Natural Language Processing Journal 11 (2025) 100160

Contents lists available at ScienceDirect

Natural Language Processing Journal

journal homepage: www.elsevier.com/locate/nlp

A multi-way parallel named entity annotated corpus for English, Tamil and Sinhala

Surangika Ranathunga [a,*], Asanka Ranasinghe [b], Janaka Shamal [b], Ayodya Dandeniya [b], Rashmi Galappaththi [b], Malithi Samaraweera [b]

[a] School of Mathematical and Computational Sciences, Massey University, Auckland, 102904, New Zealand
[b] Department of Computer Science and Engineering, University of Moratuwa, Katubedda, 10400, Sri Lanka

ARTICLE INFO

Keywords:
Named entity recognition
Pre-trained language models
Low resource languages
Sinhala
Tamil
Large language models (LLMs)

ABSTRACT

This paper presents a multi-way parallel English-Tamil-Sinhala corpus annotated with Named Entities (NEs), where Sinhala and Tamil are low-resource languages. Using pre-trained multilingual Language Models (mLMs), we establish new benchmark Named Entity Recognition (NER) results on this dataset for Sinhala and Tamil. We also carry out a detailed investigation on the NER capabilities of different types of LMs. Finally, we demonstrate the utility of our NER system on a low-resource Neural Machine Translation (NMT) task. Our dataset is publicly released: https://github.com/suralk/multiNER.

1. Introduction

Named Entity Recognition (NER) is the process of identifying Named Entities (NEs) in natural language text. An NE can be a word or a phrase, and the detected entities are categorized into predetermined categories such as person, location and organization. As an example, consider the sentence: *John works in Facebook at Los Angeles*. This sentence contains three NEs: John (person), Facebook (organization) and Los Angeles (location).

NER either acts as an intermediate step for, or helps to improve many high-level Natural Language Processing (NLP) tasks such as question answering (Lamurias and Couto, 2019), Neural Machine Translation (NMT) (Hu et al., 2022; Sulistyo et al., 2025), Information Retrieval (Guo et al., 2009) and automatic text summarization (Khademi and Fakhredanesh, 2020). NER is useful in end-user applications as well. One very good example is identifying fine-grained location names from social media posts related to rescue requests during disaster situations (Hu et al., 2024). This would enable the rescue teams to promptly attend to those rescue requests. A domain-specific example is extracting biological NEs from scientific text - accurate recognition of these entities is crucial for downstream tasks such as including gene-disease association and drug re-purposing (Jung et al., 2024).

There have been many developments with respect to NE tag sets (Li et al., 2020), tagging schemes (Ringland et al., 2019a), as well as NER techniques (Li et al., 2020). Most of the NER algorithms that produced promising results are supervised, meaning that they are trained with NE annotated datasets. While there have been efforts to produce NE annotated data (Lima et al., 2023), many low-resource languages still have little to no annotated data (Joshi et al., 2020).

Sinhala and Tamil are examples of low-resource languages (Joshi et al., 2020; Ranathunga and de Silva, 2022) with limited NE annotated data (De Silva, 2019). Manamini et al.'s (2016) is the only publicly available Sinhala NE dataset that has been manually annotated. FIRE corpus[1] is the only publicly available manually annotated NE dataset for Tamil. WikiANN (Pan et al., 2017) and LORELEI (Tracey and Strassel, 2020) are multilingual (but not multi-way pararell) NE annotated datasets that contain both Sinhala and Tamil. However, WikiANN has been automtically annotated using entity linking and LORELEI is hidden behind a paywall.

In this paper, we present a multi-way parallel English-Sinhala-Tamil NE annotated corpus that consists of 3835 sentences per language. This corpus is annotated using the CONLL03 tag set (Sang and De Meulder, 2003), which has four tags: persons (PER), locations (LOC), organizations (ORG) and miscellaneous (MISC). The corpus was annotated using Beginning-Inside-Outside (BIO) format. We provide a comprehensive analysis of this dataset and the data creation process. This dataset is publicly released.[2]

The benefit of having such multi-way parallel datasets is that they can serve as good test beds to evaluate language-specific NER capabilities of pre-trained multilingual Language Models (mLMs), which form

* Corresponding author.
E-mail address: s.ranathunga@massey.ac.nz (S. Ranathunga).

[1] http://fire.irsi.res.in/fire/2023/home
[2] https://github.com/suralk/multiNER

https://doi.org/10.1016/j.nlp.2025.100160
Received 17 December 2024; Received in revised form 4 May 2025; Accepted 28 May 2025

the basis of modern-day multilingual NLP systems. In other words, if the mLM is fine-tuned with data from individual languages, this results in language-specific NER models, which can be probed to identify how their performance varies across languages. On the other hand, fine-tuning the mLM with all the language data of the multi-way parallel corpus results in a single NER model that caters for all the languages included in the corpus. An understading on how the performance of these models varies depending on the linguistic properties of individual languages is useful in building optimal language-specific NER systems.

Despite the NER models built on pre-trained LMs outperforming the more traditional Deep Learning (DL) models such as Recurrent Neural Networks (specifically BiLSTM-CRF (Bi-Directional Long Short Term Memory with a CRF layer) (Li et al., 2020; Yadav et al., 2018)), these newer techniques have not been employed for Sinhala or Tamil NER.

In order to build the Sinhala and Tamil NER systems, we experimented with different types of pre-trained LMs. These include encoder-based LMs (eLMs): language-specific, language family-specific and multilingual, as well as Large Language Models (LLMs). Our results show that, when the language is already included in the pre-trained LM, NER systems built on that LM significantly outperform those that use the BiLSTM-CRF model. We also show that a multilingual NER model built by fine-tuning the multilingual eLM XLM-R (Conneau et al., 2019) with our multi-way parallel corpus outperforms (or is on-par with) the NER models trained for individual languages. Finally, we demonstrate the utility of the built NER models by using their output in building an NMT system. The NMT system, which was trained using the NEs identified using our NER system as an additional input significantly outperformed the baseline NMT model.

2. Related work

In this section, we provide a brief overview of NE tag sets, annotation schemes, NE annotated datasets, NER techniques based on pre-trained LMs, as well as NER research for low-resource languages. We also discuss NER research available for Sinhala and Tamil.

2.1. NE tag sets

One prominent tag set that has been widely used in NER is the CONLL03 tag set (Sang and De Meulder, 2003). This tag set has only 4 tags - Person, Location, Organization and Miscellaneous. The Co-reference and Entity Type tag set (Weischedel and Brunstein, 2005) has 12 NE types (7 numerical types, 10 nominal entity types and temporal types). WNUT2017 (Derczynski et al., 2017) corpus has 6 entity types. ACE 2005/2008 (Song et al., 2015) has identified 7 NE types, as people, organizations, locations, facilities, Geo political entities, weapons, vehicles and events. ACE was defined for the tourism domain, thus can be considered as a domain-specific tag set. While having more NE types helps in extracting more information, training Machine Learning (ML) models for the task becomes very data intensive, which in turn makes manual data annotation costly. As shown by Azeez and Ranathunga (2020a), if a sufficient annotated corpus is not created, having a fine-grained tag set is not useful, as many NE tags will not have sufficient data points to train the ML model.

2.2. Annotation schemes

Flat NE annotation schemes have been most commonly used in previous research. Alshammari and Alanazi (2021) present a comprehensive list of such annotation schemes. Some popular NE annotation schemes are IO (tags each token either as an inside tag (I) or an outside tag (O), where NEs are marked as I), IOB (which is also known as BIO - tags each token either as beginning (B) of a known NE, inside (I) it, or outside (O) of any known NE) and IOE (similar to IOB, but tags the end of an NE (E) instead of its beginning).

The main drawback of these flat annotation schemes is that they are unable to capture nested NEs. As an alternative, markup NE annotation schemes have been proposed (Mitchell et al., 2005; Ringland et al., 2019b; Marcus et al., 2011). While more comprehensive annotation schemes allow the extraction of more fine-grained information, this requires more input from humans who annotate the data and ML models need more data to learn. Therefore, still the BIO tagging scheme, which was used in the CONLL 2003 shared task (Sang and De Meulder, 2003) is being commonly used.

2.3. NER techniques

eLMs such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and their multilingual variants such as mBERT and XLM-R have been widely used for the NER task. In fact, NER was one of the evaluation tasks used for the BERT and XLM-R papers. Subsequent research has introduced various improvements on the vanilla fine-tuning of eLMs for NER (Souza et al., 2019; Fetahu et al., 2022; Huang et al., 2022; Fu et al., 2022). This line of research has outperformed the traditional DL techniques such as RNNs by a significant margin. For a comprehensive survey on the previous DL techniques for NER, we refer the reader to Yadav and Bethard (2019).

Some research that used eLMs, combined NE corpora from multiple languages (not multi-way parallel) to build a single multilingual NER model (Kulkarni et al., 2023; Shaffer, 2021), which has shown to perform better than language-specific models. This is due to the cross-lingual transfer of knowledge across languages in the eLM. Thus such multilingual eLMs are a very promising solution for low-resource languages. If there is no NE annotated data for a target language, one solution is to synthetically generate NE data using an NER model of a high-resource language (Li et al., 2021; Yang et al., 2022). Another popular solution is Transfer Learning, where an NER model fine-tuned on an eLM is used to infer NE information on a target language (Adelani et al., 2021).

With the advent of decoder-based LLMs such as Llama (Touvron et al., 2023), Mistral (Jiang et al., 2023) and Gemma (GemmaTeam et al., 2024), the task of NER has been revisited. There has been research on developing advanced prompting strategies for zero-shot NER (Xie et al., 2023), as well as different fine-tuning (Zhao et al., 2024) and decoding (Lu et al., 2024) techniques. However, these LLMs are very much English dominant and have very limited language coverage. While a few high and medium resource languages have been included in the pre-training data, low-resource languages have been largely ignored.

All Sinhala NER research except Rijhwani et al. (2020) (who used an LSTM) is based on traditional Machine Learning (ML) models such as SVM and CRF (Dahanayaka and Weerasinghe, 2014; Manamini et al., 2016; Azeez and Ranathunga, 2020b; Senevirathne et al., 2015; Wijesinghe and Tissera, 2022). Similarly for Tamil, except Anbukkarasi et al. (2022) and Hariharan et al. (2019), all the other research is based on traditional ML techniques (Murugathas and Thayasivam, 2022; Srinivasagan et al., 2014; Srinivasan and Subalalitha, 2019; Abinaya et al., 2015; Vijayakrishna and Sobha, 2008; Antony and Mahalakshmi, 2014; Abinaya et al., 2014; Gayen and Sarkar, 2014; Murugathas and Thayasivam, 2022; Theivendiram et al., 2018; Mahalakshmi et al., 2016; Malarkodi and Devi, 2020; Malarkodi et al., 2012; Ram et al., 2010).

2.4. NE annotated corpora for Sinhala and Tamil

Manamini et al. (2016) produced an NE annotated dataset[3] for Sinhala. However, it has not used a standard tagging schema such as BIO. Other Sinhala NER research has not publicly released their

[3] http://bit.ly/2XrwCoK

datasets. The FIRE corpus is the most commonly used dataset for Tamil NER. Other research (Antony and Mahalakshmi, 2014) that created NER datasets has not published their data.

WikiAnn corpus (Pan et al., 2017) is one of the efforts to build an NE annotated dataset for low resource languages. It was created by transferring annotations from English to other 282 languages through cross-lingual links in knowledge bases. Thus, the WikiAnn corpus is said to have "silver-standard" labels. Although it has both Sinhala and Tamil data, the amount is rather small, specially for Sinhala. Moreover, WikiAnn corpus has been annotated with only three major entities: Person, Organization and Location. LORELEI (Tracey and Strassel, 2020) is another multilingual NE annotated corpus that includes both Sinhala and Tamil. Unlike WikiAnn, LORELEI is a manually annotated corpus. However, it is hidden behind a paywall.

2.5. NER for low-resource languages

A pre-cursor for a successful NER system is a Named Entity annotated dataset. Table 1 lists the Named Entity annotated datasets available for some low-resource languages.

In addition to the WikiAnn and LORELEI multilingual dataets mentioned above, there exist some multilungual datasets such as MultiCoNER (Malmasi et al., 2022) and MasakhaNER (Adelani et al., 2021). MultiCoNER was built using a technique similar to WikiANN. However, data for low-resource languages has been created by Machine Translation. MasakhaNER corpus contains manually annotated data for low-resource African languages. None of these are multi-way parallel.

According to recent research, the use of eLMs have become a popular choice for NER in low-resource languages. While both monolingual eLMS, as well as multilingual eLMs such as XLM-R have been used, there is no general consensus on the best-performing eLM type. For example, Torge et al. (2023) reported that both mono-lingual and language family models outperform their multi-lingual counterpart. Similarly, Snæbjarnarson et al. (2023) showed that language family models outperform XLM-R for the low-resource Faroese language. They also demonstrated that an eLM trained on a high-resource language can produce on-par results. Subedi et al. (2024) reported that a language-specific eLM was the best in their experiments for Nepalese.

As mentioned above, multilingual NER systems built on multilingual eLMs has been quite successful in the context of low-resource languages. In addition, there have been further improvements on techniques that use eLMs. Mehari Yohannes et al. (2024) employed a collaborative learning setup that made use of two different multilingual eLMs. Sohn et al. (2024) converted the high-resource language data into a phonetic alphabet, before using it to fine-tune an eLM. The low-resource language test data was also converted into the same phonetic alphabet, to benefit better from Transfer Learning.

3. Multi-way parallel English-Tamil-Sinhala dataset

In this section, first we give a brief introduction to Sinhala and Tamil. Then we discuss the data source used to create the NE annotated dataset, the pre-processing and annotation steps we have performed, data format we used and statistics of the multi-way parallel dataset.

3.1. Sinhala and Tamil languages

Sinhala is an Indo-Aryan language spoken by approximately 16 million people, primarily in Sri Lanka. It possesses a unique alphabet and script. Tamil is a Dravidian language spoken by around 78 million people, predominantly in the Indian state of Tamil Nadu. It is widely spoken in Sri Lanka, Singapore, and Malaysia as well. The language is written using the unique Tamil script. According to Ranathunga and de Silva's (2022) language classification, Sinhala is categorized as a class 2 language and Tamil as a class 3 language. Based on Joshi et al.'s (2020) language category definition, this classification indicates that Sinhala has limited annotated datasets available. In contrast, Tamil being in class 3 indicates that it has limited labeled datasets but a more prominent web presence compared to Sinhala.

Table 1
NE annotated datasets for low-resource languages.

Language	Research	Daset size	Tags
Bhojpuri, Maithili, Magahi	Mundotiya et al. (2023)	228,373, 157,468, and 56,190 tokens	22
Faroese	Snæbjarnarson et al. (2023)	102k words	8
Tigrinya	Yohannes and Amagasa (2022)	69,309 tokens	4
Odia	Dalai et al. (2025)	6,71,354 tokens	12
Algerian	Dahou and Cheragui (2023)	220k+ tokens	3
Punjabi	Ahmad et al. (2020)	318,275 tokens	3
Urdu	Khan et al. (2022)	2161 sentences	7
Kannada	Sathyanarayanan et al. (2018)	46110 tokens	3
Icelandic	Ingólfsdóttir et al. (2020)	1 million tokens	8
Bengali	Lima et al. (2023)	10,05,791 tokens	13

3.2. Raw data

Our data comes from the multi-way parallel English-Tamil-Sinhala dataset developed by Fernando et al. (2020). This corpus contains official government documents, namely annual reports, letters and circulars. Despite being specific to the government domain, this dataset has a wide coverage mainly because of the inclusion of annual reports coming from different government institutions corresponding to Art, Media, Finance, Education, Technology, Procurement, etc.

This corpus contains duplicate sentences, unwanted long lists (with 200+ tokens per list) such as table of contents and meaningless sentences. Thus, we manually filtered 3835 Sinhala sentences[4]. During the filtering process, we removed meaningless sentences, duplicate sentences, undesirable lists and captions/headers of figures/tables from the dataset. For each Sinhala sentence in this filtered corpus, we extracted the corresponding Tamil and English sentences from Fernando et al.'s (2020) raw parallel corpus.

3.3. Data annotation

Since our dataset has only about 100k tokens per language, we did not want to go for a fine-grained NE tag set. Therefore the annotation was carried out using the CONLL03 tag set. BIO annotation scheme was used for annotation. This corpus was annotated manually using the Inception annotation tool.[5] Two independent annotators for each language were employed to annotate the dataset. Annotators were provided an in-house training and annotation guidelines. Later, in order to establish the inter-annotator agreement, two more annotators annotated about 500 tokens from each language. The inter-annotator agreement values were reported as 0.83, 0.89 and 0.88 for Sinhala, English and Tamil (respectively).

Fig. 1 is an example annotated sentence in all three languages.

3.4. Data statistics

Our final dataset consists of 3835 parallel sentences per language. Sinhala, Tamil and English vocabularies of the prepared dataset contains 11560, 19308 and 10607 distinct word tokens respectively. Table 2 shows the tag counts in each dataset. Table 3 shows the entity percentage in each dataset. Altogether, 9.74% of tokens in the corpus are NEs. This amount is in a similar range as the amount of NEs identified in the CoNLL03 dataset. Given that this dataset is from

[4] We started with Sinhala sentences because Fernando et al.'s (2020) corpus was compiled by taking Sinhala as the source.

[5] https://inception-project.github.io/downloads/

```
The [O] main [O] objective [O] in [O] this [O] project [O] is [O] the [O] displaced [O]
families [O] in [O] Kilinochchi [B-LOC] District [I-LOC] and [O] this [O] particular [O]
study [O] reveals [O] the [O] existence [O] of [O] about [O] 7000 [B-MISC] numbers [O]
of [O] displaced [O] families [O] . [O]

இந்த [O] செயற்றிட்டம் [O] கிளிநொச்சி [B-LOC] மாவட்டத்தில் [O] பாதிக்கப்பட்ட [O] குடும்பங்களை [O]
கவனத்தில் [O] கொள்கின்றது [O] மற்றும் [O] இந்தக் [O] கற்கையில் [O] பாதிக்கப்பட்ட [O] குடும்பங்களில் [O]
உள்ள [O] 7000 [B-MISC] நபர்களை [I-MISC] ஆய்வுசெய்கின்றது [O] . [O]

මෙම [O] ව්‍යාපෘතියෙහි [O] මූලික [O] අරමුණ [O] වන්නේ [O] කිලිනොච්චි [B-LOC] දිස්ත්‍රික්කයේ [I-LOC] අවතැන් [O]
වූ [O] පවුල් [O] 7000ක් [B-MISC] පමණ [O] සංඛ්‍යාවක් [O] පිළිබඳ [O] මෙම [O] සුවිශේෂී [O] අධ්‍යයනයේ [O]
දී [O] හෙළිදරව් [O] කර [O] ඇත [O] . [O]
```

Fig. 1. Sample English/Tamil/Sinhala sentences annotated with CONLL03 tag set, following the BIO format.

Table 2
Number of entities in each dataset.

Tag	English	Sinhala	Tamil
B-PER	194	194	226
I-PER	542	610	406
B-ORG	1539	1628	1420
I-ORG	3721	3666	2247
B-LOC	1511	1728	1829
I-LOC	552	780	437
B-MISC	6283	6549	6012
I-MISC	8587	7354	7543
O	82 335	73 493	62 947

official government documents, it has more location and organization NEs than person NEs. The percentage of all the other NEs being just 6.55% justifies our selection of the ConLL tagset with 4 tags - Had the Miscellaneous tag been expanded into unique NEs, each category would have a very low amount of samples annotated.

According to Table 2, we could see that B-tag counts of Person, Organization and Location entities do not tally up across languages. Following are the reasons (and an example for each) we identified for this discrepancy.

1. Translation discrepancies due to dissimilar language syntax

English: District O | of O | Ratnapura B-LOC
Sinhala: රත්නපුර B-LOC | දිස්ත්‍රික්කය I-LOC

Here, *Rathnapura District*, has been wriitten as *District of Rathnapura*, which has been marked as having one NE. However the Sinhala translation of it contains two NE tags.

2. Differences in how abbreviations and acronyms have been handled in different languages

ACCIMT B-MISC නවීන B-ORG | තාක්ෂණ I-ORG | පිළිබඳ I-ORG | ආතර් I-ORG | සී I-ORG | . I-ORG | ක්ලාක් I-ORG | ආයතනය I-ORG

The English sentence has the NE as an acronym, whereas the Sinhala version has the expanded version of it. Note that the English abbreviation ACCIMT has been annotated as B-MISC, while the corresponding Sinhala entity has been identified as an ORG.

3. Annotation mistakes by human annotators

16 B-MISC | other O | institutions O | coming O | under O | the O | Ministry B-ORG

In this phrase, the last term has been annotated as B-ORG simply because the word (incorrectly) starts with capitalization, although it does not refer to any specific ministry.

Table 3
Percentage of NEs in each dataset.

NE Type	English	Sinhala	Tamil
PER	0.18	0.19	0.26
LOC	1.38	1.73	2.10
ORG	1.63	1.41	1.63
MISC	6.55	5.76	6.92

4. Methodology

As mentioned in Section 2.3, the DL techniques used so far for Sinhala and Tamil NER are based on RNN models. Therefore we used one of these models as the baseline. To be specific, we used a Bi-LSTM CRF model, as it has been the state-of-the-art architecture before the introduction of pre-trained eLMs (Yadav and Bethard, 2019).

As for NER model implementation with eLMs, we experimented with three variants: a language-specific eLM, a language-family specific eLM, and two mLMs. As the LLM, we used Llama 3.1 8B instruct tuned model.

4.1. Bi-LSTM CRF method

We used the BiLSTM-CRF model used by Yadav et al. (2018), which has shown to be the best recurrent model for NER. In addition to character and word embeddings, this model uses prefix-suffix embeddings as the input. The model architecture is shown in Fig. 2. In our implementation, we used FastText to obtain word embeddings. We experimented with both prefix and suffix features, however only prefix features resulted in improved results, hence only that result is reported. We used three techniques to generate character embeddings:

1. Convolutional Neural Networks (CNNs) (Ma and Hovy, 2016)
2. LSTMs (Lample et al., 2016)
3. LSTM for character embedding + Affix features (Yadav et al., 2018)

4.2. Fine tuning pre-trained LMs

Multilingual eLMs: We experimented with mBERT and XLM-R. Both mBERT and XLM-R have been trained with the Masked Language Model (MLM) objective. However the creators of XLM-R claim that it is better than mBERT, and the same is confirmed by their experiment results. XLM-R has been trained with 100 languages, while mBERT has been trained with 104 languages. English and Tamil are included in both these models, however Sinhala is included only in XLM-R.
Language family-specific eLMs: We experimented with IndicBERT (Kakwani et al., 2020), a model specifically trained for 12 Indian languages, primarily from the Indo-Aryan and Dravidian language families. While Tamil is included in IndicBERT, Sinhala, despite being an Indo-Aryan language, is not.
Language-specific eLMs: For Sinhala, we used SinBERT (Dhananjaya et al., 2022), a language-specific model built following the RoBERTa

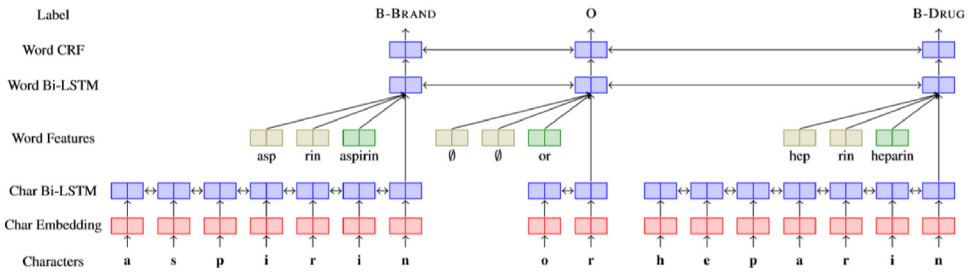

Fig. 2. Architecture of the Bi-LSTM CRF network with affix features (Yadav et al., 2018).

You are an NLP assistant whose purpose is to perform Named Entity Recognition (NER). NER involves identifying and classifying named entities in a text into predefined categories such as person names, organizations, locations, miscellaneous entities, and others. You will need to use the tags defined below:

O means the word doesn't correspond to any entity.

B-PER/I-PER means the word corresponds to the beginning of/is inside a person entity.

B-ORG/I-ORG means the word corresponds to the beginning of/is inside an organization entity.

B-LOC/I-LOC means the word corresponds to the beginning of/is inside a location entity.

B-MISC/I-MISC means the word corresponds to the beginning of/is inside a miscellaneous entity.

Do not try to answer the question! Just tag each token in the sentence

Fig. 3. The prompt used to fine-tune Llama 3.1 for the NER task.

Table 4
Language coverage in different eLMs.

	mBERT	XLM-R	IndicBERT	SinBERT	Llama 3.1
Sinhala		✓		✓	
Tamil	✓	✓	✓		
English	✓	✓	✓		✓

Table 5
Best hyper-parameter values.

Model	Learning rate
XLM-R base(Si, Ta, En)	3e−5
mBERT base(Si, Ta, En)	3e−5
SinBERT small(Si)	2.5e−4
SinBERT small(Ta, En)	2.5e−5
IndicBERT(Si, En)	9e−5
IndicBERT(Ta)	5e-5

framework. SinBERT was trained on a dataset containing 15.7 million Sinhala sentences. We could not find an equivalent language-specific eLM for Tamil.

LLM: We used Llama 3.1 8B instruc-tuned model via Unsloth. Neither Sinhala nor Tamil is included during the pre-training of this LLM.[6]

Table 4 provides an overview of the language coverage of each model. Each eLM was individually fine-tuned using our annotated dataset. A feed-forward layer was added on top of each model for classification. For the multilingual model, we combined Sinhala, English, and Tamil datasets, sampling them with a fraction of 1 to ensure a balanced mix of sentences from all three languages. In order to fine-tune Llama 3.1, the annotated dataset was converted into an instruction format. We used the prompt presented by Ahuja et al. (2023), which is shown in Fig. 3.

5. Experiment setup

For eLM fine-tuning, the hyperparameter tuning was done using Optuna Python library.[7] For all the eLMs, training batch size 8, evaluation batch size 16, 3 training epochs and a weight decay of 0.01 gave the optimal results. As shown in Table 5, only the learning rate was varied. For the LLM, we used the default hyper-parameters used by Unsloth. These values are as follows: Learning rate - 2e-4, weight decay - 0.01, batch size - 2. We used LoRA (r, alpha =16) during fine-tuning.

Evaluation was done according to the holdout method. The dataset was divided as 70-10-20 for train/validation/test sets, respectively. Eack eLm experiment was run three time with three different seeds and then the average was calculated to get the final results. However, due to computing costs, LLM experiments were run only once. All experiments were conducted on the Google Colab platform. The following GPU configurations were used for eLM experiments: CUDNN Version - 8700, Number CUDA Devices - 1, CUDA Device Name - Tesla T4, CUDA Device Total Memory [GB] - 15.8. The following configurations were used for Llama 3.1 fine-tuning: CUDA Version: 12.4, Number CUDA Devices - 1, CUDA Device Name - NVIDIA A100, CUDA Device Total Memory [GB] - 40.

6. Evaluation

Table 6 presents the final results, while Table 7 provides the tag-wise distribution of results for the XLM-R model, which gave the best results.

[6] Llama 3.1 has used pre-training data from only 7 languages: French, German, Hindi, Italian, Portuguese, Spanish, and Thai.

[7] https://optuna.org/

Table 6
Macro F1 score for different models.BLC- BiLSTM CRF, L- LSTM for char embeddings, C- CNN for char embeddings, L+A - LSTM for char embeddings with affix features.

	BLC (L)	BLC (C)	BLC (L+A)	mBERT	XLM-R	mXLM-R	IndicBERT	SinBERT	Llama3.1
Sinhala	64.91	64.67	65.66	60.53	87.71	**88.33**	49.59	83.77	46.20
Tamil	46.64	42.20	47.19	77.46	78.81	**80.23**	65.93	41.40	20.0
English	–	–	–	89.11	**89.67**	89.59	88.90	58.47	-

Table 7
Tag-wise Macro F1 score for the XLM-R model.

	PERSON	LOC	ORG	MISC	OUTSIDE
Sinhala	97.45	82.23	82.65	83.23	96.14
Tamil	79.35	81.43	69.49	76.88	94.02
English	96.6	86.18	82.92	85.57	96.71

It is important to recall that mBERT and IndicBERT were not pre-trained on Sinhala, and SinBERT was not pre-trained on Tamil or English. Lalama 3.1 was not pre-trained on either Sinhala or Tamil. Despite this, we evaluated all models across all languages.[8] If a language was not included in a particular pre-trained LM, its corresponding result is grayed out.

Among the Bi-LSTM models, the Bi-LSTM CRF with LSTM-generated input embeddings and prefix features achieved the best performance. However, its results were significantly lower than those of pre-trained LMs when the target language was included in the respective model. Note that we did not generate Bi-LSTM CRF results for English, as pre-trained LMs already achieved the highest performance for this language.

In contrast, we observed that LMs not pre-trained on a given language performed poorly for that language—for instance, mBERT's results for Sinhala and Llama 3.1 results for both Sinhala and Tamil. On a positive note, SinBERT, despite being trained on only 15.7 million Sinhala sentences, significantly outperformed the corresponding Bi-LSTM CRF results. This highlights the potential of language-specific pre-trained LMs for low-resource languages. Given that many languages have monolingual datasets, training and fine-tuning a language-specific LM for tasks such as NER can be a viable and effective approach.

Consistent with Conneau et al. (2019)'s observations, mBERT performs slightly worse than XLM-R for both English and Tamil. Meanwhile, IndicBERT lags behind both multilingual language models (mLMs), with a particularly significant drop in performance for Tamil. This is surprising, given that IndicBERT was specifically trained on Indic languages, including Tamil and other Dravidian languages. However, Kakwani et al. (2020) also reported similar findings, where IndicBERT underperformed compared to XLM-R and mBERT in multiple NLP tasks.

Another noteworthy observation is SinBERT's performance on Sinhala, which is lower than that of XLM-R. A similar trend was noted by Dhananjaya et al. (2022) in sentiment analysis. We attribute XLM-R's superior performance to its cross-lingual transfer capabilities (Asai et al., 2024). This suggests that multilingual eLMs can be a viable alternative to language-specific eLMs, especially when there is limited language-specific data available for training.

We also observed notable variations in XLM-R's performance across the three languages. Given that our dataset is multi-way parallel (i.e., the same dataset is annotated across all three languages), this variance can be explained by two key factors: the representation of each language in XLM-R and the linguistic complexity of individual languages. For example, English, which has the highest representation in XLM-R—meaning that significantly more English-language data was used during its pretraining—also achieved the highest performance. However, despite Tamil having greater representation in XLM-R than Sinhala, its performance is notably lower. This discrepancy could be due to Tamil's higher morphological complexity, as it is more agglutinative than Sinhala. Another possible explanation is that Dravidian languages (such as Tamil) are underrepresented in XLM-R compared to Indo-Aryan languages (such as Sinhala) (Ranathunga and de Silva, 2022).

We observed that the multilingually fine-tuned XLM-R (mXLM-R) achieved the best performance for both Sinhala and Tamil. However, there was a slight performance drop for English, which we attribute to the phenomenon of negative interference. This occurs when the performance of high-resource languages declines in multilingual training settings (Wang et al., 2020). Despite this minor drawback, mXLM-R remains the optimal choice for our task, as it can effectively handle all three languages within a single model.

As mentioned above, very low performance of Llama 3.1 for Sinahla and Tamil has to be due to these two languages not being included in the training data of the model. We believe that the performance of Llama 3.1 is better for Sinhala than Tamil because of the presence of Hindi, which belongs to the same language family as Sinhala.

7. Case study - improving neural machine translation with NER

Named Entity translation is challenging even to the modern day NMT systems (Läubli et al., 2020). Early solutions to this problem involve the use of external transliteration systems (Grundkiewicz and Heafield, 2018; Ameur et al., 2017; Zhang et al., 2020). However, such NE transliteration systems focus on translating individual NEs with no regard to the context of the NEs, which makes it difficult to resolve ambiguities. This is particularly problematic in highly inflected languages such as Sinhala and Tamil. Hu et al. (2022) recently proposed an alternative solution that involves pre-training a language model prior to fine-tuning it for NMT. Since their results seem to far exceed the results of NMT models that do not have explicit NE translation capabilities, we selected this model as the case study to demonstrate the usability of our NER system.

Hu et al. (2022) adopted a pre-training and fine-tuning process to implement an NMT system that has better capabilities for NE translation. They first identified NEs in a monolingual corpus and linked them to a knowledge base (KB) that contains entity translations (WikiData (Vrandečić and Krötzsch, 2014) in their case). For NE linking, they used the SLING (Ringgaard et al., 2017) entity linker. Then the entity translations in the KB were used to generate noisy code-switched data. After that, a Transformer model (Vaswani et al., 2017) was pre-trained with this noisy code-switched data using the de-noising pre-training objective. Note that this is the same objective that was used to train the encoder–decoder multilingual LM mBART (Tang et al., 2021). In de-noising pre-training, the Transformer model is taught to reconstruct an original sentence from its noised version. Finally to further improve translation of low-frequency NEs, they used a multi-task learning strategy that fine-tunes the model using both the denoising task on the monolingual data and the translation task on the parallel data. This model, known as DEEP (DEnoising Entity Pre-training) is illustrated in Fig. 4 using English and Sinhala as the example (note the code-switched data used for pre-training).

We first implemented Hu et al.'s (2022) DEEP NMT system using SLING, as well as a baseline NMT system without DEEP, for English-Sinhala (En-Si) translation. We used WikiData (Vrandečić and Krötzsch, 2014) as the monolingual dataset, and Fernando et al.'s (2020) parallel data to train the NMT model. To elaborate further, the baseline NMT system was built by simply fine-tuning a Transformer model from

[8] Except for English on Llama 3.1.

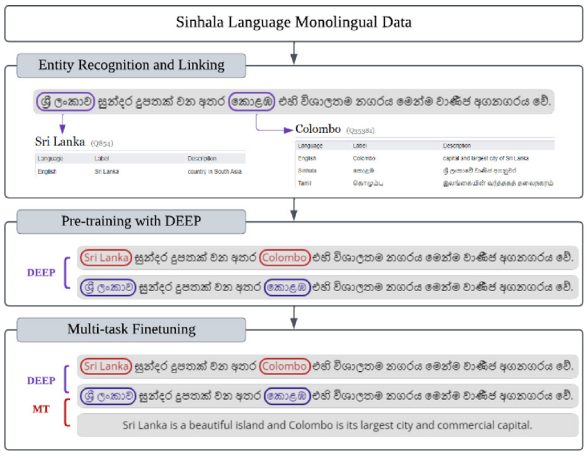

Fig. 4. DEEP (Hu et al., 2022) Architecture.

Table 8
Named entity translation results on En-Si.

Technique	BLEU	Entity translation Acc.
NMT model without DEEP	11.9	49
DEEP+SLING	11.59	47.94
DEEP+NER+Wiki data linking	**21.12**	**62.75**

scratch, similar to Hu et al. (2022). To build the DEEP NMT system, first the NEs in the Sinhala WikiData were identified and linked to their corresponding English NEs using SLING. Then the NEs in the Sinhala WikiData were replaced with the identified English NEs. These code-switched noisy sentences were used to pre-train a Transformer model using the de-noising objective. Finally, this pre-trained Transformer model was further fine-tuned with the training set of Fernando et al.'s (2020) parallel data using the NMT objective.

However, as shown in Table 8, the DEEP system with SLING lags behind the baseline system with respect to the BLEU score. This is not surprising since the SLING entity linker used in DEEP does not support Sinhala, and the WikiData corpus that was used as the KB is rather small for Sinhala.

Next, we re-implemented the DEEP system by replacing SLING with our best NER system. In other words, we used our NER system to identify NEs in Sinhala WikiData. Entity linking across the languages was done using the Pywikibot library (Anon, 2015). As shown in Table 8, the resulting NMT system significantly outperforms both baseline and DEEP with SLING (by about 9 BLEU points).

Similar to Hu et al. (2022), we also calculated entity translation accuracy. Here, we first count the number of NEs in the target side (i.e. Sinhala) of the test set used to test the NMT system. Out of these NEs, the number of NEs that got correctly translated by the NMT system is taken as the entity translation accuracy. As shown in Table 8, the NMT model that incorporates our NER output shows the highest NE translation accuracy.

8. Conclusion

In this paper, we introduced a multi-way parallel Named Entity annotated dataset for Sinhala, English, and Tamil. To the best of our knowledge, this is the first such corpus available for any language pair. We conducted a comprehensive evaluation of various pre-trained language models for the NER task. Our best-performing model is a multilingual NER system trained on the multi-way parallel corpus, demonstrating its effectiveness for this task. Additionally, we showcased the practical application of our NER system by integrating it into an English-Sinhala NMT model. Despite the promise of modern-day LLMs, our experiments with Llama 3.1 model resulted in very low results for Sinhala and Tamil, which could be due to these languages not being included in that LLM.

For future work, we plan to expand the dataset further and extend the annotation to include a broader range of Named Entities. Given that the XLM-R based NER system achieves very good performance, it is possible to use this system to annotate new data. Getting humans to clean such annotated data would be more effective than getting them to annotate data from scratch. As mentioned in Section 2, there have been work to improve LLM performance for the task of NER. In future we plan to experiment with those techniques.

CRediT authorship contribution statement

Surangika Ranathunga: Writing – review & editing, Supervision, Project administration, Methodology, Conceptualization. **Asanka Ranasinghe:** Writing – original draft, Validation, Methodology, Data curation. **Janaka Shamal:** Writing – original draft, Validation, Methodology. **Ayodya Dandeniya:** Writing – original draft, Validation, Methodology. **Rashmi Galappaththi:** Writing – original draft, Validation, Methodology. **Malithi Samaraweera:** Writing – original draft, Validation, Methodology.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT (free version) in order to improve the language quality of some of the sections. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Rameela Azeez for her initial involvement in the project, and Aravinda kankanamge, Rashad Sirajudeen and Samith Kavishke for their support in preparing experiments related to Llama 3.1.

References

Abinaya, N., John, N., Ganesh, B.H., Kumar, A.M., Soman, K., 2014. AMRITA_CEN@ FIRE-2014: named entity recognition for Indian languages using rich features. In: Proceedings of the Forum for Information Retrieval Evaluation. pp. 103–111.

Abinaya, N., Kumar, M.A., Soman, K., 2015. Randomized kernel approach for named entity recognition in tamil. Indian J. Sci. Technol. 8 (24), 7.

Adelani, D.I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S., et al., 2021. MasakhaNER: Named entity recognition for african languages. Trans. Assoc. Comput. Linguist. 9, 1116–1131.

Ahmad, M.T., Malik, M.K., Shahzad, K., Aslam, F., Iqbal, A., Nawaz, Z., Bukhari, F., 2020. Named entity recognition and classification for punjabi shahmukhi. ACM Trans. Asian Low-Resour. Lang. Inf. Process. (TALLIP) 19 (4), 1–13.

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Ahmed, M., et al., 2023. MEGA: Multilingual evaluation of generative AI. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 4232–4267.

Alshammari, N., Alanazi, S., 2021. The impact of using different annotation schemes on named entity recognition. Egypt. Inform. J. 22 (3), 295–302.

Ameur, M.S.H., Meziane, F., Guessoum, A., 2017. Arabic machine transliteration using an attention-based encoder-decoder model. Procedia Comput. Sci. 117, 287–297.

Anbukkarasi, S., Varadhaganapathy, S., Jeevapriya, S., Kaaviyaa, A., Lawvanyapriya, T., Monisha, S., 2022. Named entity recognition for tamil text using deep learning. In: 2022 International Conference on Computer Communication and Informatics. ICCCI, IEEE, pp. 1–5.

Anon, 2015. Utilizing the wikidata system to improve the quality of medical content in wikipedia in diverse languages: a pilot study. J. Med. Internet Res. 17 (5), e4163.

Antony, J.B., Mahalakshmi, G., 2014. Named entity recognition for tamil biomedical documents. In: 2014 International Conference on Circuits, Power and Computing Technologies. ICCPCT-2014, IEEE, pp. 1571–1577.

Asai, A., Kudugunta, S., Yu, X., Blevins, T., Gonen, H., Reid, M., Tsvetkov, Y., Ruder, S., Hajishirzi, H., 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 1771–1800.

Azeez, R., Ranathunga, S., 2020a. Fine-grained named entity recognition for sinhala. In: 2020 Moratuwa Engineering Research Conference. MERCon, pp. 295–300. http://dx.doi.org/10.1109/MERCon50084.2020.9185296.

Azeez, R., Ranathunga, S., 2020b. Fine-grained named entity recognition for sinhala. In: 2020 Moratuwa Engineering Research Conference. MERCon, IEEE, pp. 295–300.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2019. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

Dahanayaka, J., Weerasinghe, A., 2014. Named entity recognition for sinhala language. In: 2014 14th International Conference on Advances in ICT for Emerging Regions. ICTer, IEEE, pp. 215–220.

Dahou, A.H., Cheragui, M.A., 2023. Dzner: A large algerian named entity recognition dataset. Nat. Lang. Process. J. 3, 100005.

Dalai, T., Das, A., Mishra, T.K., Sa, P.K., 2025. OdNER: NER resource creation and system development for low-resource odia language. Nat. Lang. Process. J. 11, 100139.

De Silva, N., 2019. Survey on publicly available sinhala natural language processing tools and research. arXiv preprint arXiv:1906.02358.

Derczynski, L., Nichols, E., Van Erp, M., Limsopatham, N., 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In: Proceedings of the 3rd Workshop on Noisy User-Generated Text. pp. 140–147.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 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). Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423.

Dhananjaya, V., Demotte, P., Ranathunga, S., Jayasena, S., 2022. BERTifying Sinhala-A comprehensive analysis of pre-trained language models for sinhala text classification. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 7377–7385.

Fernando, A., Ranathunga, S., Dias, G., 2020. Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation. arXiv preprint arXiv:2011.02821.

Fetahu, B., Fang, A., Rokhlenko, O., Malmasi, S., 2022. Dynamic gazetteer integration in multilingual models for cross-lingual and cross-domain named entity recognition. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2777–2790.

Fu, Y., Lin, N., Chen, B., Yang, Z., Jiang, S., 2022. Cross-lingual named entity recognition for heterogenous languages. IEEE/ACM Trans. Audio Speech Lang. Process. 31, 371–382.

Gayen, V., Sarkar, K., 2014. An HMM based named entity recognition system for indian languages: the JU system at ICON 2013. arXiv preprint arXiv:1405.7397.

GemmaTeam, Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., et al., 2024. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295.

Grundkiewicz, R., Heafield, K., 2018. Neural machine translation techniques for named entity transliteration. In: Proceedings of the Seventh Named Entities Workshop. pp. 89–94.

Guo, J., Xu, G., Cheng, X., Li, H., 2009. Named entity recognition in query. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 267–274.

Hariharan, V., Anand Kumar, M., Soman, K., 2019. Named entity recognition in tamil language using recurrent based sequence model. In: Innovations in Computer Science and Engineering: Proceedings of the Sixth ICICSE 2018. Springer, pp. 91–99.

Hu, X., Elßner, T., Zheng, S., Serere, H.N., Kersten, J., Klan, F., Qiu, Q., 2024. DLR-GeoTweet: A comprehensive social media geocoding corpus featuring fine-grained places. Inf. Process. Manage. 61 (4), 103742.

Hu, J., Hayashi, H., Cho, K., Neubig, G., 2022. DEEP: Denoising entity pre-training for neural machine translation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1753–1766.

Huang, Y., He, K., Wang, Y., Zhang, X., Gong, T., Mao, R., Li, C., 2022. Copner: Contrastive learning with prompt guiding for few-shot named entity recognition. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 2515–2527.

Ingólfsdóttir, S.L., Guojónsson, Á.A., Loftsson, H., 2020. Named entity recognition for icelandic: Annotated corpus and models. In: Espinosa-Anke, L., Martín-Vide, C., Spasić, I. (Eds.), Statistical Language and Speech Processing. Springer International Publishing, Cham, ISBN: 978-3-030-59430-5, pp. 46–57.

Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al., 2023. Mistral 7B. arXiv preprint arXiv:2310.06825.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M., 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6282–6293.

Jung, S.J., Kim, H., Jang, K.S., 2024. Llm based biological named entity recognition from scientific literature. In: 2024 IEEE International Conference on Big Data and Smart Computing. BigComp, IEEE, pp. 433–435.

Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M.M., Kumar, P., 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In: Findings of EMNLP.

Khademi, M.E., Fakhredanesh, M., 2020. Persian automatic text summarization based on named entity recognition. Iran. J. Sci. Technol. Trans. Electr. Eng. 1–12.

Khan, W., Daud, A., Shahzad, K., Amjad, T., Banjar, A., Fasihuddin, H., 2022. Named entity recognition using conditional random fields. Appl. Sci. 12 (13), 6391.

Kulkarni, M., Preoţiuc-Pietro, D., Radhakrishnan, K., Winata, G., Wu, S., Xie, L., Yang, S., 2023. Towards a unified multi-domain multilingual named entity recognition model. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 2202–2211.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C., 2016. Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

Lamurias, A., Couto, F.M., 2019. Lasigebiotm at MEDIQA 2019: biomedical question answering using bidirectional transformers and named entity recognition. In: Proceedings of the 18th BioNLP Workshop and Shared Task. pp. 523–527.

Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., Toral, A., 2020. A set of recommendations for assessing human–machine parity in language translation. J. Artificial Intelligence Res. 67, 653–672.

Li, B., He, Y., Xu, W., 2021. Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment. arXiv preprint arXiv:2101.11112.

Li, J., Sun, A., Han, J., Li, C., 2020. A survey on deep learning for named entity recognition. IEEE Trans. Knowl. Data Eng. 34 (1), 50–70.

Lima, K.A., Md Hasib, K., Azam, S., Karim, A., Montaha, S., Noori, S.R.H., Jonkman, M., 2023. A novel data and model centric artificial intelligence based approach in developing high-performance named entity recognition for bengali language. PLoS One 18 (9), e0287818.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Lu, J., Wang, Y., Yang, Z., Liu, X., Mac Namee, B., Huang, C., 2024. PaDeLLM-NER: parallel decoding in large language models for named entity recognition. Adv. Neural Inf. Process. Syst. 37, 117853–117880.

Ma, X., Hovy, E., 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354.

Mahalakshmi, G., Antony J, B., Roshini S, B., 2016. Domain based named entity recognition using naive bayes classification. Aust. J. Basic Appl. Sci. 10 (2).

Malarkodi, C., Devi, S.L., 2020. A deeper study on features for named entity recognition. In: Proceedings of the WILDRE5–5th Workshop on Indian Language Data: Resources and Evaluation. pp. 66–72.

Malarkodi, C., Rao, P.R., Devi, S.L., 2012. Tamil NER-coping with real time challenges. In: Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages. pp. 23–38.

Malmasi, S., Fang, A., Fetahu, B., Kar, S., Rokhlenko, O., 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 3798–3809.

Manamini, S., Ahamed, A., Rajapakshe, R., Reemal, G., Jayasena, S., Dias, G., Ranathunga, S., 2016. Ananya - a named-entity-recognition (NER) system for sinhala language. In: 2016 Moratuwa Engineering Research Conference. MERCon, pp. 30–35. http://dx.doi.org/10.1109/MERCon.2016.7480111.

Marcus, R., Palmer, M., Ramshaw, R., Xue, N., 2011. Ontonotes: A large training corpus for enhanced processing. In: Olive, J., Christianson, C., McCary, J. (Eds.), Handbook of Natural LanguageProcessing and Machine Translation: DARPA GlobalAutonomous Language Exploitation.

Mehari Yohannes, H., Lynden, S., Amagasa, T., Matono, A., 2024. Semi-supervised named entity recognition for low-resource languages using dual PLMs. In: International Conference on Applications of Natural Language to Information Systems. Springer, pp. 166–180.

Mitchell, A., Strassel, S., Huang, S., Zakhary, R., 2005. Ace 2004 multilingual training corpus. Linguist. Data Consort. Phila. 1, 1–1.

Mundotiya, R., Kumar, S., Kumar, A., Chaudhary, U., Chauhan, S., Mishra, S., Gatla, P., Singh, A.K., 2023. Development of a dataset and a deep learning baseline named entity recognizer for three low resource languages: Bhojpuri, Maithili, and Magahi. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 22 (1), 1–20.

Murugathas, R., Thayasivam, U., 2022. Domain specific named entity recognition in tamil. In: 2022 Moratuwa Engineering Research Conference. MERCon, IEEE, pp. 1–6.

Pan, X., Zhang, B., May, J., Nothman, J., Knight, K., Ji, H., 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). pp. 1946–1958.

Ram, R.V.S., Akilandeswari, A., Devi, S.L., 2010. Linguistic features for named entity recognition using CRFs. In: 2010 International Conference on Asian Language Processing. IEEE, pp. 158–161.

Ranathunga, S., de Silva, N., 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. pp. 823–848.

Rijhwani, S., Zhou, S., Neubig, G., Carbonell, J.G., 2020. Soft gazetteers for low-resource named entity recognition. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8118–8123.

Ringgaard, M., Gupta, R., Pereira, F.C., 2017. SLING: A framework for frame semantic parsing. arXiv preprint arXiv:1710.07032.

Ringland, N., Dai, X., Hachey, B., Karimi, S., Paris, C., Curran, J.R., 2019a. NNE: A dataset for nested named entity recognition in english newswire. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5176–5181.

Ringland, N., Dai, X., Hachey, B., Karimi, S., Paris, C., Curran, J.R., 2019b. NNE: A dataset for nested named entity recognition in English newswire. arXiv preprint arXiv:1906.01359.

Sang, E.T.K., De Meulder, F., 2003. 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. pp. 142–147.

Sathyanarayanan, D., Ashok, A., Mishra, D., Chimalamarri, S., Sitaram, D., 2018. Kannada named entity recognition and classification using bidirectional long short-term memory networks. In: 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques. ICEECCOT, IEEE, pp. 65–71.

Senevirathne, K., Attanayake, N., Dhananjanie, A., Weragoda, W., Nugaliyadde, A., Thelijjagoda, S., 2015. Conditional random fields based named entity recognition for sinhala. In: 2015 IEEE 10th International Conference on Industrial and Information Systems. ICIIS, IEEE, pp. 302–307.

Shaffer, K., 2021. Language clustering for multilingual named entity recognition. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 40–45.

Snæbjarnarson, V., Simonsen, A., Glavaš, G., Vulić, I., 2023. Transfer to a low-resource language via close relatives: The case study on faroese. In: Proceedings of the 24th Nordic Conference on Computational Linguistics. NoDaLiDa, pp. 728–737.

Sohn, J., Jung, H., Cheng, A., Kang, J., Du, Y., Mortensen, D.R., 2024. Zero-shot cross-lingual NER using phonemic representations for low-resource languages. arXiv preprint arXiv:2406.16030.

Song, Z., Bies, A., Strassel, S.M., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., Ma, X., et al., 2015. From light to rich ERE: Annotation of entities, relations, and events. In: EVENTS@ HLP-NAACL. pp. 89–98.

Souza, F., Nogueira, R., Lotufo, R., 2019. Portuguese named entity recognition using BERT-CRF. arXiv preprint arXiv:1909.10649.

Srinivasagan, K., Suganthi, S., Jeyashenbagavalli, N., 2014. An automated system for tamil named entity recognition using hybrid approach. In: 2014 International Conference on Intelligent Computing Applications. IEEE, pp. 435–439.

Srinivasan, R., Subalalitha, C., 2019. Automated named entity recognition from tamil documents. In: 2019 IEEE 1st International Conference on Energy, Systems and Information Processing. ICESIP, IEEE, pp. 1–5.

Subedi, B., Regmi, S., Bal, B.K., Acharya, P., 2024. Exploring the potential of large language models (LLMs) for low-resource languages: A study on named-entity recognition (NER) and part-of-speech (POS) tagging for nepali language. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation. LREC-COLING 2024, pp. 6974–6979.

Sulistyo, D.A., Prasetya, D.D., Ahda, F.A., Wibawa, A.P., 2025. Pivoted low resource multilingual translation with NER optimization. ACM J. Data Inf. Qual..

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., Fan, A., 2021. Multilingual translation from denoising pre-training. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 3450–3466.

Theivendiram, P., Uthayakumar, M., Nadarasamoorthy, N., Thayaparan, M., Jayasena, S., Dias, G., Ranathunga, S., 2018. Named-entity-recognition (ner) for tamil language using margin-infused relaxed algorithm (mira). In: Computational Linguistics and Intelligent Text Processing: 17th International Conference. CICLing 2016, Konya, Turkey, April 3–9, 2016, Revised Selected Papers, Part I 17, Springer, pp. 465–476.

Torge, S., Politov, A., Lehmann, C., Saffar, B., Tao, Z., 2023. Named entity recognition for low-resource languages-profiting from language families. In: Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023. SlavicNLP 2023, pp. 1–10.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Tracey, J., Strassel, S., 2020. Basic language resources for 31 languages (plus English): The LORELEI representative and incident language packs. In: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for under-Resourced Languages (SLTU) and Collaboration and Computing for under-Resourced Languages. CCURL, pp. 277–284.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Vijayakrishna, R., Sobha, L., 2008. Domain focused named entity recognizer for tamil using conditional random fields. In: Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages.

Vrandečić, D., Krötzsch, M., 2014. Wikidata: a free collaborative knowledgebase. Commun. ACM 57 (10), 78–85.

Wang, Z., Lipton, Z.C., Tsvetkov, Y., 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. EMNLP, pp. 4438–4450.

Weischedel, R., Brunstein, A., 2005. BBN pronoun coreference and entity type corpus. Linguist. Data Consort. Phila. 112.

Wijesinghe, W., Tissera, M., 2022. Sinhala named entity recognition model: Domain-specific classes in sports. In: 2022 4th International Conference on Advancements in Computing. ICAC, IEEE, pp. 138–143.

Xie, T., Li, Q., Zhang, J., Zhang, Y., Liu, Z., Wang, H., 2023. Empirical study of zero-shot NER with ChatGPT. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 7935–7956.

Yadav, V., Bethard, S., 2019. A survey on recent advances in named entity recognition from deep learning models. arXiv preprint arXiv:1910.11470.

Yadav, V., Sharp, R., Bethard, S., 2018. Deep affix features improve neural named entity recognizers. In: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. pp. 167–172.

Yang, J., Huang, S., Ma, S., Yin, Y., Dong, L., Zhang, D., Guo, H., Li, Z., Wei, F., 2022. CROP: Zero-shot cross-lingual named entity recognition with multilingual labeled sequence translation. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 486–496.

Yohannes, H.M., Amagasa, T., 2022. Named-entity recognition for a low-resource language using pre-trained language model. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. pp. 837–844.

Zhang, Z., Hirasawa, T., Houjing, W., Kaneko, M., Komachi, M., 2020. Translation of new named entities from English to Chinese. In: Proceedings of the 7th Workshop on Asian Translation. pp. 58–63.

Zhao, J., Liu, C., Liang, J., Li, Z., Xiao, Y., 2024. A novel cascade instruction tuning method for biomedical ner. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 11701–11705.