Transforming POS Tagging for Maithili Language

Abhisek Kumar Jha, Nebojsa Kilibarda

May 3, 2024

Abstract

This is an attempt in developing a robust BERT-based classifier tailored for a low-resource natural language processing (NLP) task. Focusing on Maithili, an underrepresented language, the model employs BERT's contextual embeddings and integrates a Conditional Random Field (CRF) layer for sequential labeling, enhancing its ability to capture intricate linguistic nuances. Leveraging a manually annotated a Maithili corpus containing 52,190 words, this classifier endeavors to achieve outstanding performance despite the limited linguistic resources available. By pushing the boundaries of NLP in underrepresented languages, this initiative underscores the importance of inclusivity and accessibility in language technology development.We ran experiments and achieved an accuracy of 85%

1 Introduction

The goal of this project is to compare effectiveness of the transformer to various other RNN models on a limited resource natural language processing task. We created a classifier to perform POS (Part-Of-Speech) tagging. The target language for our research is Maithili, a language that despite being spoken by roughly 40 million people, is still underrepresented in the NLP community. We found two works that a big part of our work stemmed from, published by the same group. They created the initial dataset for Maithili POS tagging. [PS20] This is the dataset we used. Their latest research set the benchmark of 91.57% using the Bi-GRU-CRF model. [SAH22] We also looked into research on utility of CRF with Transformers. [ZW19]

2 Corpus and POS Tagset

The tagged corpus comprises a total of 52,190 words, which are divided into two parts: the training corpus, consisting of 48,007 words, and the validation corpus, consisting of 4,183 words. The sentences in the validation corpus are randomly selected from the entire corpus and are not included in the training data. Table 1 presents detailed statistics of the annotated corpus.

Table 1: Statistics of the Maithili corpus.

Corpus	Number of Sentences	Number of Words	Number of Unique Words
Train	2298	48007	9231
Validation	162	4183	1687
Total	2460	52190	9856

In part-of-speech tagset for Maithili, we have 27 unique symbols, each serving to categorize words into distinct linguistic classes. These symbols include NN, representing Singular Common Nouns, NNS for Plural Nouns, NNP denoting Proper Nouns, NNL indicating Name Spatial and Temporal, and NTP signifying Name Title Person. Furthermore, we have designated symbols such as PRN for Personal Pronouns, PRO for Pronouns, PRF for Reflexive Pronouns, and PRQ for Question Words. Other categories encompass DEM for Demonstratives, MOD for Modals, VM for Main Verbs, VN for Dependent Verbs, and VAUX for Auxiliary Verbs. Additionally, we have included JJ for Adjectives, JJC for Quantifiers, RB for Adverbs, PP for Prepositions, and CC for Conjunctions. Completing the set are symbols like CND for Conditionals, IN for Interjections, NEG for Negations, CD for Cardinals, OD for Ordinals, FN for Foreign Words, SYM for Symbols, and PUNC for Punctuation.

3 Transformer

3.1 BERT

To get a baseline evaluation of the transformer approach, we initially fine-tuned the BERT model to perform the POS tagging task, using the findings in the landmark research paper, 'Attention is all you need'. [VSP+17] We were interested in how the baseline model would stack up to the RNN-based model created by the SoTA research. Their approach was a Bi-GRU-CRF model, which yielded a 91.53% accuracy using the Maithili dataset. The base BERT model we built was capable of 81.97%.

3.2 BERT + CRF

Seeing that the transformer is only accurate to 82% with fine tuning that we were able to perform, we decided to look at a different approach. The problem transformers can potentially have in a task such as POS tagging is that outputs are conditionally independent. For example, if there are four tags a, A, b and B, and b can only be output after a, a transformer may not access this information and might produce sequence 'Bb' or 'Ab' because of it's inability to model the outputs sequentially. We thought that adding a CRF layer on top of the transformer would help increase the accuracy by ensuring that the output sequences remain syntactically valid. While this did improve accuracy, it was not enough to be competitive with the SoTA model, coming in at 82.85% accuracy, around 8% short of the SoTA.

Table 2: Accuracy by varying the size of the training data.

Training Data size (in words)	Accuracy (in %)
10000	46.78
20000	64.62
30000	72.59
40000	78.22
48007	82.85

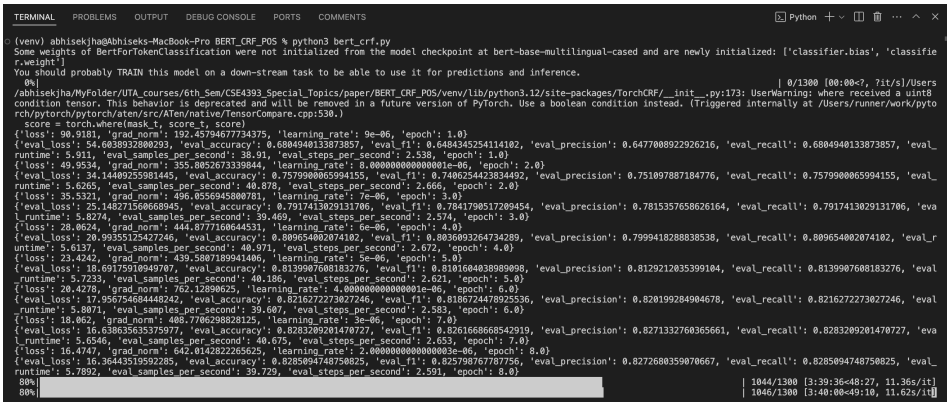

Figure 1: Terminal Result

3.3 Future Exploration

The problem in our approach is not with the model. Our findings indicate that the problem lies in the amount of data. In order for this model to become competitive with the RNN based models that rule the sphere at the moment, more data is necessary, according to the current findings in Transformers. There are multiple ways in which this can be accomplished, some of which involve back-translation, training on syntactically similar languages, etc. Our goal was to test the power of this model on limited resources and attempt to create a model that is able to perform the task successfully even with little data. For this reason, generating more data was not of interest to us in our research. We will keep looking at ways to optimize the model architecture so as to become better suited for smaller datasets

and hopefully able to take the torch from RNN's even on smaller datasets such as the Maithili POS tagging problem.

References

[PS20] Ankur Priyadarshi and Sujan Kumar Saha. Towards the first maithili part of speech tagger: Resource creation and system development. *Computer Speech & Language*, 62:101054, 2020.

[SAH22] SUJAN KUMAR SAHA. A study on the performance of recurrent neural network based models in maithili part of speech tagging. 2022.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[ZW19] Linhao Zhang and Houfeng Wang. Using bidirectional transformer-crf for spoken language understanding. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I 8*, pages 130–141. Springer, 2019.