Bidirectional English-Nepali Machine Translation System for Legal Domain

[1]Shabdapurush Poudel, [1]Bal Krishna Bal and [2]Praveen Acharya
[1]Kathmandu University, Dhulikhel, Nepal | [2]Dublin City University, Ireland

poudelshabda@gmail.com, bal@ku.edu.np, acharyaprvn@gmail.com

INTRODUCTION

- Deep Learning based Machine Translation models are performing better and have successful advancement in the translation domain.
- Movement from Statistical Based Machine Translation model to Neural Machine translation has not been smooth for the Nepali Language.
- Neural Machine Translation in the legal domain has not been explored previously.
- Translation in Nepali Legal Domain requires proper digital dataset and footprint, and most of the previous works are focused in the general domain of the language.
- Previous works are based on SMT and NMT model, and domain specific models with specific datasets are yet to be explored.

OBJECTIVES

- To explore Transformer based NMT models.
- To develop domain specific dataset in Legal Domain.
- To develop an efficient model better suited for Nepali legal domain.
- To identify various issues and challenges that arise while creating the dataset and models.
- To evaluate the previous general domain datasets performance and accuracy against legal terminologies.

RESEARCH METHODOLOGY

- Through previous literature and review, we identified the state of NMT, and areas that can be further improved through our contribution.
- We created a domain specific dataset of size 125,000 Nepali - English parallel sentences.
- We trained a transformer-based model from scratch for this specific domain-based translation purpose in Nepali English language pair.
- We analyzed the output and findings of the fine tuned model and also identified the areas for improvements.

Dataset Collection

- Documents provided by Legal firms upon NDA.
- Supreme Court Site for public legal proceedings.
- Scraping news site for English and Nepali pair news, consisting of legal terminologies.

Dataset

- Created Dataset from scratch for Legal Domain.

CORPUS Source	CORPUS Size (Parallel Sentences)
Manually Translated Documents	60k
Legal Websites	25k
News Sites	40k

Table 1 : *Data Source and CORPUS Size*

EVALUATION

- Evaluated our model using BLEU score for performance of the model in both general and legal domain.
- For final evaluation, legal professionals were also given access to the model, to try and rate the translation.

Model		Nepali - > English		English - > Nepali	
Model	Domain	Legal	General	Legal	General
NMT Model		7.98	13.67	6.63	9.47
RNN Model		6.19	-	5.89	-

Table 2 : *BLEU score of RNN and NMT model*

DATA PREPROCESSING

NORMALIZATION AND TOKENIZATION

- Used IndicNLP to Normalize and Tokenize Nepali sentences.
- Used Sacremoses library for English sentences.

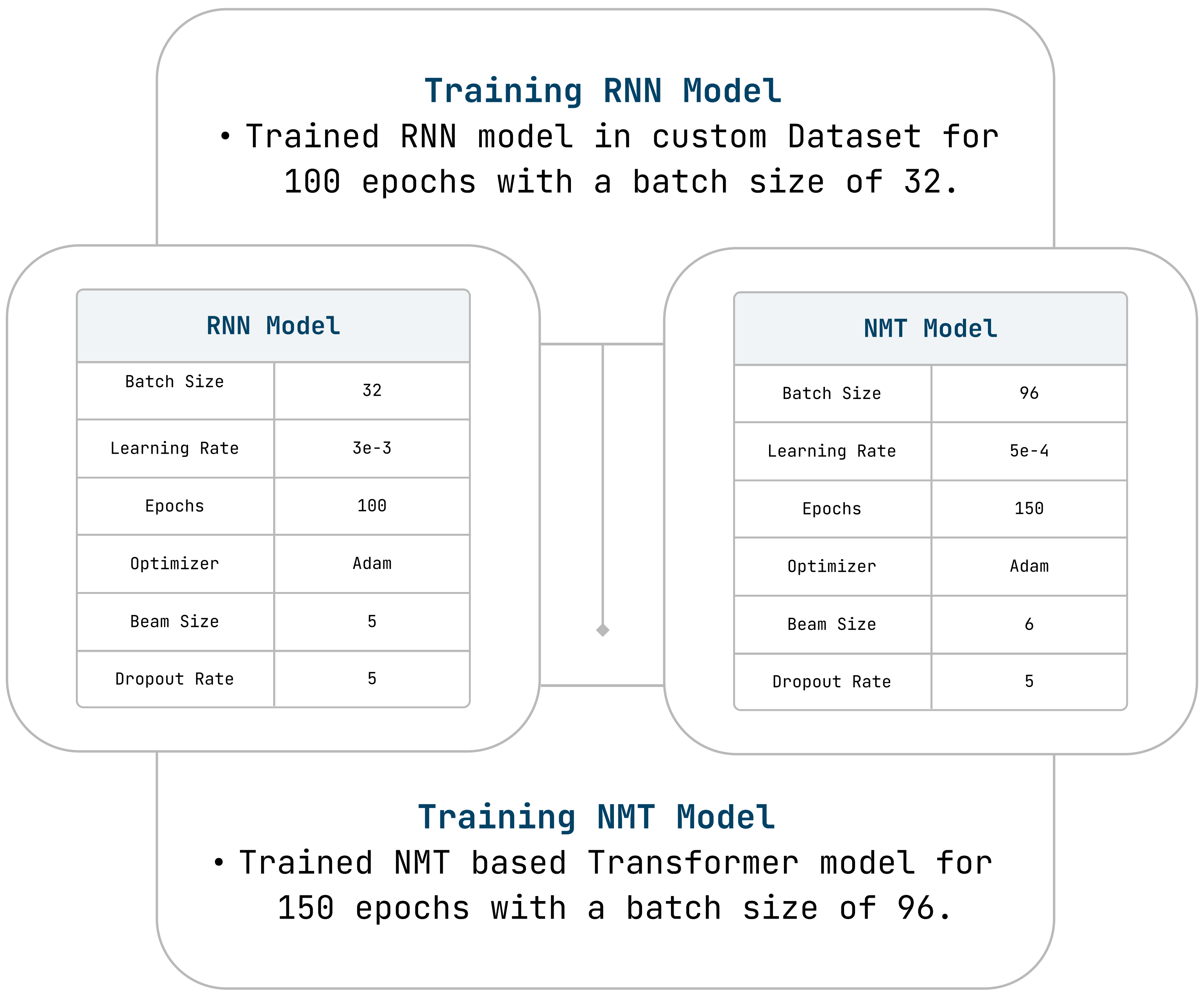
VOCABULARY BUILDING

- Used BPE to learn legal vocabulary of size about 10K.
- Used Sentencepiece library to learn BPE for source language (Nepali).

Used Alignment tools to align the sentence pair.

TRAINING

- Used RNN model initially to train the dataset.
- Used NMT based 6 encoder-decoder layer transformer model to train the dataset.



Training RNN Model
- Trained RNN model in custom Dataset for 100 epochs with a batch size of 32.

RNN Model	
Batch Size	32
Learning Rate	3e-3
Epochs	100
Optimizer	Adam
Beam Size	5
Dropout Rate	5

NMT Model	
Batch Size	96
Learning Rate	5e-4
Epochs	150
Optimizer	Adam
Beam Size	6
Dropout Rate	5

Training NMT Model
- Trained NMT based Transformer model for 150 epochs with a batch size of 96.

Figure 1 : *Training Parameters for models*

CONCLUSION & FUTURE WORKS

- Being first work in domain specific work in Nepali language, the results set a baseline for future works.
- Creating a proper domain specific dataset for translation model, can help further improve the performance and quality of the model.
- Need to work on date conversion from English Gregorian and Nepali Bikram-Sambat calendars.
- Can further improve the fluency of translation, given proper resources.
- Handle O-O-V words to improve the quality of translated documents.

ACKNOWLEDGEMENTS

We would like to express our sincere gratitude to law firms (NDA) of Nepal along with students of AI, Kathmandu University for providing us with data and helping us with cleaning and creating parallel datasets.

SIGUL 2024