L3Cube-MahaNER: A Marathi Named Entity Recognition Dataset and BERT models

Anonymous ACL submission

Abstract

Named Entity Recognition (NER) is a basic NLP task and finds major applications in conversational and search systems. It helps us identify key entities in a sentence useful for the downstream application. NER or similar slot filling systems for popular languages have been heavily used in commercial applications. In this work, we focus on Marathi, an Indian language, spoken prominently by the people of Maharashtra state. Marathi is a low resource language and still lacks useful NER resources. We present L3Cube-MahaNER, the first major gold standard named entity recognition dataset in Marathi. We also describe the manual annotation guidelines followed during the process. In the end, we also benchmark the dataset on different CNN, LSTM, and Transformer based models.

1 Introduction

A principal technique of information extraction is Named Entity Recognition. It is an integral part of natural language processing systems. The technique involves the identification and categorisation of the named entity. These categories include entities like people's names, locations, numerical values and temporal values. NER has a myriad of applications like customer service, text summarization etc. Through the years, a large amount of work has been done for Named Entity Recognition in the English language. Through the years, a large amount of work has been done for NER in the English language. The work is very mature and the functionality comes out of the box with NLP libraries like NLTK(Bird et al., 2009) and spacy(Honnibal and Montani, 2017). In contrast, limited work is done in the Indic languages like Hindi and Marathi. (Patil et al., 2016) addresses the problems faced by Indian languages like the presence of abbreviations, ambiguities in named entity categories, different dialects, spelling variations and the presence of foreign words. (Shah, 2016) elaborates on these issues along with others like the lack of well-annotated data, fewer resources and tools etc. Furthermore, the existing resources for NER in Marathi like the released by (Murthy et al., 2018) titled IIT Bombay Marathi NER Corpus has various limitations like the presence of foreign language words and the existence of about 39 percent of sentences with O tags only. Apart from that, many datasets aren't available publicly or contain fewer sample sentences.

In this paper, we present our dataset L3Cube-MahaNER. This dataset has been compiled in-house at L3Cube. It is the largest publicly available dataset for Marathi NER. We have annotated the dataset manually in order to contribute to the resources available for NER on Marathi. It contains 25,000 manually tagged sentences categorized according to the entity classes. These entities annotated in the dataset include names of locations, organizations, people and numeric quantities like time, measure and other entities like dates and designations. The paper also depicts the dataset statistics and the guidelines that have been followed while tagging these sentences.

In this paper, we also present the results of deep-learning models like CNN, LSTM, BiLSTM transformers like mBERT, IndicBert(Kakwani et al., 2020), xlm-Roberta, Roberta-Marathi, MahaBERT, MahaROBERTa, MahaALBERTA and character-based models that have been trained on the L3Cube-MahaNER dataset. We experiment on all major multi-lingual and Marathi BERT models to establish a benchmark for future comparisons.

2 Related Work

Named Entity Recognition is a concept that originated at the Message Understanding Conferences (Grishman and Sundheim, 1996) in 1995. Machine learning techniques and linguistic techniques were the two major techniques used to perform NER. Handmade rules (Abdallah et al., 2012) developed by experienced linguists were used in the linguistic techniques. These systems, which included gazetteers, dictionaries, and lexicalized grammar, demonstrated good accuracy levels in English. However, these strategies had the disadvantage of being difficult to transfer to other languages or professions. Decision Trees (Paliouras et al., 2000), Conditional Random Field, Maximum Entropy Model (Bender et al., 2003), Hidden Markov Model, Support Vector Machine were included in machine learning techniques. To attain better competence, these supervised learning algorithms make use of massive volumes of NE annotated data.

A comparative study by training the models on the same data using Support Vector Machine (SVM) and Conditional Random Field(CRF) was carried out by (Krishnarao et al., 2009). It was concluded that the CRF model was superior. A more effective hybrid system consisting of the Hidden Markov Model, a combination of handmade rules and MaxEnt was introduced by (Srihari, 2000) for performing NER. Deep learning models were then utilized to complete the NER problem as technology progressed.Convolutional Neural Network (CNN) (Albawi et al., 2017), Long-Short Time Memory(LSTM) (Hochreiter and Schmidhuber, 1997), Bi-directional Long-Short Time Memory(BiLSTM) (Yang and Xu, 2020), Transformers were among the most popular models.

NER for Indian languages is a comparatively difficult task due to a lack of capitalization, spelling variances, and uncertainty in the meaning of words. The structure of the language is likewise difficult to grasp. Furthermore, the lack of a well-ordered labelled dataset makes advanced approaches such as deep learning methods difficult to deploy. (Bhattacharjee et al., 2019) has described various problems faced while implementing NER for Indian languages.

(Murthy et al., 2018) in 2018 introduced Marathi annotated dataset named IIT Bombay Marathi NER Corpus for Named Entity Recognition consisting of 5591 sentences and X tags. They considered 3 main categories named Location, Person, Organization for training character-based model on the dataset. They made use of multilingual learning to jointly train models for multiple languages, which in turn helps in improving the NER performance of one of the languages.

(Pan et al., 2017) in 2017 released a dataset named WikiAnn NER Corpus consisting of 14,978 sentences and 3 tags labelled namely Organization, Person and Location. It is a however a silver-standard dataset for 282 different languages including Marathi. This project aims to create a cross-lingual name tagging and linking framework for Wikipedia's 282 languages.

3 Compilation of dataset

3.1 Data Collection

Our dataset consists of 25,000 sentences in the Marathi language. We have used the base sentences from the L3Cube-MahaCorpus, which is a monolingual Marathi dataset created inhouse by L3Cube.

The sentences in the dataset are in the Marathi language with minimal appearance of English words and numerics as present in the original news. However, while annotating the dataset, these English words have not been considered as a part of the named entity categories. Furthermore, the dataset does not preserve the context of the news, such as the publicating profiles, regions, and so on.

3.2 Dataset Annotation

We have manually tagged the entire dataset into eight named entity classes. These classes include Person (NEP), Location(NEL), Organization(NEO), Measure(NEM), Time(NETI), Date(NED), and Designation(ED). While tagging the sentences, we established an annotation guideline to ensure consistency. Firstly, the sentences were relieved of any contextual associations. Then, the approach for the contents of the named entity classes was decided as follows. Proper nouns involving persons' names are tagged as NEP and places are tagged as NEL. All kinds of organizations like companies, councils, political parties and government departments are tagged as NEO. Numeric quantities of all kinds are tagged as NEM with respect to the context. Furthermore, temporal values like time are tagged as NETI and dates are tagged as NED. Apart from that, individual titles and designations, which precede proper nouns in the sentences are tagged as ED. Despite maintaining these guidelines, some entities had ambiguous meanings and were difficult to tag. In these circumstances, we resolved the intricacies unanimously.

3.3 Dataset Statistics

Dataset	Sentence Count	Tag Count
Train	21500	26502
Test	2000	2424
Tune	1500	1800

Table 1: Count of sentences and tags in the dataset.

Tags	Train	Test	Tune
NEM	7263	651	512
NEP	6856	594	450
NEL	4946	450	321
NEO	3476	324	227
NED	2363	240	174
ED	934	90	70
NETI	569	58	39

Table 2: Count of individual tags of L3Cube-MahaNER.

For more clarity, some example sentences with tagged entities are mentioned in Table 4.

4 Experimental Techniques

4.1 Model Architectures

In deep learning, the models are trained using large datasets consisting of labeled data and neural network topologies that learn features from the data effectively, without the need for feature extraction to be done manually.

Similarly, the transformer aims to address sequence-to-sequence problems while also resolving long-range relationships in natural language processing. The transformer model contains a ”self attention” mechanism that examines the relationship between all of the words in a phrase. It provides differential weightings to indicate which phrase components are most significant in determining how a word should be read. Thus the transformer identifies the context that assigns each word in the sentence its meaning. The training time also is lowered as the feature enhances parallelization.

CNN: A single 1D convolution is used to pass 300-dimensional word embeddings. These embeddings are fed into a Conv1D layer having 512 filters. A single dense layer of size 8 is subjected as output each time. The activation function used is relu. All the models have the same optimizer and loss functions. The optimizer used is RMSPROP. We have experimented with Indic fastText embeddings[a] and embedding layers with random initializations.

LSTM: We have used a basic LSTM model in which we have passed the word embeddings of 300 dimensions. Along with that, a single Bi-Lstm having 512 hidden units is used. A single dense layer of size 8 is subjected as output each time. Additionally, experiments using embeddings mentioned in the CNN section are also performed.

BiLSTM: It is analogous to the CNN model with the single 1D convolution substituted by a Bi-LSTM layer. An embedding vector of dimension 300 is used in this model. Along with that, a batch size of 16 is used. Additionally, the experiments performed with the embeddings in the previous section are executed here as well.

BERT: BERT(Devlin et al., 2019) is a Google-developed transformer-based approach for NLP pre-training that was inspired by pre-training contextual representations. It's a deep bidirectional model, which means it's trained on both sides of a token's context. BERT's most notable feature is that it can be fine-tuned by adding a few output layers.

Model	**F1**	**Precision**	**Recall**	**Accuracy**
mBERT	85.3	82.83	97.94	96.92
Indic BERT	86.56	85.86	87.27	97.15
Xlm-Roberta	85.69	84.21	87.22	97.07
Roberta-Marathi	83.86	82.22	85.57	96.92
MahaBERT	**86.80**	84.62	89.09	97.15
MahaRoBERTa	86.60	84.30	89.04	97.24
MahaAlBERT	85.96	84.32	87.66	97.32
CNN	79.5	82.1	77.4	97.28
LSTM	74.9	84.1	68.5	94.89
BILSTM	**80.4**	83.3	77.6	94.99

Table 3: F1 score(macro), precision and recall of various transformer and normal models using the Marathi dataset.

mBert: mBERT(Pires et al., 2019), which stands for multilingual BERT, is the next step in constructing models that understand the meaning of words in context. A deep learning model was built on 104 languages by concurrently encoding all of their information on mBERT.

ALBERT: ALBERT(Lan et al., 2020) is a transformer design based on BERT that requires many less parameters than the current state-of-the-art model BERT. These models can train around 1.7 times quicker than BERT models and have greater data throughput than BERT models. IndicBERT is a multilingual ALBERT model that includes 12 main Indian languages and was trained on large-scale datasets. Many public models, such as mBERT and XLM-R, have more parameters than IndicBERT, although the latter performs exceptionally well on a wide range of tasks.

RoBERTa: RoBERTa(Liu et al., 2019) is an unsupervised transformers model that has been trained on a huge corpus of English data. This means it was trained exclusively on raw texts, with no human labelling, and then utilised an automated approach to generate labels and inputs from those texts. The multilingual model XLM-RoBERTa has been trained in 100 languages. Unlike certain XLM multilingual models, it does not require lang tensors to detect which language is being used. It can also deduce the correct language from the supplied ids.

5 Results

In this study, we have experimented with various model architectures like CNN, LSTM, BiLSTM, and transformers like BERT, Roberta to perform named entity recognition on our dataset. This section presents the F1 scores attained by training these models on our dataset. These results have been reported in Table 3. Among the CNN and LSTM based models, the BiLSTM model with the trainable word embeddings gives the best results on the L3Cube-MahaNER dataset. Moreover, the MahaBERT model, which has mBERT as its base architecture, yields the best results amongst transformers. The LSTM and the Roberta-Marathi models report the lowest scores among all models. In general, the monolingual Marathi BERT models based on MahaBERT perform slightly better than their multi-lingual counterparts. The results also show the importance of using subword-based approaches over word-based models.

6 Conclusion

In this paper, we hold forth to the problem of scarcity of annotated corpora and hence present L3Cube-MahaNER which is by far the largest dataset for Marathi Named Entity recognition, containing 25000 distinct sentences. We achieved results using deep learning models such as CNN, LSTM, BiLSTM, and transformers in BERT as listed above, to set the basis for future work. We observed the highest scores on MahaBERT and BiLSTM for our dataset. We believe that our corpus will

Sentence	Tag
कोलकाता आणि दक्षिण भारतातूनही सुपारी नागपुरात येत	NEL O NEL NEL O NEL O
या हल्ल्यात काश्मीर पोलिसांच्या एका जवानाने तर सीआरपीएफच्या दोन जवानांनी आपले प्राण गमावलेत	O O NEO NEO NEM O O NEO NEM O O O O
दरम्यान राज्यातील सरकारच्या स्थैर्यावर नारायण राणे यांनी याआधीही प्रश्नचिन्ह उपस्थित केलं होतं	O O O O NEP NEP O O O O O O
विरोधी पक्षनेते देवेंद्र फडणवीस यांनीही हे सरकार अंतर्विरोधातून कोसळेल असा दावा केला आहे	O ED NEP NEP O O O O O O O O O

Table 4: Sample Tagged Sentences

play a pivotal role in expanding conversational AI for the Marathi Language.

References

Sherief Abdallah, Khaled Shaalan, and Muhammad Shoaib. 2012. Integrating rule-based system with classification for arabic named entity recognition. volume 7181, pages 311–322.

Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6.

Oliver Bender, Franz Josef Och, and Hermann Ney. 2003. Maximum entropy models for named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 148–151.

Krishnanjan Bhattacharjee, Shiva Karthik S, Swati Mehta, Ajai Kumar, Ria Mehta, Dweep Pandya, Pratik Chaudhari, and Devika Verma. 2019. Named entity recognition: A survey for indian languages. In *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, volume 1, pages 217–220.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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 the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Awaghad Ashish Krishnarao, Himanshu Gahlot, Amit Srinet, and D. S. Kushwaha. 2009. A comparative study of named entity recognition for hindi using sequential learning algorithms. In *2009 IEEE International Advance Computing Conference*, pages 1164–1169.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

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

Rudra Murthy, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2018. Judicious selection of training data in assisting language for multilingual neural NER. In *Proceedings of the 56th*

Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 401–406.

Georgios Paliouras, Vangelis Karkaletsis, Georgios Petasis, and Constantine D. Spyropoulos. 2000. Learning decision trees for named-entity recognition and classification. In *In ECAI Workshop on Machine Learning for Information Extraction*.

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.

Nita Patil, Ajay Patil, and Pawar B.V. 2016. Issues and challenges in marathi named entity recognition. *International Journal on Natural Language Computing*, 5.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert?

Hinal Shah. 2016. Study of named entity recognition for indian languages. *International Journal of Information Sciences and Techniques*, 6.

Rohini Srihari. 2000. A hybrid approach for named entity and sub-type tagging. In *Sixth Applied Natural Language Processing Conference*, pages 247–254, Seattle, Washington, USA. Association for Computational Linguistics.

Gang Yang and Hongzhe Xu. 2020. A residual bilstm model for named entity recognition. *IEEE Access*, 8:227710–227718.