scheduler. For each training, we used the same set of hyperparameters. The experiments were conducted for 20 epochs with a batch size of 16 and a learning rate of 1e-5.

5.3 Results

In Tables 10 and 11, we provide the performance results for mBERT, XLM-RoBERTa, IndicBERT, and MuRIL on our dataset. XLM-RoBERTa achieved the highest F1 score of 78.51%, followed by MuRIL and mBERT with an F1 score of 77.79% and 77.70% respectively. IndicBERT has the lowest performance, with an F1 score of 28.89%. XLM-RoBERTa performed very well in identifying the entity YEAR, achieving an outstanding F1 score of 91.69%, but showed lower performance for the entity FAC, with an F1 score of 45.26%. Figure 1 represents the confusion matrix of the XLM-RoBERTa model. Errors have been observed in tagging a NE as not being a NE, except for the tags YEAR and DATE. The maximum errors are observed for the tag B-FAC. Additionally, mislabeling of B-FAC as B-LOC, I-ENT as I-LOC and I-PER, and I-MISC as I-LOC has been noted. A more detailed analysis of the model is conducted using the nervaluate package. Table 12 provides additional details for the evaluation schema, which are Strict, Exact, and Partial for all NE tags. According to the Strict evaluation method, a model prediction is considered correct only when the predicted entity label and the predicted entity string match the ground truth exactly; otherwise, it is considered incorrect. The Exact evaluation schema focuses solely on the accuracy of the predicted entity string boundaries, disregarding the entity type. The Partial evaluation schema combines aspects of the Strict and Exact evaluation. Unlike the Strict and Exact, the Partial method considers partial matches as incorrect.

Model	**P**(%)	**R**(%)	**F1**(%)
mBERT	72.35	83.89	77.70
XLM-RoBERTa	72.55	85.53	78.51
IndicBERT	23.48	37.56	28.89
MuRIL	72.55	83.83	77.79

Table 10: F1 score, precision (P), and recall (R) of various models on GUIT-AsTourNE dataset.

	mBERT	XLM-RoBERTa	IndicBERT	MuRIL
LOC	82.03	83.45	24.78	82.71
ORG	66.42	71.90	9.51	67.54
PER	75.37	76.82	18.56	73.98
ENT	66.85	67.52	8.19	64.77
FAC	55.55	45.26	5.7	54.88
YEAR	94.67	91.69	60.82	94.94
DATE	70.96	53.65	4.98	64.70
MISC	60.44	58.15	8.93	60.57

Table 11: The NE class wise F1(%) score of various models on GUIT-AsTourNE dataset.

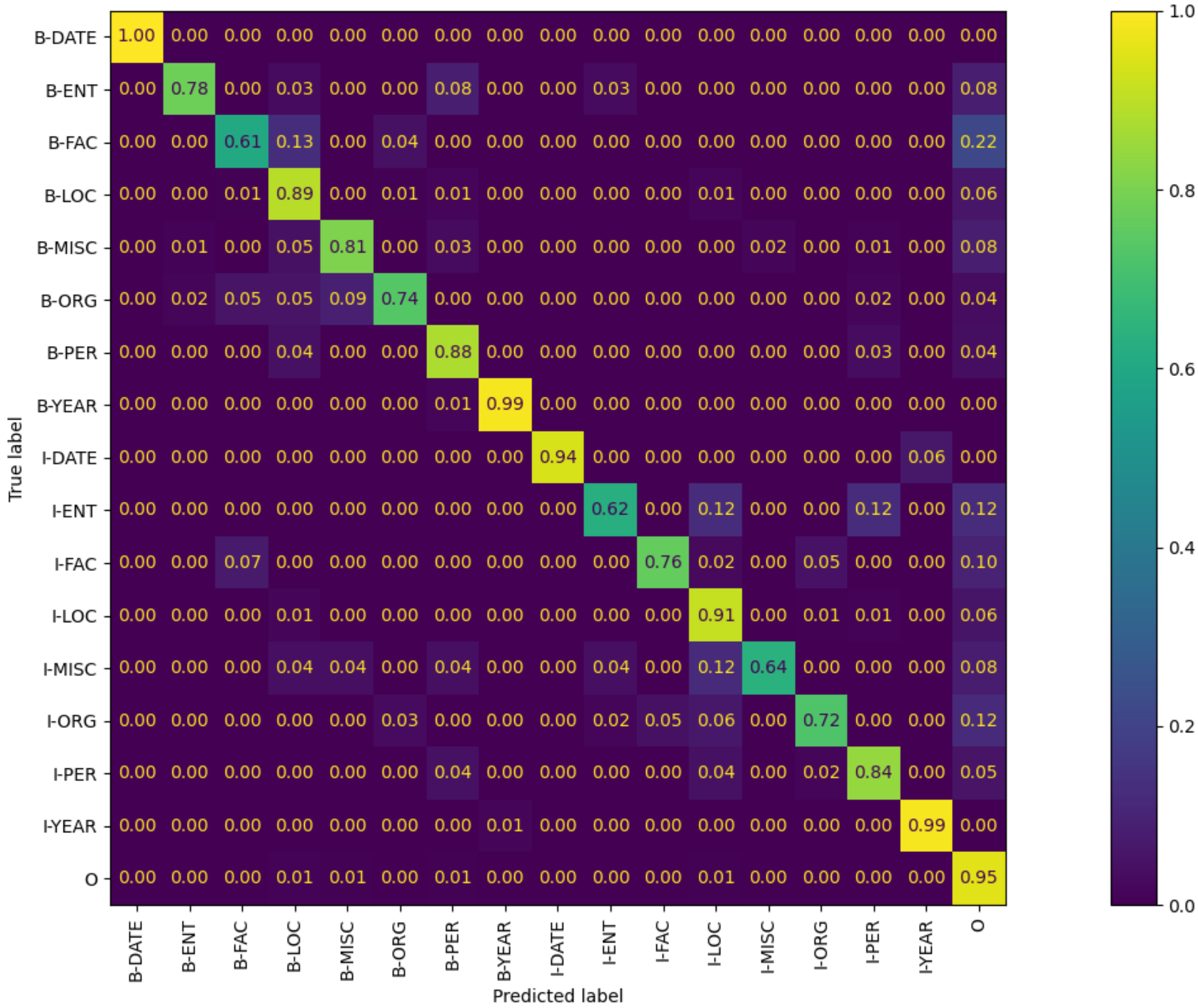

Figure 1: Confusion Matrix for XLM-RoBERTa on GUIT-AsTourNE dataset

6 Conclusion

In this paper, we present a new NE dataset, GUIT-AsTourNE, for Assamese in the tourism domain, annotated into eight NE classes. We discuss the NE class, annotation guidelines, and annotation process in detail. We analyse the annotation quality by calculating the IAA between the annotator and validators. First, the annotation is performed by one annotator. Then, we validate the annotation by two validators. After that, we find the conflicted token between the annotator and the validators. We seek the help of a linguist to resolve these conflicted tokens. The final dataset contains 7166 sentences, 94604 tokens and 9151 entities. We fine-tuned transformer-based lan-

Evaluation Scheme	NE Class	Error Type					F1 (%)
		Correct	Incorrect	Partial	Missing	Spurious	
Strict	LOC	636	86	0	53	142	77.60
	ORG	37	19	0	1	12	59.20
	PER	238	42	0	11	87	72.34
	ENT	23	10	0	3	13	56.09
	FAC	13	6	0	4	7	53.06
	YEAR	59	9	0	0	7	82.51
	DATE	7	6	0	0	5	45.16
	MISC	77	25	0	6	88	51.67
Exact	LOC	651	71	0	53	142	79.43
	ORG	42	14	0	1	12	67.19
	PER	250	30	0	11	87	75.98
	ENT	26	7	0	3	13	63.41
	FAC	15	4	0	7	23	61.22
	YEAR	59	9	0	0	7	82.51
	DATE	7	6	0	0	5	45.16
	MISC	85	17	0	6	88	57.04
Partial	LOC	651	0	71	53	142	83.77
	ORG	42	0	14	1	12	78.39
	PER	250	0	30	11	87	80.54
	ENT	26	0	7	3	13	71.95
	FAC	15	0	4	4	7	69.38
	YEAR	59	0	9	0	7	88.81
	DATE	7	0	6	0	5	64.51
	MISC	77	25	0	6	88	62.75

Table 12: Evaluation result of XLM-RoBERTa on GUIT-AsTourNE datset

guage models like mBERT, XLM-RoBERTa, IndicBERT, and MuRIL. For this, we split our data into train, dev and test and performed the experiments by keeping the same hyper-parameter for all the experiments. We observed the highest F1 score of 78.51% on XLM-RoBERTa. Also, the performance of mBERT and MuRIL is almost similar. In the future, we plan to extend this dataset to other NLP tasks like relation extraction.

Acknowledgements

The Linguistic works including Validations have been carried out in the Centre for R&D in Digital Enablement of Local Languages, Department of Information Technology, Gauhati University.

References

Ekaterina Artemova, Maxim Zmeev, Natalia Loukachevitch, Igor Rozhkov, Tatiana Batura, Vladimir Ivanov, and Elena Tutubalina. 2022. Runne-2022 shared task: Recognizing nested named entities. *arXiv preprint arXiv:2205.11159*.

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

Diganta Baishya and Rupam Baruah. 2024. Part-of-speech tagging for low resource languages: Activation function for deep learning network to work with minimal training data. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. Nosta-d named entity annotation for german: Guidelines and dataset. In *LREC*, pages 2524–2531.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

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.

Dhomas Hatta Fudholi, Annisa Zahra, Septia Rani, Sheila Nurul Huda, Irving Vitra Paputungan, and Zainudin Zukhri. 2023. Bert-based tourism named entity recognition: making use of social media for travel recommendations. *PeerJ Computer Science*, 9:e1731.

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*.

Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard. 2011. Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview. In *Proceedings of the 5th linguistic annotation workshop*, pages 92–100.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274.

Barathi Ganesh HB, Soman KP, Reshma U, Mandar Kale, Prachi Mankame, Gouri Kulkarni, and Anitha Kale. 2018. Information extraction for conversational systems in indian languages-arnekt iecsil. In *Proceedings of the 10th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 18–20.

Ahmad Fathan Hidayatullah, Rosyzie Anna Apong, Daphne Teck Ching Lai, and Atika Qazi. 2022. Extracting tourist attraction entities from text using conditional random fields. In *2022 IEEE 7th International Conference on Information Technology and Digital Applications (ICITDA)*, pages 1–6. IEEE.

George Hripcsak and Adam S Rothschild. 2005. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298.

Mustafa Jarrar, Muhammad Abdul Mageed, Mohammed Khalilia, Bashar Talafha, Abdelrahim Elmadany, Nagham Hamad, et al. 2023. Wojoodner 2023: The first arabic named entity recognition shared task. In *Proceedings of ArabicNLP 2023*, pages 748–758.

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*.

Mayank Kejriwal. 2022. Knowledge graphs: Constructing, completing, and effectively applying knowledge graphs in tourism. In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, pages 423–449. Springer.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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

Nandana Mahanta, Sourish Dhar, and Sudipta Roy. 2016. Entity recognition in assamese text. In *2016 International Conference on Communication and Electronics Systems (ICCES)*, pages 1–5.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M Khapra, Pratyush Kumar, Rudra Murthy V, and Anoop Kunchukuttan. 2022. Naamapadam: A large-scale named entity annotated data for indic languages. *arXiv preprint arXiv:2212.10168*.

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

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, Vancouver, Canada. Association for Computational Linguistics.

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

Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. 2022b. Aspos: Assamese part of speech tagger using deep learning approach. in 2022 ieee/acs 19th international conference on computer systems and applications (aiccsa).

Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. 2023. Part-of-speech tagger for assamese using ensembling approach. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(10):1–22.

Dhrubajyoti Pathak, Sukumar Nandi, and Priyankoo Sarmah. 2024. Evaluating performance of pre-trained word embeddings on assamese, a low-resource language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6418–6425.

Bornali Phukon, Akash Anil, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2021. Synonymy expansion using link prediction methods: A case study of assamese wordnet. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–21.

Pattabhi RK Rao, CS Malarkodi, R Vijay Sundar Ram, and Sobha Lalitha Devi. 2015. Esm-il: Entity extraction from social media text for indian languages@ fire 2015-an overview. In *FIRE workshops*, pages 74–80.

Navanath Saharia, Dhrubajyoti Das, Utpal Sharma, and Jugal Kalita. 2009. Part of speech tagger for assamese text. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, pages 33–36.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Khurniawan Eko Saputro, Sri Suning Kusumawardani, and Silmi Fauziati. 2016. Development of semi-supervised named entity recognition to discover new tourism places. In *2016 2nd International Conference on Science and Technology-Computer (ICST)*, pages 124–128. IEEE.

Shikhar Kr Sarma, R Medhi, M Gogoi, Utpal Saikia, et al. 2010. Foundation and structure of developing an assamese wordnet. In *Proceedings of 5th international conference of the global WordNet Association*.

Jumi Sarmah, Shikhar Kumar Sarma, and Anup Kumar Barman. 2019. Development of assamese rule based stemmer using wordnet. In *proceedings of the 10th Global WordNet Conference*, pages 135–139.

Satoshi Sekine and Hitoshi Isahara. 2000. Irex: Ir & ie evaluation project in japanese. In *LREC*, pages 1977–1980.

Padmaja Sharma, Utpal Sharma, and Jugal Kalita. 2012. Suffix stripping based ner in assamese for location names. In *2012 2nd National Conference on Computational Intelligence and Signal Processing (CISP)*, pages 91–94.

Padmaja Sharma, Utpal Sharma, and Jugal Kalita. 2014. Named entity recognition in assamese using crfs and rules. In *2014 International Conference on Asian Language Processing (IALP)*, pages 15–18.

Padmaja Sharma, Utpal Sharma, and Jugal Kalita. 2016. Named entity recognition in assamese: A hybrid approach. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2114–2120.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ere: Annotation of entities, relations, and events. In *Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98.

Gitimoni Talukdar, Pranjal Protim Borah, and Arup Baruah. 2014. Supervised named entity recognition in assamese language. In *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, pages 187–191.

Gitimoni Talukdar, Pranjal Protim Borah, and Arup Baruah. 2018. Assamese named entity recognition system using naive bayes classifier. In *Advances in Computing and Data Sciences*, pages 35–43, Singapore. Springer Singapore.

Kuwali Talukdar and Shikhar Kumar Sarma. 2023. Upos tagger for low resource assamese language: Lstm and bilstm based modelling. In *2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, pages 1–6. IEEE.

Kuwali Talukdar, Shikhar Kumar Sarma, Farha Naznin, and Ratul Deka. 2024. Deep learning based upos tagger for assamese religious text. *International Journal of Religion*, 5(4):163–170.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*, volume 2008, pages 96–101.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, 17.

Annisa Zahra, Ahmad Fathan Hidayatullah, and Septia Rani. 2022. Bidirectional long-short term memory and conditional random field for tourism named entity recognition. *Int J Artif Intell ISSN*, 2252(8938):1271.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.