Proceedings of the WILDRE-6 Workshop @LREC2020, pages 97–101
Marseille, 20 June 2022

L3Cube-MahaCorpus and MahaBERT: Marathi Monolingual Corpus, Marathi BERT Language Models, and Resources

Raviraj Joshi
Indian Institute of Technology Madras, Tamilnadu, India
L3Cube Pune, Maharashtra, India
ravirajoshi@gmail.com

Abstract

We present L3Cube-MahaCorpus a Marathi monolingual data set scraped from different internet sources. We expand the existing Marathi monolingual corpus with 24.8M sentences and 289M tokens. We further present, MahaBERT, MahaAlBERT, and MahaRoBerta all BERT-based masked language models, and MahaFT, the fast text word embeddings both trained on full Marathi corpus with 752M tokens. We show the effectiveness of these resources on downstream Marathi sentiment analysis, text classification, and named entity recognition (NER) tasks. We also release MahaGPT, a generative Marathi GPT model trained on Marathi corpus. Marathi is a popular language in India but still lacks these resources. This work is a step forward in building open resources for the Marathi language. The data and models are available at https://github.com/l3cube-pune/MarathiNLP.

Keywords: Marathi Monolingual Corpus, Deep Learning, Marathi NLP, Transformers, Marathi BERT, Marathi Word Embeddings, Text Classification, NER, GPT

1. Introduction

Pre-trained language models based on BERT have been widely used in NLP applications (Wolf et al., 2019; Qiu et al., 2020). These language models are fine-tuned on the target task and are reported to provide superior results. The target tasks include text classification, named entity recognition (NER), parts of speech (POS) tagging, dependency parsing, natural language inference (NLI), etc (Otter et al., 2020). The BERT-based models can be trained using un-supervised large text corpus using masked language modeling objective and next sentence prediction tasks.

The mono-lingual and multi-lingual masked language models have been very popular recently. The multilingual language models provide significant benefits for low resource languages by leveraging the learning from high resource text (Pires et al., 2019). However, models trained on a single language are shown to perform better than multi-lingual models on target tasks in corresponding language (Straka et al., 2021). Previous works have built BERT based language models in German, Vietnamese, Arabic, Dutch, French, Hindi, Bengali, etc (Scheible et al., 2020; Nguyen and Nguyen, 2020; Le et al., 2019; Delobelle et al., 2020; Abdul-Mageed et al., 2020; Jain et al., 2020). In this work, we focus on building monolingual corpus and BERT based language model in Marathi. Marathi is a low-resource Indian language and is native to the state of Maharashtra.

Marathi is the third most popular language in India after Hindi and Bengali (Kulkarni et al., 2021a; Joshi et al., 2019). It is spoken by around 83 million people in India. Despite huge representation, in terms of speaking diaspora, the language resources have not received adequate attention for the Marathi language. The language resource in the simplest form is a monolingual corpus. However, even monolingual corpus for Indian languages is mostly biased towards Hindi. This can be seen from the fact that the recently released IndicNLP data set has 62.9M Hindi sentences and only 9.9M Marathi sentences (Kakwani et al., 2020). There is a strong need to develop language resources for Marathi starting from building a monolingual corpus.

In this work, we add to the existing monolingual corpus by building L3Cube-MahaCorpus[1]. The data has been scraped from various internet sources. The corpus for Indian languages has mostly been exclusively dominated by news sources. We specifically consider this bias and also include sentences from non-news sources. L3Cube-MahaCorpus adds 24.8M sentences and 289M tokens (5.3 GB) to the existing Marathi monolingual datasets. After combining this with the existing Marathi corpus there is a total of 57.2M sentences and 752M tokens (13 GB).

We further introduce MahaBERT[2], MahaRoBERTa[3], MahaAlBERT[4][5], and MahaGPT[6] all Transformer BERT based Marathi language models trained on the full Marathi monolingual corpus. The BERT models are trained using masked language modeling objectives. These models are further evaluated on downstream tasks of text classification and named entity recognition (NER) in Marathi. We also release MahaFT, the fast text word embedding trained on the full Marathi Corpus. The dataset and resources are publicly shared to facilitate further research in Marathi NLP. The main contributions of this work are:

- We present L3Cube-MahaCorpus, a Marathi

[1] https://github.com/l3cube-pune/MarathiNLP
[2] https://huggingface.co/l3cube-pune/marathi-bert
[3] https://huggingface.co/l3cube-pune/marathi-roberta
[4] https://huggingface.co/l3cube-pune/marathi-albert
[5] https://huggingface.co/l3cube-pune/marathi-albert-v2
[6] https://huggingface.co/l3cube-pune/marathi-gpt

monolingual corpus with 24.8M sentences and 289M tokens.

- We introduce MahaBERT, MahaAlBERT, and MahaRoBERTa, the BERT variations trained on a full corpus with 752M tokens. We also release MahaGPT, a Marathi generative pre-trained transformer model trained on the full corpus.
- Finally, we release MahaFT, Marathi fast text embeddings trained on the full corpus.

2. Related Work

In this section, we review different unsupervised and supervised data sets in the Marathi language. A summary of publicly available Marathi monolingual corpus and classification data sets is provided in Kulkarni et al. (2021a). The main sources include Wikipedia text, CC-100 Dataset (Wenzek et al., 2019), OSCAR Corpus (Suárez et al., 2019), and IndicNLP Corpus (Kakwani et al., 2020). The wiki dataset consists of 85k cleaned Marathi articles. The other sources are multi-lingual datasets with Marathi as one of the languages. The CC-100 monolingual data set consists of around 50 million tokens for the Marathi language. The OSCAR corpus consists of around 82 million tokens in Marathi. The IndicNLP is perhaps the largest non-wiki source and consists of 142 million tokens.

There are limited resources for supervised tasks in Marathi. The text classification data set includes IndicNLP News Article Dataset (Kakwani et al., 2020), iNLTK Headline Dataset (Arora, 2020), L3CubeMahaSent (Kulkarni et al., 2021b). The IndicNLP News Article Dataset is a news article classification dataset in Marathi consisting of 4779 records. The iNLTK Headline Dataset categorizes news headlines and consists of 12092 records. The L3CubeMahaSent is a sentiment classification dataset in Marathi and consists of 16000 records. Another data set for Marathi NER was introduced in (Murthy et al., 2018). It consists of 5591 sentences and 3 named entities as target labels. Moreover, some hate speech detection datasets have also been released in Marathi (Gaikwad et al., 2021; Mandl et al., 2021; Pawar and Raje, 2019).

In this work, we have utilized all publicly available Marathi monolingual corpus along with the L3Cube-MahaCorpus to train the language models and word embedding. These models are evaluated on the three classification tasks and a NER task.

3. Curation of Dataset

The L3Cube-MahaCorpus is collected from news and non-news sources. The major chunk of the data is scraped from the Maharashtra Times website [7]. The non-news sources were taken from a collection website [8]. The data set was scraped using the BeautifulSoup library along with the use of Selenium for dynamic pages. The final data set was shuffled and de-duplicated. The de-duplication was also performed with the existing monolingual data set. The L3Cube-MahaCorpus adds 17.6 M sentences (212 M tokens) from the news sources and 7.2 M sentences (76.4 M tokens) from the non-news sources. These are made available separately as well. Overall it adds 24.8 M sentences and 289 M tokens. When combined with the existing monolingual dataset, we now have 57.2 M sentences and 752 M tokens in the Marathi language. These statistics are also described in Table 1.

Dataset	#tokens	#sentences
L3Cube-MahaCorpus (News)	212	17.6
L3Cube-MahaCorpus (Non-news)	76.4	7.2
L3Cube-MahaCorpus	289	24.8
Full Marathi Corpus	752	57.2

Table 1: Dataset Statistics (in millions).

4. Pre-trained Resources

The full Marathi monolingual corpus is used to train Transformer based masked language models and FastText word embeddings.

4.1. Transformer Models

The BERT represents a deep bi-directional Transformer based model trained using a large unlabelled corpus. These pre-trained models have been shown to produce state-of-the-art results on a variety of downstream tasks. There are different variations of BERT models like AlBERT and RoBERTa which are also considered in this work. From the multilingual perspective, there are three main models which can also be used with the Marathi language. These include multilingual-BERT (Devlin et al., 2019), XLM-R based on RoBERTa (Conneau et al., 2019), and IndicBERT (Kakwani et al., 2020) based on AlBERT. These three models are fine-tuned on monolingual Marathi corpus and released as a part of this work. All the models are trained for 2 epochs with standard hyper-parameters and masked language modeling objective only. The learning rate used is 2e-5 with a batch size of 64.

- mBERT[9]: It is a BERT-base vanilla model pre-trained on 104 languages using masked language modeling (MLM) and next sentence prediction (NSP) objective. The Marathi was one of the languages used in pre-training.
- XLM-RoBERTa[10]: It is a RoBERTa based model pre-trained on 100 languages using MLM objec-

[7] https://maharashtratimes.com/

[8] http://www.netshika.com/sangrah.html

[9] https://huggingface.co/bert-base-multilingual-cased

[10] https://huggingface.co/docs/transformers/model_doc/xlmroberta

Model	L3CubeMahaSent	News Articles	News Headlines	Marathi NER
mBERT	80.4	97.6	90.6	58.35
indicBERT	83.3	98.7	93.7	60.79
XLM-R	82.0	98.5	92.5	62.32
MahaBERT	82.8	98.7	94.4	62.57
MahaAlBERT	**83.7**	**99.1**	**94.7**	60.00
MahaRoBERTa	83.4	98.5	94.2	**64.34**
FB-FT + KNN	73.6	99.1	88.8	-
INLP-FT + KNN	74.9	98.9	90.7	-
MahaFT + KNN	75.1	98.9	91.2	-

Table 2: The results for different models on classification and NER tasks. The numbers for classification task L3CubeMahaSent, News Articles, and News Headlines represent the classification accuracy. The numbers for the Marathi NER task represent the macro-f1 score. The FB-FT is Marathi fast text embeddings trained on Wiki and Common Crawl Corpus released by Facebook used along with KNN(k=4). The INLP-FT represents the Marathi fast text embeddings released by IndicNLP Suite. The MahaFT are Marathi fast text embeddings released as a part of this work.

tive. The model is shown to outperform mBERT on different tasks. Even this model contains Marathi as one of the pre-training languages. The RoBERTa mainly modifies the hyper-parameters used in the original BERT and gets rid of the NSP task (Liu et al., 2019).

- IndicBERT[11]: It is a multi-lingual AlBERT model exclusively pre-trained on 12 Indian languages. The AlBERT is a lite version of the BERT model (Lan et al., 2019). It uses parameter reduction techniques like repeated layers to reduce the memory footprint. The model has been shown to work well on most of the Indic NLP tasks (Joshi et al., 2021; Kulkarni et al., 2021b; Velankar et al., 2021; Nayak and Joshi, 2021).

4.2. FastText Word Embeddings

Pre-trained word embeddings are commonly used to initialize the embedding layer of the neural networks. These distributed representations are trained on large unlabeled corpus and are useful for many downstream tasks. The FastText word embeddings are popular for morphologically rich languages (Bojanowski et al., 2017). It represents the word as a bag of character n-grams thus avoiding any out of vocabulary word. We train the FastText model on the Marathi monolingual corpus using standard hyper-parameters. A skip-gram model is trained with a window size of 5, 10 negative samples per instance, and 10 epochs.

4.3. Marathi GPT

GPT2 is a generative transformer model trained using causal language modeling (CLM) objective (Radford et al., 2019). It is also a class of self-supervised models trained to predict the next work on the unsupervised data. We train a standard GPT2 model with 12 layers and 768 internal dimension on Marathi Corpus for 5 epochs with a learning rate of 2e-5. We use a custom BPE-based tokenizer with a vocab size of 50257.

[11] https://huggingface.co/ai4bharat/indic-bert

5. Down Stream Tasks

- **IndicNLP News Article Classification**: The task consists of Marathi news articles classified as sports, entertainment, and lifestyle. There are 3823 train, 479 test, and 477 validation examples.

- **iNLTK Headline Classification**: In this classification task the Marathi news headlines are categorized as entertainment, sports, and state. The dataset consists of 9672 train, 1210 test, and 1210 validation examples.

- **L3CubeMahaSent Sentiment Analysis**: The sentiment analysis task consists of Marathi tweets categorized as positive, negative, and neutral. The dataset consists of 12114 train, 2250 test, and 1500 validation examples.

- **Marathi Named Entity Recognition**: This is a Marathi entity recognition task where each token in the sentence is categorized as Location, Person, and Organization. The dataset consists of 3588 train, 1533 test, and 470 validation examples.

5.1. Results

The L3Cube-MahaCorpus along with other publicly available Marathi corpus is used to train three variations of BERT using MLM objective. These variations are based on base-BERT, AlBERT, and RoBERTa architecture and are termed as MahaBERT, MahaAlBERT, and MahaRoBERTa respectively. The multilingual versions of these architectures mBERT, indicBERT based on AlBERT, and XLM-R based on RoBERTa are also used for baseline comparison. The

multilingual versions are fine-tuned on the Marathi corpus to get the Marathi BERT models. Similar hyperparameters are used for MLM pre-training of all these models. The results are described in Table 2. Note that the results for base models may be slightly different than ones reported in the original work as they were re-computed using a common setup and hyperparameters. These models are evaluated on three classification datasets and one named entity recognition dataset. For the classification task, the pre-trained models are further fine-tuned by the addition of a dense layer on top of [CLS] token embedding. The NER task is formulated as a token classification task and all token embeddings are passed through the dense layer for classification. Overall the monolingual versions of models perform better than the multi-lingual versions.
The fast text word embeddings trained on full Marathi corpus termed as MahaFT are evaluated on the classification datasets. In this setup, word embeddings are averaged to get the sentence representation. A KNN classifier with k=4 is used for the classification of the averaged fast text embedding. These Marathi word embeddings are compared against two other publicly available variations. The FB-FT represents Marathi fast text embeddings trained on Wiki and Common Crawl Corpus released by Facebook. The INLP-FT was released as part of IndicNLP suite. The MahaFT performs competitively with other word embeddings. Overall we show the resources released as a part of this work either perform competitively with or better than the currently available alternatives for the Marathi language.

6. Conclusion

In this paper, we have presented L3Cube-MahaCorpus, MahaBERT, and MahaFT. The MahaCorpus, is a Marathi monolingual corpus and is a significant addition to the existing monolingual corpus. The Marathi BERT is trained in three different flavors namely MahaBERT, MahaRoBERTa, and MahaAlBERT. The MahaFT is the Marathi fast text word embeddings. These resources are exclusively trained on Marathi monolingual corpus. The models are evaluated on downstream Marathi classification and NER tasks. The models are shown to work better than their multi-lingual counterparts.

Acknowledgments

Multiple L3Cube Pune, student groups have contributed to this work. We would like to thank Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, and Gayatri Kshirsagar for their contribution. We also thank groups Algorithm_Unlock and Bits_To_Bytes for their support.

7. Bibliographical References

Abdul-Mageed, M., Elmadany, A., and Nagoudi, E. M. B. (2020). Arbert & marbert: deep bidirectional transformers for arabic. *arXiv preprint arXiv:2101.01785*.

Arora, G. (2020). inltk: Natural language toolkit for indic languages. *arXiv preprint arXiv:2009.12534*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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

Delobelle, P., Winters, T., and Berendt, B. (2020). Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

Devlin, J., Chang, M.-W., Lee, K., and 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)*, pages 4171–4186.

Gaikwad, S. S., Ranasinghe, T., Zampieri, M., and Homan, C. (2021). Cross-lingual offensive language identification for low resource languages: The case of marathi. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 437–443.

Jain, K., Deshpande, A., Shridhar, K., Laumann, F., and Dash, A. (2020). Indic-transformers: An analysis of transformer language models for indian languages. *arXiv preprint arXiv:2011.02323*.

Joshi, R., Goel, P., and Joshi, R. (2019). Deep learning for hindi text classification: A comparison. In *International Conference on Intelligent Human Computer Interaction*, pages 94–101. Springer.

Joshi, R., Karnavat, R., Jirapure, K., and Joshi, R. (2021). Evaluation of deep learning models for hostility detection in hindi text. In *2021 6th International Conference for Convergence in Technology (I2CT)*, pages 1–5. IEEE.

Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N., Bhattacharyya, A., Khapra, M. M., and Kumar, P. (2020). inlpsuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4948–4961.

Kulkarni, A., Mandhane, M., Likhitkar, M., Kshirsagar, G., Jagdale, J., and Joshi, R. (2021a). Experimental evaluation of deep learning models for marathi text classification. *arXiv preprint arXiv:2101.04899*.

Kulkarni, A., Mandhane, M., Likhitkar, M., Kshirsagar, G., and Joshi, R. (2021b). L3cubemahasent: A marathi tweet-based sentiment analysis dataset. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 213–220.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2019). Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

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

Mandl, T., Modha, S., Shahi, G. K., Madhu, H., Satapara, S., Majumder, P., Schaefer, J., Ranasinghe, T., Zampieri, M., Nandini, D., et al. (2021). Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indoaryan languages. *arXiv preprint arXiv:2112.09301*.

Murthy, R., Kunchukuttan, A., and Bhattacharyya, P. (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.

Nayak, R. and Joshi, R. (2021). Contextual hate speech detection in code mixed text using transformer based approaches. *arXiv preprint arXiv:2110.09338*.

Nguyen, D. Q. and Nguyen, A. T. (2020). Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

Otter, D. W., Medina, J. R., and Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):604–624.

Pawar, R. and Raje, R. R. (2019). Multilingual cyberbullying detection system. In *2019 IEEE International Conference on Electro Information Technology (EIT)*, pages 040–044. IEEE.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., and Boeker, M. (2020). Gottbert: a pure german language model. *arXiv preprint arXiv:2012.02110*.

Straka, M., Náplava, J., Straková, J., and Samuel, D. (2021). Robeczech: Czech roberta, a monolingual contextualized language representation model. *arXiv preprint arXiv:2105.11314*.

Suárez, P. J. O., Sagot, B., and Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Velankar, A., Patil, H., Gore, A., Salunke, S., and Joshi, R. (2021). Hate and offensive speech detection in hindi and marathi. *arXiv preprint arXiv:2110.12200*.

Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. (2019). Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.