SIGUL 2022

SIGUL 2022 Workshop - June 24-25, 2022

Nepali Encoder Transformers:

An Analysis of Auto Encoding Transformer Language Models for Nepali Text Classification

Utsav Maskey, Manish Bhatta, Shiva Raj Bhatta, Sanket Dhungel, Bal Krishna Bal

Information and Language Processing Research Lab
Department of Computer Science and Engineering
Kathmandu University

Contents

Introduction

- The Transformer architecture has become the go-to method for neural language modeling.
- However, low-resource languages have not received enough attention and the study is far from sufficient.
- We focus on the factors that are unique and applicable to the Nepali language, a low to medium resourced language that must be considered when modeling.

Related Work

- **XLM-RoBERTa**: XLM-RoBERTa (Conneau et al., 2020) is a multi-lingual model that is trained on large dataset of filtered CommonCrawl data covering 100 languages.
- **Indic Transformers:** This work focuses on modeling medium-resourced languages including Hindi, Bengali and Telugu languages using transformers. (Jain et al., 2020)
- **NepaliBERT:** It trains BERT based language model for Nepali Language. (Rajan, 2021), (Pudasaini, 2022). We compare the performance of these models in this work.

Nepali Language

- The Nepali language is spoken by more than 16 million people all over the world. ("Nepali language - Wikipedia", 2022)
- It is written in the Devanagari script, and follows Subject Object Verb (SOV) pattern.
- The Nepali Language follows the Abugida writing system.

Letters	Count	Examples
Independent Consonants	33	क, ख, ग, घ, … ज्ञ
Independent Vowels	11	आ, इ, उ, ओ, ए, …
Dependent Vowels	10	ा, ि, ु, ो, े, …
Dependent Symbols		्, ँ, ं,, ः
Punctuation		। ; ? / : , etc.
Digits	9	१, २, ३, ४, ५, ६, ७, ८, ९

Example:

Compound Letter:	लु	(ल + ु)
Word:	फ्लु	(फ + ् + ल + ु)
Sentence:	मलाई फ्लु लागेको छ ।	

Tokenization

- It is the process of converting words and character splittings into machine understandable tokens.
- Unicode Normalizations which benefit languages with latin alphabets, causes issues on Abugida languages.
- This issue was later realized and resolved in Multilingual BERT but was overlooked in nepaliBERT (Pudasaini, 2022).

	Before Tokenization	Tokenized
Decomposition:	फ्लु (फ + ् + ल + ु)	फल (फ+ ल)
Meaning:	Flu	Fruit

Table: Unicode Normalization removes the dependent symbols and causes ambiguity.

Input Text : "फ्लुको कारणले हुने पहिलोनेपाली भवकृष्ण भट्टराई"	
Tokenizer	Tokenized output
Shushant/ nepaliBERT	['फल', '##को', 'कारण', '##ल', 'ह', ##न','पहिलो', '##न', '##पाली', 'भव', '##क', '##षण', 'भट', '##टर', '##ाई']
R4J4N/ NepaliBERT	['फ्लु', '##को', 'कारणले', 'हुने', 'पहिलो', '##नेपाली', 'भव', '##कृष्ण', 'भट्टराई']
Sentence Piece Model [Ours]	['_फ्लु', 'को', '_कारणले', '_हुने', '_पहिलो', 'नेपाली', '_', 'भव', 'कृष्ण', '_भट्टराई']

Table: Comparison of tokenizers.

Masked Language Modeling

- **Dataset used:**
  - OSCAR (Suárez et al., 2020)
  - cc100 dataset (Conneau et al., 2020)
  - iNLTK dataset (Arora, 2020)

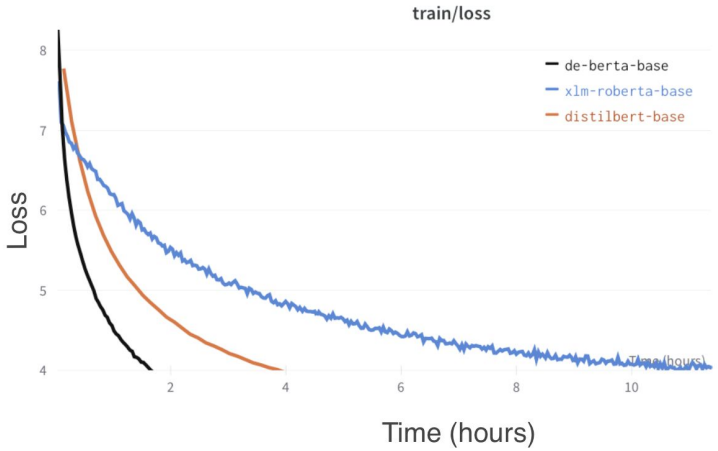

Figure: Language Models Training Feasibility Test

- **Models considered for training:**
  - De-berta-base (P. He et al., 2020)
  - Distilbert-base (Sanh et al., 2019)
  - XLM-roberta (XLM-R) (Conneau et al., 2020)

Masked Language Modeling

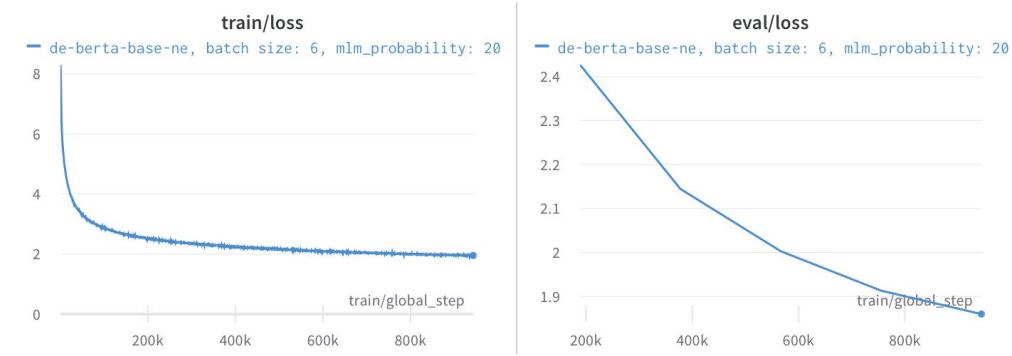

Figure: Training of DeBERTa Model

Model	Train/loss	Batch size	Perplexity (eval)
Distilbert-base	2.6412	28	12.3802
De-berta-base	**1.9375**	6	**6.4237**

Table: Summary of LM training for 5 epochs with MLM probability of 20%.

We trained two auto-encoding transformer models, the DeBERTa model that focuses on attaining the best performance, and the DistilBERT model which focuses on being lightweight

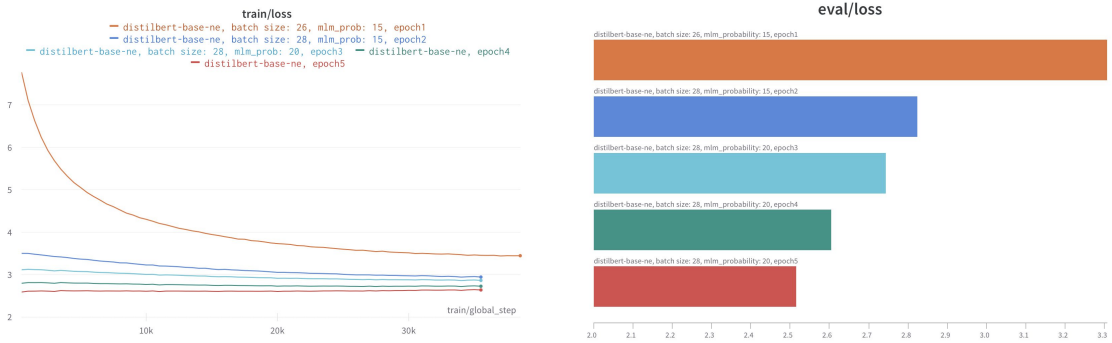

Figure: Training of DistilBERT Model

Downstream Classification Comparison

The classification task performance evaluation is performed on the “16 Nepali News” dataset (Chaudhary & Sabin, 2017). The dataset consists of approximately 14,364 Nepali language news documents, partitioned (unevenly) across 16 different newsgroups.

Model	Epoch	Train steps	Highest Accuracy
deberta-base [ours]	3	4845	**88.93%**
distilbert-base [ours]	3	**1212**	88.31%
nepaliBERT	4	3231	85.96%
NepaliBERT	6	3230	81.05%
XLM-Roberta	5	8075	84.02%

Table: Highest accuracy attained by the models

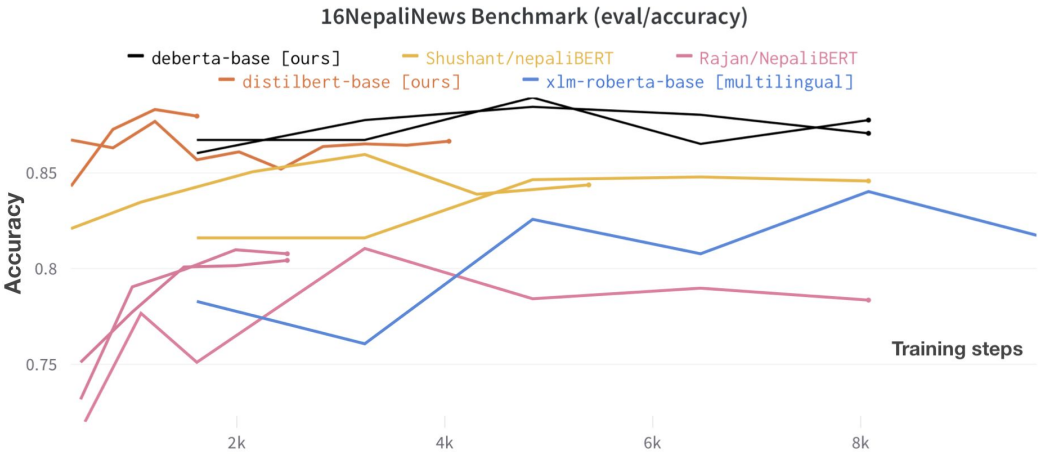

Figure: Evaluating language models on “16 Nepali News” Dataset. Training are performed with varying hyperparameters. Each progression in the x-axis represents an Epoch.

Challenges of Modeling URLs

- **Tokenizer:** Language-specific pre-processing of text benefits the training of language models.
- **MLM with Progressive Masking:** Models are trained for multiple epochs by gradually increasing the number of masked tokens on every preceding epoch.

Conclusion

- We analyzed the need and effectiveness of pre-trained Transformer language models for Nepali.
- We trained two auto-encoding transformer models: The DeBERTa model, which focuses on achieving the best performance, and the DistilBERT model, which focuses on being lightweight.
- We looked into language specific details and low-resource language modeling techniques and investigated useful training methods.

References

- Arora, G. (2020). i{NLTK}: Natural Language Toolkit for Indic Languages. Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS), 66–71.
- Conneau, A., Khandelwal, K., Goyal, N. (2020). Unsupervised Cross-lingual Representation Learning at Scale.
- He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654.
- Jain, K., Deshpande, A., Shridhar, K., Laumann, F., & Dash, A. (2020). Indic-transformers: An analysis of transformer language models for Indian languages.
- Nepali language - Wikipedia. En.wikipedia.org. (2022). Retrieved 22 May 2022, from https://en.wikipedia.org/wiki/Nepali_language
- Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451.
- Pudasaini, S. (2022). Pretraining Nepali Masked Language Model using BERT Architecture. 3rd International Conference on Natural Language Processing, Information Retrieval, and AI
- Rajan. (2021). NepaliBERT. https://huggingface.co/Rajan/NepaliBERT
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing.

Thank you