Neural Machine Translation: Hindi ⇔ Nepali

Sahinur Rahman Laskar, Partha Pakray and **Sivaji Bandyopadhyay**
Department of Computer Science and Engineering
National Institute of Technology Silchar
Assam, India
{sahinurlaskar.nits, parthapakray, sivaji.cse.ju}@gmail.com

Abstract

With the extensive use of Machine Translation (MT) technology, there is progressively interest in directly translating between pairs of similar languages. Because the main challenge is to overcome the limitation of available parallel data to produce a precise MT output. Current work relies on the Neural Machine Translation (NMT) with attention mechanism for the similar language translation of WMT19 shared task in the context of Hindi-Nepali pair. The NMT systems trained the Hindi-Nepali parallel corpus and tested, analyzed in Hindi ⇔ Nepali translation. The official result declared at WMT19 shared task, which shows that our NMT system obtained Bilingual Evaluation Understudy (BLEU) score 24.6 for primary configuration in Nepali to Hindi translation. Also, we have achieved BLEU score 53.7 (Hindi to Nepali) and 49.1 (Nepali to Hindi) in contrastive system type.

1 Introduction

MT acts as an interface, which handles language perplexity issues using automatic translation in between pair of diverse languages in Natural Language Processing (NLP). Although, corpus-based based MT system overcome limitations of rule-based MT system such as dependency on linguistic expertise, the complexity of various tasks of NLP and language diversity for Interlingua-based MT system (Dave et al., 2001). But it needs sufficient parallel corpus to get optimize MT output. The NMT falls under the category of corpus-based MT system, which provides better accuracy than Statistical Machine Translation (SMT), corpus-based MT system. The NMT system used to overcome the demerits of SMT, such as the issue of accuracy and requirement of large datasets. Recurrent Neural Network (RNN) encoder-decoder NMT system, which assists encoding of a variable-length source sentence into a fixed-length vector and same is decoded to generate the target sentence (Cho et al., 2014). The simple RNN adopted Long Short Term Memory (LSTM), which is a gated RNN used to improve the translation quality of longer sentences. The importance of LSTM component is to learn long term features for encoding and decoding. Besides, LSTM, other aspects that improve the performance of the NMT system like the requirement of test-time decoding using beam search, input feeding using attention mechanism (Luong et al., 2015). The reason behind the massive unfolding of the NMT system over SMT is the ability of context analysis and fluent translation (Mahata et al., 2018; Pathak and Pakray, 2018; Pathak et al., 2018).

Motivated by the merits of the NMT over other MT systems and the importance of direct translation in between pairs of similar languages, current work has investigated similar language pair namely, Hindi-Nepali, for translation from Hindi to Nepali and vice-versa using the NMT system. Due to lack of background work of similar language pair translation, the specific translation work for Hindi ⇔ Nepali is still in its infancy. To examine the efficiency of our NMT systems, the predicted translations exposed to automatic evaluation using the BLEU score (Papineni et al., 2002).

The rest of the paper is structured as follows: Section 2, details of the system description is presented. Section 3, result and analysis are discussed and lastly, Section 4, concludes the paper with future scope.

2 System Description

The key steps of system architecture are data preprocessing, system training and system testing and same have been described in the subsequent subsections. We have used OpenNMT (Klein et al.,

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 202–207
Florence, Italy, August 1-2, 2019.

2017) and Marian NMT (Junczys-Dowmunt et al., 2018) toolkit to train and test the NMT system. The OpenNMT, an open source toolkit for NMT, which prioritizes efficiency, modularity and support significant research extensibility. Likewise, Marian, a research-friendly tookit based on dynamic computation graphs written in purely C++, which achieved high training and translation speed for NMT.

2.1 Data Preprocessing

During the preprocessing step, source and target sentences of raw data are tokenized using Amun toolkit and makes a vocabulary size of dimension 66000, 50000 for Nepali-Hindi parallel sentence pairs, which indexes the words present in the training process. All unique words are listed out in dictionary files. The details of the data set are discussed next.

Data The NMT system has been trained using parallel source-target sentence pairs for Hindi and Nepali, where Hindi and Nepali are the source and target language and vice-versa. The training corpus has been compiled manually by back-translation using Google translator[1] from the Wikipedia source of Hindi language,[2] Nepali language,[3] and source of Bible[4] and as well as dataset provided by the WMT19 organizer (Barrault et al., 2019). The test data provided by the organizer for Hindi to Nepali translation consists of 1,567 number of instances and for Nepali to Hindi translation consists of 2,000 number of instances, have been used to check the translational effect of the trained system. Also, validate using a subset of training corpus containing 500 instances. The details of the corpus statistics are shown in Table 1. The NMT system has been trained and tested in three different configurations such as Run-1, Run-2, and Run-3 using primary and contrastive system type, which are summarized in Table 2 and 3.

2.2 System Training

After preprocessing the data, the source and target sentences were trained using our NMT systems for translation prediction in case of both Hindi to Nepali and Nepali to Hindi. Our NMT systems adopted OpenNMT and Marian NMT to train parallel training corpora using sequence-to-sequence RNN having attention mechanism. In NMT system architecture, encoder and decoder are the main components of the system. The encoder consists of a two-layer network of LSTM units, having 500 nodes in each layer, which transforms the variable length input sentence of the source language into a fixed size summary vector. After that, a two-layer LSTM decoder having 500 hidden units, process the summary vector (output of encoder) to generate target sentence as output. Multiple Graphics Processing Units (GPU) were used to increase the performance of training. The minimum batch size is set to 2000 for memory requirements, a drop out of 0.1 and enable layer normalization, which guarantees that memory will not grow during training that result in a stable training run.

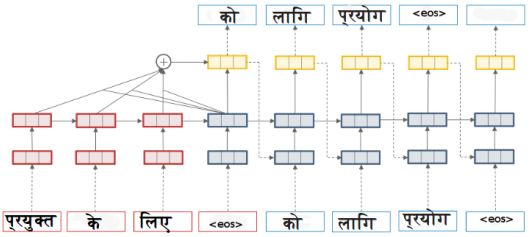

Figure 1: NMT System Architecture.

NMT System with Attention Mechanism The main disadvantage of the basic encoder-decoder model is that it transforms the source sentence into a fixed length vector. Therefore, there is a loss of information in case of a long sentence. The encoder is unable to encode all valuable information into the summary vector. Hence, an attention mechanism is introduced to handle such an issue. The encoder design is the main difference between basic encoder-decoder model and attention model. In the attention model, a context vector is taken as input by the decoder, unlike a summary vector in the basic encoder-decoder model. The context vector is computed using convex coefficients, are called attention weights, which measure how much important is the source word in the generation of the current target word.

Figure 1 presents the NMT system architecture, where attention mechanism and input feeding are used to translate Hindi source sentence "प्रयुक्त के लिए" into the Nepali target sentence "को लागि प्रयोग" (Luong et al., 2015). Here, $< eos >$ marks the end of a sentence.

[1] https://translate.google.com/
[2] https://en.wikipedia.org/wiki/Hindi
[3] https://en.wikipedia.org/wiki/Nepali_language
[4] https://www.bible.com

Nature of corpus	Name of Corpus/Source	Number of instances
Training	WMT19 Organizer	65,505
	Bible + Wikipedia (using Back-translation)	1,81,368
	Total	2,46,873
Test	Hindi to Nepali	1,567
	Nepali to Hindi	2,000
Validation	WMT19 Organizer	500

Table 1: Corpus Statistics.

Configuration	Tools	Training Data (No. of instances)
Primary (NMT-1): Run-1	Marian NMT	65,505 (WMT19 Organizer)
Contrastive (NMT-2): Run-2	OpenNMT	1,33,526 (65,505: WMT19 Organizer + Bible + Wikipedia)
Contrastive (NMT-3): Run-3	Marian NMT	2,46,873 (65,505: WMT19 Organizer + Bible + Wikipedia)

Table 2: Different configuration, tools and training data used for Hindi-Nepali Translation.

Configuration	Tools	Training Data (No. of instances)
Primary (NMT-1): Run-1	Marian NMT	65,505 (WMT19 Organizer)
Contrastive (NMT-2): Run-2	Marian NMT	1,33,526 (65,505: WMT19 Organizer + Bible + Wikipedia)
Contrastive (NMT-3): Run-3	OpenNMT	2,46,873 (65,505: WMT19 Organizer + Bible + Wikipedia)

Table 3: Different configuration, tools and training data used for Nepali-Hindi Translation.

2.3 System Testing

During system testing phase, the trained system is carried out on test sentences as mentioned in Section 2.1 provided by the WMT19 organizer for predicting translations.

3 Result and Analysis

The official results of the competition are reported by WMT19 organizer (Barrault et al., 2019) and the same are presented in Table 4, 5, 6 and 7 respectively.

A total of six, five teams participated in Hindi to Nepali and Nepali to Hindi translation using primary and contrastive system type. In the primary system type of Hindi to Nepali translation, our NMT system attained a lower BLEU score and a higher BLEU score in Nepali to Hindi translation than other participated teams. However, in both directions of Hindi-Nepali translation under contrastive configuration our system (Marian) obtained excellent BLEU score 53.7 (Hindi to Nepali), 49.1 (Nepali to Hindi). Moreover, it has been observed that our system's BLEU score of Marian outperforms OpenNMT in both directions of Hindi-Nepali translation under contrastive as well as primary configuration.

Analysis To analyze the best and worst performance of our NMT system, considered the sample sentences from test data provided by the organizer and predicted target sentences on the same test data by our NMT system and Google translator. In the case of a short, medium, long sentences of best performance are given in Table 8, our NMT system provides a perfect prediction like Google translation for the given test sentences. In Table 9, the worst case prediction sentences are presented. In Segment Id = 136, our NMT system's prediction is wrong. The predicted target sentence is in a different language in Segment Id = 25 and also, in case of a long sentence as given in Segment Id = 153, the prediction is not precise. However, Google translation yields accurate prediction in the same sentences.

Segment Id=306: Source Language=Nepali, Target Language=Hindi	
Source Test Sentence	तपाईं यो खाता मेट्न निश्चित हुनुहुन्छ ?
Generated Target Sentence	क्या आप निश्चित हैं कि आप इस खाते को विलोपित करना चाहते हैं?
Google Translation	क्या आप वाकई इस खाते को हटाना चाहते हैं ?

Segment Id=1: Source Language=Nepali, Target Language=Hindi	
Source Test Sentence	तपाईंले उपलब्ध गराउनु भएको इमेल ठेगाना वैध छैन ।
Generated Target Sentence	आपने जो ईमेल पता दिया है वह वैध नहीं है.
Google Translation	आपके द्वारा प्रदान किया गया ईमेल पता मान्य नहीं है।

Segment Id=43: Source Language=Nepali, Target Language=Hindi	
Source Test Sentence	तपाईंले समर्थन गर्ने लगइन विधि प्रयोग गरी रहनुभएको र तपाईंको पासवर्ड सही रुपले लेख्नु भएको निश्चित गर्न जाँच गर्नुहोस् । याद गर्नुहोस्, धेरै पासवर्डहरू सम्वेदनशील हुन्छन्; तपाईंको क्याप्स लक चालु भैरहेको हुनसक्छ ।
Generated Target Sentence	जांच करें कि आपके कूटशब्द की वर्तनी शुद्ध है और फिर आप एक समर्थित लॉगिन विधि का प्रयोग कर रहे हैं. याद रखें कि बहुत से कूटशब्द कुछ बड़े-छोटे अक्षरों के प्रति संवेदनशील होते हैं; आपका कैप्सलाक ऑन हो सकता है.
Google Translation	यह पुष्टि करने के लिए जांचें कि आप लॉगिन विधि का उपयोग कर रहे हैं और अपना पासवर्ड सही तरीके से लिख रहे हैं। याद रखें, कई पासवर्ड पर्यायवाची हैं; आपका कैप्स लॉक चालू हो सकता है।

Table 8: Best Performance examples in Nepali to Hindi translation.

Segment Id=136: Source Language=Nepali, Target Language=Hindi	
Source Test Sentence	मेनु सिर्जना गर्न असफल (सम्भवत% 1 छुटेको वा बिग्रिएको हुनुपर्छ)
Generated Target Sentence	पियर से प्रमाण पत्र लेने में असमर्थ
Google Translation	मेनू बनाने में विफलता (शायद% 1 को हटाया जाना चाहिए या क्षतिग्रस्त होना चाहिए)

Segment Id=25: Source Language=Nepali, Target Language=Hindi	
Source Test Sentence	डा. जिव फाइल "% 1" म्याक्रो फाइल भएकाले यसले आकृति समावेश गर्दैन ।
Generated Target Sentence	Die Audiodatei für dieses Ereignis existiert nicht.
Google Translation	डॉ। लाइव फ़ाइल "% 1" एक मैक्रो फ़ाइल है क्योंकि इसमें आकृति शामिल नहीं है।

Segment Id=153: Source Language=Nepali, Target Language=Hindi	
Source Test Sentence	या त फोल्डरहरूलाई छुट्टाछुट्टै चयन गरेर, र/वा सबै स्थानीय फोल्डरहरू, सबै टाढाको फोल्डरहरू, वा दुवै चयन गरेर तपाईंले स्रोतको रूपमा कम्तीमा पनि एउटा फोल्डरलाई निर्दिष्ट गर्नु पर्दछ ।
Generated Target Sentence	आप कम से कम एक फ़ोल्डर को एक स्रोत के रूप में जरूर निर्दिष्ट करें. या तो फ़ोल्डर को निजी रूप से चुनते हुये, और / या सारे स्थानीय फोल्ड को चुनते हुये, सारे दूरस्थ फ़ोल्डर , या दोनों.
Google Translation	आपको या तो फ़ोल्डर को अलग-अलग, और / या सभी स्थानीय फ़ोल्डर, सभी दूर फ़ोल्डर, या दोनों का चयन करके स्रोत के रूप में कम से कम एक फ़ोल्डर निर्दिष्ट करना होगा।

Table 9: Worst Performance examples in Nepali to Hindi translation.

Moreover, the BLEU scores of the test set translated by the Google translator with the test set provided by the organizer show close to each other for both target language Hindi and Nepali, as shown in Table 10.

Target Language	BLEU Score
Hindi	0.405171
Nepali	0.332679

Table 10: BLEU scores of Hindi and Nepali target language for test data and test set translation by Google translator.

Team	BLEU Score	Type	System
Panlingua-KMI	11.5	Primary	PBSMT
CMUMEA N	11.1	Primary	AUGTRAN
TeamZeroGang	8.2	Primary	-
NITS-CNLP	**3.7**	**Primary**	**NMT-1 (Marian)**

Table 4: BLEU scores result of participated teams at WMT19 shared task in Hindi to Nepali translation.

Team	BLEU Score	Type	System
NITS-CNLP	**24.6**	**Primary**	**NMT-1 (Marian)**
CMUMEA N	12.1	Primary	AUGTRAN
Panlingua-KMI	9.8	Primary	PBSMT
TeamZeroGang	9.1	Primary	-
CFILT_IITB	2.7	Primary	WITH MONOLINGUAL

Table 5: BLEU scores result of participated teams at WMT19 shared task in Nepali to Hindi translation.

Team	BLEU Score	Type	System
NITS-CNLP	**53.7**	**Contrastive**	**NMT-3 (Marian)**
TeamZeroGang	8.2	Contrastive	-
NITS-CNLP	**3.6**	**Contrastive**	**NMT-2 (OpenNMT)**
CFILT_IITB N	3.5	Contrastive	Basic

Table 6: BLEU scores result of participated teams at WMT19 shared task in Hindi to Nepali translation.

Team	BLEU Score	Type	System
NITS-CNLP	**49.1**	**Contrastive**	**NMT-3 (Marian)**
TeamZeroGang	9.1	Contrastive	-
Panlingua-KMI	4.2	Contrastive	NMT
Panlingua-KMI	3.6	Contrastive	NMT-Transformer
NITS-CNLP	**1.4**	**Contrastive**	**NMT-2 (OpenNMT)**

Table 7: BLEU scores result of participated teams at WMT19 shared task in Nepali to Hindi translation.

4 Conclusion and Future Scope

In this work, our NMT systems adopted attention mechanism to predict translation of similar language pair namely, Hindi to Nepali and vice-versa. In the current competition, in primary configuration, our NMT system obtained BLEU score 24.6 in Nepali to Hindi translation and BLEU score 3.7 in Hindi to Nepali translation. On the other hand, in contrastive configuration, our NMT system acquired BLEU score 53.7 (Hindi to Nepali), 49.1 (Nepali to Hindi). However, close analysis of generated target sentences on given test sentences remarks that our NMT systems need to improve in case of wrong translation, translation in a different language. Moreover, BLEU scores presented in Table 10, pointed out that is case of both target language Hindi and Nepali, the scores are in relatively stable in both directions of Hindi-Nepali translation like our systems (both Marian and OpenNMT) in contrastive configuration (as mentioned in Table 6 and 7) but unlike in primary configuration (Marian) (as mentioned in Table 4 and 5). Hence, more experiments and comparative analysis will be needed in future work to reason about Marian outperforms OpenNMT in both directions i.e. Hindi to Nepali and Nepali to Hindi translation. In the future work, more number of instances in Hindi-Nepali pair, different Indian similar language pair like Bengali-Assamese, Telugu-Kannada, Hindi-Punjabi, shall be considered for machine translation, which may be possible to overcome the limitation of available parallel data to produce precise MT output.

Acknowledgement

Authors would like to thank WMT19 Shared task organizers for organizing this competition and also, thank Centre for Natural Language Processing (CNLP) and Department of Computer Science and Engineering at National Institute of Technology, Silchar for providing the requisite support and infrastructure to execute this work.

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation, Volume 2: Shared Task Papers*, Florence, Italy. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Shachi Dave, Jignashu Parikh, and Pushpak Bhattacharyya. 2001. Interlingua-based english–hindi machine translation and language divergence. *Machine Translation*, 16(4):251–304.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Sainik Kumar Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2018. Mtil2017: Machine translation using recurrent neural network on statistical machine translation. *Journal of Intelligent Systems*, pages 1–7.

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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Amarnath Pathak and Partha Pakray. 2018. Neural machine translation for indian languages. *Journal of Intelligent Systems*, pages 1–13.

Amarnath Pathak, Partha Pakray, and Jereemi Bentham. 2018. English–mizo machine translation using neural and statistical approaches. *Neural Computing and Applications*, 30:1–17.