Efforts Towards Developing a Tamang Nepali Machine Translation System

Binaya K. Chaudhary[1] **Bal Krishna Bal**[2] **Rasil Baidar**[1]

Information and Language Processing Research Lab
Department of Computer Science and Engineering
Kathmandu University
Dhulikhel, Kavre, Nepal
[1]{binayachaudari, rasilgrt}@gmail.com
[2]bal@ku.edu.np

Abstract

The Tamang language is spoken mainly in Nepal, Sikkim, West Bengal, some parts of Assam, and the North East region of India. As per the 2011 census conducted by the Nepal Government, there are about 1.35 million Tamang speakers in Nepal itself. In this regard, a Machine Translation System for Tamang-Nepali language pair is significant both from research and practical outcomes in terms of enabling communication between the Tamang and the Nepali communities. In this work, we train the Transformer Neural Machine Translation (NMT) architecture with attention using a small hand-labeled or aligned Tamang-Nepali corpus (15K sentence pairs). Our preliminary results show BLEU scores of **27.74** for the Nepali→Tamang direction and **23.74** in the Tamang→Nepali direction. We are currently working on increasing the datasets as well as improving the model to obtain better BLEU scores.

1 Introduction

Machine Translation (MT) Systems today represent one of the biggest achievements of human mankind in the field of Artificial Intelligence (AI) and Language Technologies (LT). With the advancement of Deep Neural Networks (DNN) and abundance of data, today's MT Systems already claim more than 90% accuracy in the translation thus opening doors for the adoption and use of highly reliable translation systems.

These systems serve multiple folds: (1) bridge the language barrier between different language speaking communities; (2) preserve languages that are not as popularly spoken or used in the younger generations and many more. If we talk about the Tamang language, native to the Tibeto-Burman group of the Sino-Tibetan language family and spoken by about 1.35 million native speakers in Nepal according to the 2011 census by the Nepal Government[1], then we can say that it can greatly benefit from MT systems. This is further enunciated by the fact that Tamang is one of the languages which is highly affected by the latest migration trends, both within and outside the country, in quest of better lives and opportunities. Consequently, the language is not being spoken or learned by the younger generation and is being limited to the older generations thus creating a fear of extinction. On the contrary, in areas with dense Tamang populations in Nepal the primary language of communication is Tamang and even the medium of teaching is Tamang in these areas. Needless to note, MT Systems can open up the Tamang community from such areas to the outer world via the knowledge sources available in Nepali and English languages on the Internet.

In this paper, we discuss our efforts on developing a Tamang↔Nepali Machine Translation System. The biggest contribution of this work is the development of the parallel corpus of 15,000 parallel sentences in Tamang and Nepali from scratch. Moreover, we have adopted the Transformer architecture (Vaswani et al., 2017) for a low-resource language pair (Tamang-Nepali) with quite a few language-based customizations of hyper-parameters. We have achieved a BLEU score of **27.74** for Nepali→Tamang and a BLEU score of **23.74** for Tamang→Nepali on the test sets.

2 Challenges in Parallel Corpus Development

This is the first work of any kind in terms of Nepali↔Tamang Machine Translation; there is no related work in this regard. When we started working on the project, we did not have any parallel

[1] https://unstats.un.org/unsd/demographic-social/census/documents/Nepal/Nepal-Census-2011-Vol1.pdf

Proceedings of the 17th International Conference on Natural Language Processing, pages 281–286
Patna, India, December 18 - 21, 2020.

sentences for the Tamang-Nepali pair. Hence, we started to look for any available resources for the language pair. The problem with Tamang is that there is still no consensus in the community regarding the script for the language. There are a significant number of people in the Tamang community advocating for the Tamyig script which is very close to Tibetan script. However, since the Tamyig script is mostly confined to scriptures, holy books of the Lamas and not taught widely in schools, a large chunk of the Tamang community still uses the Devanagari script for writing and print media. Hence, we can find quite a few publications in Tamang written in the Devanagari script. Based on the popular usage and availability, we also decided to opt for Devanagari script-based Tamang language texts. Our very first preliminary attempt included seeing if we could use the Bible translations in Nepali[2] and Tamang[3] as the parallel sentences. However, the resource had its own set of issues

- the quality of the translation was not good,
- sentence-level alignment of translations did not yield satisfactory results,
- there was an issue of inconsistent use of the Eastern and Western Tamang, dialects of the Tamang language in Nepal.

So, we dropped the idea of using this resource. We also tried using Tamang songs and their translations but this attempt was not fruitful as the translations lacked wide coverage of texts of varying complexity in the languages. Finally, we established contact with Tamang linguists and experts as well as activists who had been working for the Tamang language in different capacities for long and we decided to initially develop 15,000 parallel sentences for Tamang-Nepali. These sentences range from simple to medium and complex sentences and have been taken from different sources like day-to-day spoken communication texts, child storybooks, general articles from Tamang language magazines, literary articles, etc.

[2] https://www.bible.com/bible/1711/MAT.1

[3] https://www.bible.com/bible/1177/MAT.1

3 Crowdsourcing the Development of Parallel Corpus

In order to make the process of the parallel corpus development more managed, we developed a tool, namely the Corpus Development Software which is a platform for submitting translations to the pre-selected sentences for the source language, as one sentence at a time. This approach prevents the misalignment of the sentences in due course of translation as the task demands one sentence in the target for each source sentence. The tool does not just provide an interface for submission of the translations but also lets us know the assigned translator regarding the deadline and what the status of the assigned task is ("Not Started", "In Progress", "Completed", "Past due" etc.). Once the translator submits the task, then the system similarly assigns the task to the reviewer and accordingly sets the status of the task for the review phase. Only after the reviewer submits the task, it becomes part of the finalized parallel corpus. We present the high-level workflow of the Corpus Development tool in Figure 1.

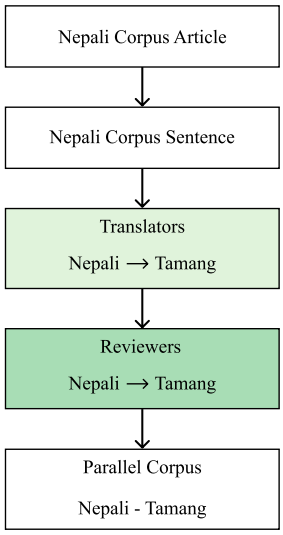

Figure 1: Workflow of the Corpus Development tool

4 Choosing the Right MT System Architecture

We studied a few state-of-the-art NMT architectures in due course of finalizing the MT architecture for our system. The first one included Google's Neural Machine Translation system (GNMT) (Wu et al., 2016), developed by Google. It is an LSTM network with 8 layers of encoders and decoders

with an attention mechanism. GNMT (Wu et al., 2016) tackles some of the well-known problems in NMT such as slower training, ineffectiveness in dealing with rare words, etc.

Similarly, Gehring et al. (2017) introduced a NMT architecture based on Convolutional Neural Network (CNN) for sequence-to-sequence learning that outperformed GNMT (Wu et al., 2016) on WMT'14 English to German translation and WMT'14 English-French.

Johnson et al. (2017) suggest a simple solution for a multilingual translation system using a single NMT model. The proposed model is identical to the GNMT system (Wu et al., 2016) with some optional connections on the actual network and few modifications in the input sequence. Johnson et al. (2017) describes the improvements in translation quality of low resourced language pairs when low-resourced language pairs and high resourced language pairs are mixed into a single model.

For this research, we follow the recent state-of-the-art Transformer (Vaswani et al., 2017) NMT architecture to create a fully supervised Neural Machine Translation system for developing a translation system for the Tamang-Nepali language pair. The core concept behind the Transformer model is self-attention—the ability to look at the multiple positions of an input sequence to understand/compute the representation of the sequence. This approach has proven to be better than recurrent methods, popularly being used for sequence-to-sequence learning. Transformer architecture addresses the recursion problem and allows parallelization, therefore reducing training time and increasing the performance making it cheaper and quicker to train. It can handle longer-range dependencies than most other translation models.

5 Experimental Setup

In this section, we present the experimental settings used for training the MT architecture and models for experimenting and reporting the results. The models which varied with different vocabulary of pre-selected words (1K, 2.5K and 5K) were trained using *fairseq*[4] toolkit (version 0.9.0) (Ott et al., 2019). Google Colab[5] (free-tier) is used for training the MT architectures with hardware acclerator set to GPU.

5.1 Experiment Settings

The Transformer model is a general sequence to sequence model with self-attention. We pass the input sentence through 5 layers of encoder stacked on top of each other that generates an output for each word/token in the sentence and 5 decoder layers to the encoder's output with its own input (self-attention) to predict the next word. The model handles variable-sized input using self-attention heads. We use 8 attention heads for both the encoder and the decoder. Similarly, the number of embedding dimensions and inner-layer dimensions used are 512 and 2048, respectively. We train the model with a learning rate of $7e{-}4$, the minimum learning rate being $1e{-}9$ for 150 epochs with a batch size of 64, updating a checkpoint after every 10 epochs. We set dropout, weight decay, and label smoothing to be 0.4, 10^{-4}, and 0.2 respectively. Adam optimizer with betas (0.9, 0.98) is used to optimize the model.

5.2 Data and Pre-processing

The corpus is developed from scratch accruing precise data directly from community linguists, which contains Nepali (Nep) and Tamang (Tamg) translation aligned at the sentence level; though the corpus is directly collected from community linguists, many instances of repeated sentence pairs were found, all such repeated sentence pairs are removed programmatically keeping alignment intact. After the removal of repeated sentence pairs, there are around 8243 + 3622 training sentence pairs, where 3622 sentence pairs are used for validation/development. We report results on the test set containing 3023 sentence pairs. The **Train**, **Test** and **Valid** dataset split is shown in Table 2.

	Sentence Pairs
Train	8243
Test	3023
Valid	3622

Table 2: Train, Test and Valid dataset split

IndicNLP[6] library (Kunchukuttan, 2020) is used to normalize and tokenize both Nepali and Tamang language texts. Translation, although being an open vocabulary problem, it is not possible to feed all the possible words in a language into a model.

[4] https://github.com/pytorch/fairseq
[5] https://colab.research.google.com/
[6] https://github.com/anoopkunchukuttan/indic_nlp_library

Translation Direction	Test	Valid	Vocabulary Size
Nepali→Tamang	**27.74**	29.41	1000
Tamang→Nepali	**23.74**	24.91	
Nepali→Tamang	27.33	29.07	2500
Tamang→Nepali	22.41	23.75	
Nepali→Tamang	26.13	27.53	5000
Tamang→Nepali	22.08	22.95	

Table 1: BLEU scores for Test and Valid set trained for 150 epochs using multiple vocabulary sizes.

Sennrich et al. (2015) describes the capability of open-vocabulary translation by encoding rare and unknown words as a sequence of subword units. For learning the vocabulary of source and target language, we use BPE[7] (Gage, 1994). BPE is the standard technique in Neural Machine Translation (NMT) and has been applied successfully in many systems. Ding et al. (2019); Gupta et al. (2019) argues that the impact of vocabulary size is significant in a low resourced dataset and that optimal BPE for Transformer architectures is small for low resource languages. Thus, based on the findings of Ding et al. (2019); Gupta et al. (2019), we chose the vocabulary size as $\leq$ 5000. Tokens not included in the vocabulary are replaced by universal tokens <unk>. Sentencepiece[8] library (Kudo and Richardson, 2018) is used to learn BPE in both the source and the target languages.

6 Results

After training all the models for 150 epochs, the best performing checkpoint[9] is used for each model with beam size of 5 (Yang et al., 2018) and length penalty of 1.2 to measure the translation quality using the BLEU[10] metric. The BLEU score (Papineni et al., 2002) is a general way to evaluate the performance of machine translation system when there can be multiple right outputs. Sacrebleu[11] (version 1.4.10) (Post, 2018) is used to compute the BLEU scores. BLEU scores obtained after training on multiple vocabulary sizes (1K, 2.5K and 5K) are shown in Table 1.

The model trained with vocabulary size of 1000 performs better among all and was able to obtain a BLEU score of **27.74** in Nepali→Tamang direction and **23.74** in Tamang→Nepali direction. Translation examples by the system generated by applying the best performing model are as follows:

Nepali→Tamang

Source	बच्चाहरुको निहुमा मुखमा आएजति शब्दमा अल्झेको मायाको आमा पनि घरको झ्यालमा बसेर दुईटी बच्चीलाई हेरिरहेकी थिईन् ।
Reference	कोलाकादेला निउरि सुङरि खातेदोना छिकरि हाल्बा मायाला आमा नोन दिमला झ्यालरि चिसि कोला ङिदान च्यासि चिबा मुबा ।
System	कोलाकादेला कुरि हापख्याइरि छिकरि आझोबा मायाला नोन दिमला झ्यालरि चिसि कोला ङहिदा च्याचिबा मुबा ।

Source	क्या महान् ईश्वर जसले थर्थरी कामेर संत्रस्त भएका अर्जुनमा शौर्य भर्दछ र कस्तो ठुलो योद्धा जो समाप्त भैसकेको शत्रुमाथि आक्रमण गर्दछ !
Reference	तिला ग्रेन ला थेसे लोङसि लगलग दार्बा अर्जुनरि पराक्रम युला ओम खाराङबा ग्रेन योद्धा जो जिनजिन्बा सत्तुरफिरि आक्रमण लाला !
System	क्या महान ग्लुसि निससे च्याङना गे लासि संत्रस्त ताबा अर्जुनरि शौर्य भर्दाम्ला ओम खाराङबा ठुलो चु जोद्धा चु जोप्त तासि जिन्बा ल्हुइरि आक्रमण लाला !

Source	त्यो इ.पू. २०० र सन् ३०० का बीचको अवधि– पनि चाखलाग्दो छ र यस अवधिका महत्वपूर्ण लक्षणहरु छन् ।
Reference	थे इ.पू. २०० थेन सन् ३०० ला गुङला दुइ– नोन चाखलाग्दो मुला ओम चु दुइला महत्वपूर्ण लक्षणजुगु मुला ।
System	थे इ . पू . २०० थेन सन् ३००० दुइला अवधि– नोन चालाबा मुला ओम चु अवधिला महत्वपूर्णजुगु लक्षणजुगु मुला ।

[7]Byte-Pair Encoding: A data compression technique

[8]https://github.com/google/sentencepiece

[9]Fairseq saves best performing checkpoint as "checkpoint_best.pt"

[10]BLEU: Bilingual Evaluation Understudy

[11]https://github.com/mjpost/sacreBLEU

Tamang→Nepali

Source	चुरि पाङ्बा मुतेबा अहंकार छ्याम चु वक्तव्यसे देन्बा दुइदा उन्बा मुला ।
Reference	यहाँ व्यक्त सम्पूर्ण अहंकार सहित यस वक्तव्यले वास्तविक अवस्थालाई देखाएको छ ।
System	यहाँ बोल्ने सबै अहंकार सबसँगै वक्तव्यले साँच्चालाई देखाएको छ ।

Source	शताब्दीयौं हेन्सेला गीताला अस्ङख्य थेन ल्होलो किसिमला फ्राल्तामन्हाङरि खाजिबाइ याखार्गिसे नोन चु मूलभूत रुपरेखादा अतिक्रमण लाबा आरे ।
Reference	शताब्दीयौंदेखिका गीताका असङ्ख्य र विविध किसिमका व्याख्यामध्ये कुनै एउटाले पनि यस मूलभूत रूपरेखालाई अतिक्रमण गरेको छैन ।
System	शताब्दीयौंदेखि गीताको अस्ख्य र विभिन्न किसिमको फ्याल्तमा आइसकेको एउटा मूलभूत रूपले पनि यो मूलभूत रूपलाई अतिक्रमण गर्ने छैन ।

Source	विडम्बनावस्, कम्युनिष्ट हिन्ना बिबा थेनोन जीवजुगुला चोथेबा जमात तिनि थोक हिसाबसे सढाउ फुमरि स्खलित ताजिबा मुला ।
Reference	विडम्बनावस्, कम्युनिष्ट हौं भन्ने उनै जीवहरूको यत्रो जमात आज थोक हिसाबले सढाउ अण्डामा स्खलित भइरहेको छ ।
System	विडम्बनावस् , कम्युनिष्ट हो भन्ने तिनै जीवनहरूको यत्र ज्ञान आज लिएर हिसाबले सढाउ फुत्तमा स्खलित भएका छन् ।

7 Discussion

For a low-resource language like Tamang, which has very negligible digital footprints, the achieved BLEU scores are quite encouraging. Besides, the gathered dataset (approx. 15K) can serve as a benchmark data for the Tamang - Nepali MT. We consider this as a significant contribution and achievement for the language pair. The deployed MT system is made available in the link[12] and is open for testing. We are really encouraged by the preliminary feedback that we have received from the community.

8 Future Plans and Road Map

We are in touch with the linguists and experts from the community and working towards further increasing the dataset. Similarly, we will also be looking further into how the MT model can be further optimized via hyper-parameter tuning, vocabulary size modification, changes in the learning rate etc.

The Tamang community has been involved in the testing of the system throughout and they are quite satisfied with the gradual improvement of the quality of the system. We plan to increase the size of the parallel sentences to 100K via crowdsourcing and other methods. We also plan to extend the bi-lingual MT system to a trilingual one incorporating English to the system so that the Tamang and the Nepali community benefit from the knowledge sources available in English on the Internet.

[12] https://translation.ilprl.ku.edu.np/

Acknowledgments

We would like to thank the **Information and Language Processing Research Lab, Kathmandu University** for providing the research opportunity and platform for this work. Similarly, our sincere thanks goes to Mr. Amrit Yonjan-Tamang and his team from Tamang Nangkhor for their support in the parallel corpus development. We would also like to extend our gratitude to Mr. Dawa Tamang, Virginia, US for providing support to the work.

References

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning.

Rohit Gupta, Laurent Besacier, Marc Dymetman, and Matthias Gallé. 2019. Character-based nmt with transformer.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/

indic_nlp_library/blob/master/docs/indicnlp.pdf.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3054–3059, Brussels, Belgium. Association for Computational Linguistics.