BERTODIA: BERT PRE-TRAINING FOR LOW RESOURCE ODIA LANGUAGE

Shantipriya Parida[a] Satya Prakash Biswal
Biranchi Narayan Nayak Mael Fabien
Esaú VILLATORO-TELLO[b] Petr Motlicek

Idiap-RR-16-2021

Version of OCTOBER 29, 2021

[a]Idiap Research Institute
[b]Idiap

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

BertOdia: BERT pre-training for low resource Odia language

Shantipriya Parida[1*], Satya Prakash Biswal[2], Biranchi Narayan Nayak[3], Maël Fabien[1,4], Esaú Villatoro-Tello[1,5], Petr Motlicek[1], and Satya Ranjan Dash[6]

[1] Idiap Research Institute, Martigny Switzerland
{firstname.lastname}@idiap.ch
[2] The University Of Chicago, Chicago, USA
sbiswal@chicagobooth.edu
[3] Capgemini Technology Services India Limited, Bangalore, India
biranchi125@gmail.com
[4] École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
mael.fabien@epfl.com
[5] Universidad Autónoma Metropolitana, Mexico City, Mexico
evillatoro@cua.uam.mx
[6] KIIT University, Bhubaneswar, India
sdashfca@kiit.ac.in

Abstract. Odia language is one of the 30 most spoken languages in the world. It is spoken in the Indian state called Odisha. Odia language lacks online content and resources for natural language processing (NLP) research. There is a great need for a better language model for the low resource Odia language, which can be used for many downstream NLP tasks. In this paper, we introduce a Bert-based language model, pre-trained on 430'000 Odia sentences. We also evaluate the model on the well-known Kaggle Odia news classification dataset (BertOdia: *96*%, RoBERTaOdia: *92*%, and ULMFit: *91.9*% classification accuracy), and perform a comparison study with multilingual Bidirectional Encoder Representations from Transformers (BERT) supporting Odia. The model will be released publicly for the researchers to explore other NLP tasks.

Keywords: Low Resource · BERT · RoBERTa.

1 Introduction

Odia[7] is one of the oldest languages in India as it dates back to the 10th century CE. It originated from Ardhamagadhi Prakrit. This is the mother tongue of around 31 million people living in Odisha state. The term Odia comes from the ancient Sanskrit Odra. The Odrakas are described in Mahabharata as great warriors who fought in that battle. Odia is one of the six Classical languages (Sanskrit, Tamil, Telugu, Kannada, and Malayalam) identified by the Indian

[*] Corresponding author
[7] https://en.wikipedia.org/wiki/Odia_language

Government. Odia language has more than 2 lakh manuscripts documented, which makes it the second-highest manuscript holder among Indian languages.

Odia is agglutinative.[8] It differentiates between plural and singular number; male and female gender; second, first, and third persons. But it does not have gender biasing in verbs, nouns, or pronouns like other languages, which reduces the complexity. Odia language allows compounding but does not allows elision. It has 6 vowels 28 consonants, 9 diphthongs, 0 ending with consonants, and 4 semivowels. Odia's vocabulary is influenced by Sanskrit, and is also a little influenced by Arabic, Persian, Austronesian languages as the Kalinga empire (Odisha's ancient name) was connected to other kingdoms.[9] Odia script is an Abugida, that is written from left to right.

In recent years there is a growing interest in the NLP community using pre-trained language models for various NLP tasks where the models are trained in a semi-supervised fashion to learn a general language model [5]. A better language model is the key component of the automatic speech recognition system (ASR) [19]. Building a language model is a challenging task in the case of low resource languages where the availability of contents is limited [1]. Researchers proposed many techniques for the low resource NLP tasks such as feature engineering, and knowledge transfer across domain [9, 10]. However, these approaches do not use a pre-trained general language model, rather they perform pre-training for each task individually. We focus on building a general language model using the limited resources available in the low resource language which can be useful for many language and speech processing tasks.

Our key contribution includes building a language-specific BERT model for this low resource Odia language and as per our best knowledge, this is the first work in this direction. The overall architecture of the proposed model is shown in Figure 1.

2 Related Work

Low resource languages have been drawing the attention of several recent works in language model pre-training [3]. Although Multilingual Bidirectional Encoder Representations from Transformers (M-BERT)[10] successfully covers 104 languages and is being further extended, a large number of languages are still not covered. It is one of the research trends to extend M-BERT for low resource languages [24]. Some researchers found that language-specific BERT models perform better compared to multilingual BERT models [14].

The sufficient availability of online contents remains one of the major challenges for many low resource languages including Odia [16]. Recently, a few research projects were initiated to build a language model for many low-resource Indian languages including Odia. In particular, the Natural Language Toolkit

[8] https://www.mustgo.com/worldlanguages/oriya/

[9] https://www.nriol.com/indian-languages/oriya-page.asp

[10] https://github.com/google-research/bert/blob/master/multilingual.md

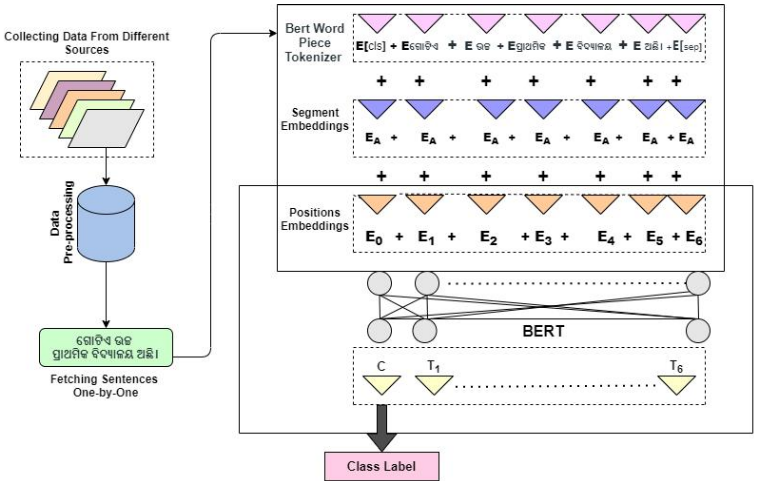

Collecting Data From Different Sources
Data Pre-processing
ଗୋଟିଏ ଉଚ୍ଚ ପ୍ରାଥମିକ ବିଦ୍ୟାଳୟ ଅଛି।
Fetching Sentences One-by-One
Bert Word Piece Tokenizer
E[cls] + Eଗୋଟିଏ + E ଉଚ୍ଚ + Eପ୍ରାଥମିକ + E ବିଦ୍ୟାଳୟ + E ଅଛି। + E[sep]
Segment Embeddings
EA + EA + EA + EA + EA + EA + EA
Positions Embeddings
E0 + E1 + E2 + E3 + E4 + E5 + E6
BERT
C
T1
T6
Class Label

Fig. 1. The Proposed Model: Visualisation of an our experimental model used for the Single Sentence Classification Task with Bert Embedding Layers.

for Indic Languages (iNLTK)[11] released different language models for the Indian languages including Odia using 17K Odia Wikipedia articles, and the model is tested on the classification task using IndicNLP News Article Classification Dataset - Oriya [2]. There is also the multilingual IndicBERT[12] model based on BERT that supports 12 Indian languages including Odia available in Huggingface transformers library. It also has IndicGLUE, a natural language understanding benchmark for the evaluation of a few tasks for Odia [8].

3 Data Source

We have collected monolingual Odia text from the recently released OdiEnCorp 2.0 [17].[13] In addition to this, we have used Odia corpus from OSCAR [15].[14] We also included in our dataset the parallel corpus sources by the Center for Visual Information Technology (CVIT).[15] This contains both CVIT PIB [v0.2] (Sentence aligned parallel corpus between 11 Indian Languages, crawled and extracted from the press information bureau website) and CVIT MKB [v0.0] (The Prime Minister's speeches - Mann Ki Baat, on All India Radio, translated into

[11] https://github.com/goru001/inltk

[12] https://github.com/AI4Bharat/indic-bert

[13] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3211

[14] https://oscar-corpus.com/

[15] http://preon.iiit.ac.in/~jerin/bhasha/

many languages). Finally, we added a Kaggle dataset scraped from Wikipedia. [16]

All the aforementioned datasets were merged into a single training file and then deduped to remove any duplicate sentences. The statistics of the dataset are shown in Table 1. This dataset also covers different domains such as the Bible, Wikipedia articles, literature websites, government portals, literature, and dictionaries.

Source	Sentences	Unique Odia tokens
OdiEnCorp2.0	97,233	1,74,045
CVIT PIB	58,461	66,844
CVIT MKB	769	3,944
OSCAR	1,92,014	6,42,446
Wikipedia	82,255	2,36,377
Total Deduped	430,732	1,123,656

Table 1. Dataset statistics.

4 Training and Evaluation

There are different variants of BERT [4] available for the language model such as ”A Lite BERT” (ALBERT) allowing to fit the model into memory and to increase the speed of BERT [12], StructBERT considering word level and sentence level ordering [6], and RoBERTa [13] that iterates on BERT’s pre-training procedure, including training the model longer, with bigger batches over more data, removing the next sentence prediction objective, training on longer sequences and dynamically changing the masking pattern applied to the training data.

In our experiment, we built the language model based on BERT and RoBERTa. We used a single GPU (NVIDIA Tesla V100-SXM2-16GB) for training our models. We train for 30 epochs with a training time of around 12 hours. During training, we did not consider upper and lower case letters, since the Odia language does not distinguish between them.[17] The configuration parameters are shown in Table 2.

4.1 BERT/RoBERTa Model Training

We explored training both BERT and RoBERTa models on the same dataset. RoBERTa is built on BERT’s language masking strategy, wherein the system learns to predict intentionally hidden sections of text within otherwise unannotated language examples.

[16] https://www.kaggle.com/disisbig/odia-wikipedia-articles

[17] https://fontmeme.com/odia/

Parameter	BERT	RoBERTa
Learning Rate	5e-5	5e-5
Training Epochs	5	10
Dropuout Prob	0.1	0.1
MLM Prob	0.1	0.2
Self attention layer	6	6
Attention head	12	12
Hidden layer size	768	768
Hidden layer Activation	gelu	gelu
Total parameters	84M	125M

Table 2. Training Configurations.

Model	Task	Accuracy
BertOdia	Text Classification	**96.0**
RoBERTaOdia	Text Classification	92.0
ULMFiT	Text Classification	91.9

Table 3. BertOdia Performance

We used Huggingface's interface to train both BERT and RoBERTa models. For the RoBERTa model, we chose to train a byte-level Byte-Pair Encoding (BPE) tokenizer (the same as GPT-2), with the same special tokens as RoBERTa. We arbitrarily picked its vocabulary size to be 52,000. The advantage of using a byte-level BPE (rather than a WordPiece tokenizer that was used in BERT) is that it will start building its vocabulary from an alphabet of single bytes [21]. Hence, all words will be decomposed into tokens which were reflected in the results we obtained. For the BERT model, we chose to train a WordPiece tokenizer. Again we arbitrarily picked vocabulary size to be 52,000.

The hyper-parameters are shown in Table 2. The learning curve for BertOdia training is shown in Figure 2.

4.2 Model Fine Tuning

To evaluate the models, we fine-tuned the BERT/RoBERTa models on a downstream task i.e. classification task. We used the Kaggle Odia news classification dataset[18] for this task. This data set is scraped from Odia daily new papers. It has headlines and the section of the newspaper from which they are scrapped. The dataset contains 3 classification labels (sports, business, entertainment) for the headlines and is divided into train/test sets. We used the same tokenizer that was trained and saved earlier. After tokenization, special tokens were added. Then the sentences were padded to 512 blocks. We trained the final layer using the classification dataset.

[18] `https://www.kaggle.com/disisbig/odia-news-dataset`

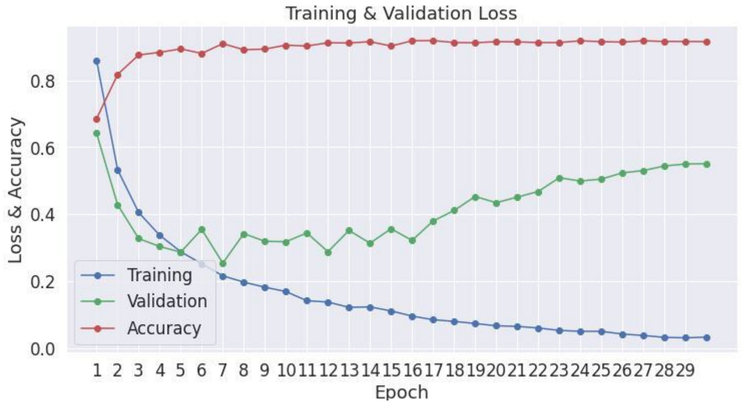


Fig. 2. Training: Performance vs epochs.

4.3 ULMFiT Model Training

We also used a pre-trained ULMFiT model to benchmark against our BERT-based models [7]. This model uses AWD_LSTM architecture. It also uses the default parameters and a drop_mult of *0.3*. The model is trained on Odia Wikipedia articles[19] and available on github.[20] We fine-tuned the model for the downstream classification task with the same news classification data used for the BERT and RoBERTa models.

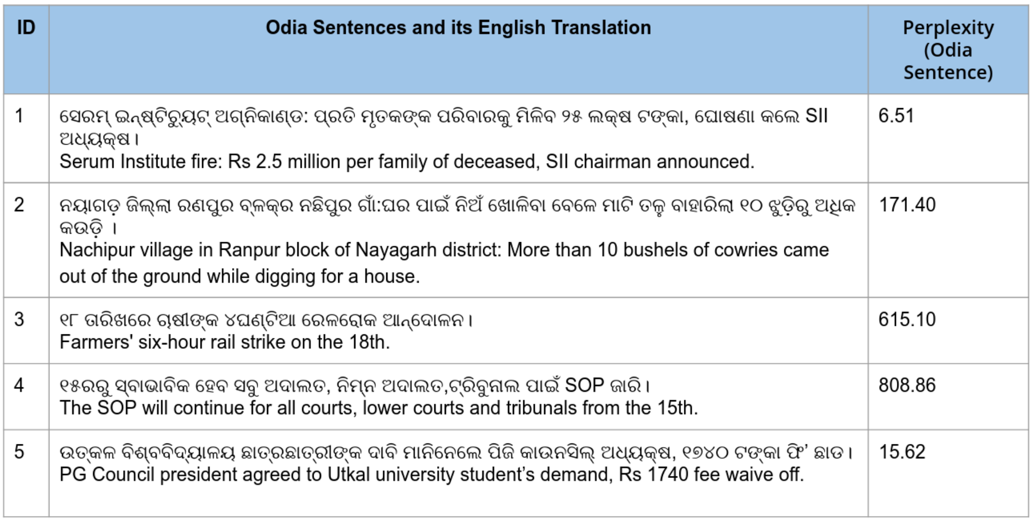
The evaluation results of our models on the news classification are shown in Table 3. The BertOdia model performance outperformed RoBERTaOdia by up to 4.5% (relative) and RoBERTaOdia reached performances on this task compared to the performance of ULMFiT.

4.4 Language Model Evaluation

To evaluate the BERT model, we have used the perplexity (PPL) score using BERT masking approach [20, 22]. It is important to notice that although the PPL metric applies specifically to classical language models, it is not well defined for masked language models like BERT. When dealing with masked languages, PPL can be thought of as an evaluation of the model's ability to predict uniformly among the set of specified tokens in a corpus, meaning that the tokenization procedure has a direct impact on a model's perplexity. Additionally, when working with approximate models (e.g., BERT), we typically have a constraint on the number of tokens the model can process, e.g., 512 tokens. Thus, when computing the PPL of BERT, the input sequence is typically broken into subsequences equal to the model's maximum input size, considering a sliding-window strategy.

[19] `https://www.kaggle.com/disisbig/odia-wikipedia-articles`

[20] `https://github.com/goru001/nlp-for-odia`



ID	Odia Sentences and its English Translation	Perplexity (Odia Sentence)
1	ସେରମ୍ ଇନ୍ଷ୍ଟିଚ୍ୟୁଟ୍ ଅଗ୍ନିକାଣ୍ଡ: ପ୍ରତି ମୃତକଙ୍କ ପରିବାରକୁ ମିଳିବ ୨୫ ଲକ୍ଷ ଟଙ୍କା, ଘୋଷଣା କଲେ SII ଅଧ୍ୟକ୍ଷ। Serum Institute fire: Rs 2.5 million per family of deceased, SII chairman announced.	6.51
2	ନୟାଗଡ଼ ଜିଲ୍ଲା ରଣପୁର ବ୍ଲକ୍‌ର ନଛିପୁର ଗାଁ:ଘର ପାଇଁ ନିଅଁ ଖୋଳିବା ବେଳେ ମାଟି ତଳୁ ବାହାରିଲା ୧୦ ଝୁଡ଼ିରୁ ଅଧିକ କଉଡ଼ି । Nachipur village in Ranpur block of Nayagarh district: More than 10 bushels of cowries came out of the ground while digging for a house.	171.40
3	୧୮ ତାରିଖରେ ଚାଷୀଙ୍କ ୪ଘଣ୍ଟିଆ ରେଳରୋକ ଆନ୍ଦୋଳନ। Farmers' six-hour rail strike on the 18th.	615.10
4	୧୫ରରୁ ସ୍ୱାଭାବିକ ହେବ ସବୁ ଅଦାଲତ, ନିମ୍ନ ଅଦାଲତ,ଟ୍ରିବୁନାଲ ପାଇଁ SOP ଜାରି। The SOP will continue for all courts, lower courts and tribunals from the 15th.	808.86
5	ଉତ୍କଳ ବିଶ୍ୱବିଦ୍ୟାଳୟ ଛାତ୍ରଛାତ୍ରୀଙ୍କ ଦାବି ମାନିନେଲେ ପିଜି କାଉନସିଲ୍ ଅଧ୍ୟକ୍ଷ, ୧୭୪୦ ଟଙ୍କା ଫି' ଛାଡ। PG Council president agreed to Utkal university student's demand, Rs 1740 fee waive off.	15.62

Fig. 3. Perplexity score for sample Odia sentences.

This involves repeatedly sliding the context window so that the model has more context when making each prediction.

Accordingly, we evaluated our model in a small set of unseen data (500 sentences) extracted from a news website.[21] For all the considered sentences, the length is equal to or less than 512 tokens. Hence, considering the approach described above, we obtained a mean score of *PPL = 134.97* with a standard deviation of *196.98*. The perplexity score differs (low-high) based on the sentences as depicted in Figure 3.

In [18] authors describe the training process performed on Embeddings from Language Models (ELMo) for many Indian languages. Among other languages, they report a perplexity of (*PPL=975*) for the Odia language model. Although we can not make a direct comparison, our trained model obtains a better performance for this morphologically rich language [11].

4.5 IndicGLUE tasks

We also benchmarked our model against a few IndicGLUE tasks [8]. Despite our model being trained on 6% of data used for training IndicBert, we got comparable results as shown in Table 4.

For the Cloze-style Multiple-choice QA task, we feed the masked text segment as input to the model and at the output, we have a softmax layer that predicts a probability distribution over the given candidates. We fine-tune the model using cross-entropy loss with the target label as 1 for the correct candidate and 0 for

[21] `http://news.odialanguage.com/`

Model	Article Genre Classification	Cloze-style multiple-choice QA
XLM-R	97.07	35.98
mBERT	69.33	26.37
IndicBERT base	97.33	**39.32**
IndicBERT large	**97.60**	33.81
BertOdia	96.90	23.00

Table 4. Comparison of BertOdia with IndicBERT. BertOdia was trained on 6% of the data of IndicBERT.

the incorrect candidates [8]. The code for the fine-tuning model is given in the IndicGLUE github.[22]

For the Article Genre Classification task we used the IndicGLUE dataset for news classification.[23] It is generated from an Odia daily newspaper source except it has one more genre called crime. More details of the IndicGLUE datasets could be found on its website.[24]

5 Conclusion

In this paper, we presented BertOdia, a pre-trained Odia language model which can be useful for many language and speech processing tasks for this low resource language. BertOdia will be the first language-specific BERT model in Odia which can be used by researchers for many language and speech processing tasks. Our studies will help researchers working on low-resource languages. The code and dataset are available at:

```
https://colab.research.google.com/gist/satyapb2002/
aeb7bf9a686a9c7294ec5725ff53fa49/odiabert_languagemodel.ipynb
```

Future work will include:

- Incorporating OdiEnCorp1.0 [25] [16] and IndicCorp data sets [26] which have around 8.2M sentences combined. We want to include a large volume of Odia text covered in a variety of domains to build a robust language model for better performance. Even the IndicBERT large has poor performance on Assamese and Odia-the two languages due to the smallest corpora sizes as compared to other Indian languages [8].
- Developing NLP datasets (natural language inference, question answering, next sentence prediction) for the Odia language.

[22] https://github.com/AI4Bharat/indic-bert/tree/master/fine_tune

[23] https://storage.googleapis.com/ai4bharat-public-indic-nlp-corpora/evaluations/inltk-headlines.tar.gz

[24] https://indicnlp.ai4bharat.org/indic-glue/

[25] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2879

[26] https://indicnlp.ai4bharat.org/corpora/

- Preparing a dashboard similar to GLUE [23][27] and IndicGLUE [28], called "OdiaGLUE" for evaluating the model on various natural language understanding (NLU) tasks specific for Odia language.
- Enriching OdiaGLUE with more NLU tasks as the recent IndicGLUE supports limited NLU tasks for Odia.

Acknowledgments

The authors Shantipriya Parida, Maël Fabien and Petr Motlicek were supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

The authors do not see any significant ethical or privacy concerns that would prevent the processing of the data used in the study. The datasets do contain personal data, and these are processed in compliance with the GDPR and national law.

Esaú Villatoro-Tello, was supported partially by Idiap Research Institute, SNI CONACyT, and UAM-Cuajimalpa Mexico.

References

1. Adams, O., Makarucha, A., Neubig, G., Bird, S., Cohn, T.: Cross-lingual word embeddings for low-resource language modeling. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 937–947 (2017)
2. Arora, G.: inltk: Natural language toolkit for indic languages. In: Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS). pp. 66–71 (2020)
3. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: Practical ML for Developing Countries Workshop@ ICLR 2020 (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186 (2019)
5. Grießhaber, D., Maucher, J., Vu, N.T.: Fine-tuning bert for low-resource natural language understanding via active learning. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 1158–1171 (2020)
6. Hazem, A., Bouhandi, M., Boudin, F., Daille, B.: Termeval 2020: Taln-ls2n system for automatic term extraction. In: Proceedings of the 6th International Workshop on Computational Terminology. pp. 95–100 (2020)
7. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 328–339 (2018)

[27] `https://gluebenchmark.com/`

[28] `https://indicnlp.ai4bharat.org/indic-glue/`

8. Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M.M., Kumar, P.: IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In: Findings of EMNLP (2020)
9. Kocmi, T., Parida, S., Bojar, O.: CUNI NMT system for WAT 2018 translation tasks. In: Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation. Association for Computational Linguistics, Hong Kong (1–3 Dec 2018), https://www.aclweb.org/anthology/Y18-3002
10. Korzeniowski, R., Rolczynski, R., Sadownik, P., Korbak, T., Mozejko, M.: Exploiting unsupervised pre-training and automated feature engineering for low-resource hate speech detection in polish. Proceedings ofthePolEval2019Workshop p. 141 (2019)
11. Kumar, S., Kumar, S., Kanojia, D., Bhattacharyya, P.: "a passage to india": Pre-trained word embeddings for indian languages. 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. 352–357 (2020)
12. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. In: International Conference on Learning Representations (2019)
13. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
14. Martin, L., Muller, B., Suárez, P.J.O., Dupont, Y., Romary, L., de la Clergerie, É.V., Seddah, D., Sagot, B.: Camembert: a tasty french language model. In: ACL 2020-58th Annual Meeting of the Association for Computational Linguistics (2020)
15. Ortiz Suárez, P.J., Romary, L., Sagot, B.: A monolingual approach to contextualized word embeddings for mid-resource languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1703–1714. Association for Computational Linguistics, Online (Jul 2020), https://www.aclweb.org/anthology/2020.acl-main.156
16. Parida, S., Bojar, O., Dash, S.R.: Odiencorp: Odia–english and odia-only corpus for machine translation. In: Smart Intelligent Computing and Applications, pp. 495–504. Springer (2020)
17. Parida, S., Dash, S.R., Bojar, O., Motlıcek, P., Pattnaik, P., Mallick, D.K.: Odiencorp 2.0: Odia-english parallel corpus for machine translation. In: LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020. p. 14
18. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237 (2018)
19. Raju, A., Filimonov, D., Tiwari, G., Lan, G., Rastrow, A.: Scalable multi corpora neural language models for asr. Proc. Interspeech 2019 pp. 3910–3914 (2019)
20. Salazar, J., Liang, D., Nguyen, T.Q., Kirchhoff, K.: Masked language model scoring. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2699–2712 (2020)
21. Sennrich, R., Haddow, B., Birch, A.: Neural Machine Translation of Rare Words with Subword Units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL (2016)

22. Wang, A., Cho, K.: Bert has a mouth, and it must speak: Bert as a markov random field language model. In: Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation. pp. 30–36 (2019)
23. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 353–355. Association for Computational Linguistics, Brussels, Belgium (Nov 2018). https://doi.org/10.18653/v1/W18-5446, https://www.aclweb.org/anthology/W18-5446
24. Wang, Z., Mayhew, S., Roth, D., et al.: Extending multilingual bert to low-resource languages. arXiv preprint arXiv:2004.13640 (2020)