Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2024, 11(11):52-55**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Indic Language Named Entity Recognition Using Small and Large Language Models

Rahul Kavi

Independent Researcher, USA

Abstract: Language Models are very good at NLP tasks such as named entity recognition (NER). Relatively limited research has been done in this area of NER with languages in the Indian sub-continent (Indic Languages). This work addresses the gap in exploring the use of medium and large language models to detect named entities in text in Indian languages (with non-Latin script). The WikiANN dataset is chosen as it is a widely used, popular named entity recognition dataset (available in over 200 languages). A subset of Indic languages such as Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil and Telugu are chosen. The chosen language models include DistilBERT and XLM-RoBERTa. These models have been pre-trained on the above-chosen Indic languages. Fine-tuning for these models on the above nine languages is performed to assess the performance of named entities such as Location, Person, and Organization. DistilBERT is a medium-sized model that can be trained for NER tasks. XLM-RoBERTa is a large language model widely used for several NLP tasks (including NER). The performance of these language models is computed entity-wise for all nine languages. The DistilBERT model performs well with a macro F1 score of 0.78. The XLM-RoBERTa model beats the DistilBERT model with a performance of 0.85 macro F1 score.

Keywords: BERT, Language Models, Named Entity Recognition, Natural Language Processing.

1. Introduction

Named Entity Recognition (NER) is among the most essential NLP tasks. Models to solve this task are widely used in industry NLP pipelines. This task involves taking input text and categorizing each word into pre-determined named entities, e.g., Person, Organization, Location, Date, Event, Artifact, etc. The entities provide a vital preliminary understanding of the underlying text (that this was run on) and provide helpful insights into the content of the data. Classical approaches to train and detect named entities include extracting word tokens, part of speech tags, etc. Feature vectors are hand-engineered, and classification approaches such as Support Vector Machine (SVM) and Conditional Random Fields (CRF) are commonly used in traditional approaches to solving NER tasks. Named Entity datasets and trained models on Indic languages are a recent development (10 years ago) compared to the rest of the industry. With the onset of smartphones, content in non-Latin scripts has increased exponentially. According to a recent report, the preferred language for Indian users to use on the internet is non-English (almost 58%) compared to English (around 42%) [1]. However, there is a need for more (quality) labeled data for these models on Indic languages. With the recent trends in AI being specific to non-Indic languages, there is an immense need for models and datasets for Indic languages. More widely used Large Language Models (out of the box), such as LLAMA3 (instruct) and Mistral (instruct) are primarily designed to be used in the English language (some variants also include European languages). These models must be fine-tuned to be used for languages that it wasn't designed for. This process may be cumbersome, and it may affect model efficiency. Other simpler models (in masked language modeling) based on transformers (BERT-like) are widely used for language classification, named entity recognition models. Several of these models are designed to work with Indic languages out of the box (e.g., DistilBERT and XLM-RoBERTa). The performance of

transformer-based language models is evaluated on the WikiANN dataset. This dataset is also known as the PAN-X (based on Wikipedia articles). This dataset contains Indic languages such as Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil and Telugu. The DistilBERT and XLM-RoBERTa support all nine chosen Indic languages. These models are fine-tuned and compared entity-wise (Person: PER, Organization: ORG, and Location: LOC). The performance of these models is also broken down in terms of macro-average F1 score. This analysis is done for all nine languages (on DistilBERT and XLM-RoBERTa). The findings are presented in the conclusion section.

2. Related Work

Named Entity Recognition is a widely studied topic in Natural Language Processing. Hidden Markov Models (HMMs) were one of the first approaches applied to solve NER problems [2]. This relied on features such as: If the word is a two/four-digit number, If a word contains the date, If a given word is the first word in a sentence; If a word is in lower case. These hand-engineered features were English-focused. Zhou et al. [3] proposed a HMM Chunk Tagger (based on which an NER system was built). In 2004, Zhao et al. [4] proposed a word-similarity-based approach used along with HMMs to build a NER system for biomedical texts. Kazama et al. [5] proposed an SVM-based approach to solve NER tasks. Mayfield [6] proposed an SVM-Lattice approach (without relying on language-specific features) to address NER tasks. Das et al. [7] proposed a CRF-based approach to solve NER classification for English and Indian languages. Shobana et al. [8] proposed CRF-based NER classification using word-based features (such as word prefixes, suffixes, context word features, and part of speech information). This information was used to categorize named entities in geological text (e.g., country, state, city, island, river, etc). In the mid-2010s, Neural Nets and Deep Learning-based algorithms were widely considered as the state of the art. Chowdhury et al. [9] designed a bi-directional RNN (Recurrent Neural Network) based approach to evaluate the performance of NER on Chinese medical records. Li et al. [10] developed an LSTM (Long Short-Term Memory) approach to solve i2b2 NER challenges. GRU (Gated Recurrent Unit) based approaches were also widely applied. Chiu et al. [11] explored the word and character-level feature-based bidirectional LSTM-CNN approach to solve NER tasks. After introducing transformers, BERT-like [12] models became very popular in the NLP literature. A multilingual version of BERT is widely used for named entity recognition tasks. Hakala et al. [13] applied Multilingual BERT on the Pharma/Bio-medical dataset in Spanish. Arkhipov et al. [14] showed how multilingual BERT extended to Slavic languages (using unsupervised pre-training). BERT was combined with a CRF layer to perform multilingual NER. Work in NER tasks in Indic languages has been relatively new. Litake et al. [15] used BERT and other variants (XLM-RoBERTa) to perform NER tasks (in low-resource languages such as Marathi and Hindi). Dhamecha et al. [16] used mBERT, IndicBERT, and RoBERTa variants of BERT to fine-tune and perform NER tasks on several open-source datasets. Bahad et al. [17] performed multilingual NER analysis on four Indic languages using HiNER [18] and IndicNER models [19]. Most of the variants of BERT are pre-trained on new languages (with Masked Language Modeling task) with an appropriate tokenizer to extend the model to a new language. However, relatively speaking, there are fewer quality datasets for Indic languages.

3. Current Approach

The dataset used for this task/experiment is the WikiANN (NER) dataset. Originally, the dataset consisted of over 200+ languages of articles from Wikipedia with PER, ORG, and LOC entity types. For the sake of this study, we focus on nine Indic languages (Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil and Telugu). Specifically, we chose a balanced version of the WikiANN dataset available on HuggingFace (this has over 170+ languages). In this dataset, training samples per Indic language vary. However, the test and validation sets are properly balanced [20]. The dataset distribution is shown in Table 1.

Table 1: WikiANN dataset distribution (with chosen Indic Languages)

Language	Train	Test	Validation
Bengali	10000	1000	1000
Gujarati	100	100	100
Hindi	5000	1000	1000

Kannada	100	100	100
Malayalam	10000	1000	1000
Marathi	5000	1000	1000
Punjabi	100	100	100
Tamil	15000	1000	1000
Telugu	1000	1000	1000

The dataset is imbalanced based on the training set across languages. One can see that the sample count for specific languages (e.g., Bengali, Tamil, Malayalam, and Hindi) is higher than the rest of the languages. In terms of evaluation, we focus on macro average F1 (for model performance across languages) and weighted average F1 score to evaluate the performance. Regarding models, we do not perform any masked language model pre-training on the above languages. The chosen models for fine-tuning and evaluation are DistilBERT and XLM-RoBERTa. DistilBERT [21] is a smaller and more efficient version of BERT, which is a multilingual model. XLM-RoBERTa is a large multilingual model [22]. These two models support all the nine chosen indic languages (for this study). XLM-RoBERTa was initially trained in mock language modeling. However, it can be fine-tuned for text classification and NER tasks. In this study, DistilBERT and XLM-RoBERTa models were trained with five epochs (starting with original weights). Fine-tuning was performed on A100 GPUs.

Table 2: Language-wise Macro F1 performance of DistilBERT and XLM-RoBERTa

Language	DistilBERT	XLM-RoBERTa
Bengali	0.95	0.97
Gujarati	0.57	0.77
Hindi	0.88	0.90
Kannada	0.60	0.72
Malayalam	0.84	0.86
Marathi	0.86	0.89
Punjabi	0.72	0.86
Tamil	0.87	0.88
Telugu	0.78	0.82
Average	0.78	0.85

DistilBERT performs very efficiently across all languages (except Gujarati and Kannada, which is primarily due to a lack of training data). XLMRobBERTa is remarkably robust across all languages despite fewer samples. This is significantly better than DistilBERT (language-wise performance) in Table 2. The average performance of XLMRoBERTa is around 0.85 (macro F1) compared to 0.78 for DistilBERT.

In Table 3, the performance of DistilBERT and XLM-RoBERTa is represented across all languages in terms of entity (PER, LOC, and ORG). The performance of LOC and PER is quite decent. However, the performance of the ORG entity lags in both DistilBERT and XLM-RoBERTa.

Table 3: Entity-wise performance of F1 score across models

Entity	DistilBERT	XLM-RoBERTa
LOC	0.82	0.83
ORG	0.69	0.68
PER	0.85	0.87
Avg	0.78	**0.79**

4. Conclusion

The performance evaluation of DistilBERT and XLMRoBERTa was done on the Indic subset of the WikiANN dataset. XLMRobeta performs the best overall in terms of entity-wise comparison and language-wise comparison. DistilBERT is very close to XLMRoberta in terms of performance. In terms of overall balanced

performance (memory footprint or efficiency vs. accuracy), DistilBERT is a good choice for NER tasks as it is relatively faster in terms of inference and faster to train compared to XLMRoBERTa.

References

[1]. Singh, A., Khaitan, A., Suresh, R., & Pilla, A. S. (2017). Indian languages–Defining India's Internet.

[2]. Bikel, D. M., Schwartz, R., & Weischedel, R. M. (1999). An algorithm that learns what's in a name. Machine learning, 34, 211-231.

[3]. Morwal, S., Jahan, N., & Chopra, D. (2012). Named entity recognition using hidden Markov model (HMM). International Journal on Natural Language Computing (IJNLC) Vol, 1.

[4]. Zhao, S. (2004). Named entity recognition in biomedical texts using an HMM model. In Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP) (pp. 87-90).

[5]. Isozaki, H., & Kazawa, H. (2002). Efficient support vector classifiers for named entity recognition. In COLING 2002: The 19th International Conference on Computational Linguistics.

[6]. Mayfield, J., McNamee, P., & Piatko, C. (2003). Named entity recognition using hundreds of thousands of features. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 (pp. 184-187).

[7]. Das, A., & Garain, U. (2014). Crf-based named entity recognition@ icon 2013. arXiv preprint arXiv:1409.8008.

[8]. Sobhana, N., Mitra, P., & Ghosh, S. K. (2010). Conditional random field based named entity recognition in geological text. International Journal of Computer Applications, 1(3), 143-147.

[9]. Chowdhury, S., Dong, X., Qian, L., Li, X., Guan, Y., Yang, J., & Yu, Q. (2018). A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. BMC bioinformatics, 19, 75-84.

[10]. Liu, Z., Yang, M., Wang, X., Chen, Q., Tang, B., Wang, Z., & Xu, H. (2017). Entity recognition from clinical texts via recurrent neural network. BMC medical informatics and decision making, 17, 53-61.

[11]. Chiu, J. P., & Nichols, E. (2016). Named entity recognition with bidirectional LSTM-CNNs. Transactions of the association for computational linguistics, 4, 357-370.

[12]. Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[13]. Hakala, K., & Pyysalo, S. (2019, November). Biomedical named entity recognition with multilingual BERT. In Proceedings of the 5th workshop on BioNLP open shared tasks (pp. 56-61).

[14]. Arkhipov, M., Trofimova, M., Kuratov, Y., & Sorokin, A. (2019, August). Tuning multilingual transformers for language-specific named entity recognition. In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing (pp. 89-93).

[15]. Litake, O., Sabane, M., Patil, P., Ranade, A., & Joshi, R. (2022). Mono vs multilingual BERT: A case study in hindi and marathi named entity recognition. arXiv preprint arXiv:2203.12907.

[16]. Dhamecha, T. I., Murthy V, R., Bharadwaj, S., Sankaranarayanan, K., & Bhattacharyya, P. (2021). Role of language relatedness in multilingual fine-tuning of language models: A case study in indo-aryan languages. arXiv preprint arXiv:2109.10534.

[17]. Bahad, S., Mishra, P., Arora, K., Balabantaray, R. C., Sharma, D. M., & Krishnamurthy, P. (2024). Fine-tuning Pre-trained Named Entity Recognition Models For Indian Languages. arXiv preprint arXiv:2405.04829.

[18]. Murthy, R., Bhattacharjee, P., Sharnagat, R., Khatri, J., Kanojia, D., & Bhattacharyya, P. (2022). Hiner: A large hindi named entity recognition dataset. arXiv preprint arXiv:2204.13743.

[19]. Mhaske, A., Kedia, H., Doddapaneni, S., Khapra, M. M., Kumar, P., Murthy V, R., & Kunchukuttan, A. (2022). Naamapadam: a large-scale named entity annotated data for Indic languages. arXiv preprint arXiv:2212.10168.

[20]. WikiANN HuggingFace. https://huggingface.co/datasets/unimelb-nlp/wikiann

[21]. Sanh, V. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[22]. Conneau, A. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.