# A Comparative Study of SMT and NMT: Case Study of English-Nepali Language Pair

*Praveen Acharya*[1], *Bal Krishna Bal*[2]

[1]Kathmandu University, Nepal
[2]Kathmandu University, Nepal
acharyaprvn@gmail.com, bal@ku.edu.np

## Abstract

Machine Translation is one of the major problems in the field of Natural Language Processing. In due course, many approaches have been applied in order to solve this problem ranging from the traditional rule-based approach, statistical methods to the more recent neural network based methods.

Neural network based methods have produced comparable results to that of the existing phrase based model and in some language pairs they have even outperformed the latter.

A huge amount of parallel corpus is required for both the SMT and NMT models in order to produce reasonable results. For some language pairs, the required data is readily available but for many others it is not necessarily the case.

This research work focuses on the comparative study of how SMT and NMT based machine translation models perform and compare to each other in case where the language pair is under resourced in terms of the availability of the parallel corpus.

**Index Terms**: SMT, NMT, English-Nepali language pair

## 1. Introduction

Machine translation is a field of study focused on primarily building a system that is capable of translating from one language to another accurately and without losing information. Traditional systems using dictionaries and rule based ordering of words have not been able to achieve the required level of accuracy. Hence, corpus and statistical techniques have led to developing systems that result in better translations.

The process of translation usually consists of reading the input language (source language); applying certain approaches to translate the language and produce the corresponding output language (target language). Current machine translation systems are nowhere near of providing quality translations as compared to a human translator. But; there has been improvement with the current statistical machine translation system and more recent neural machine translation that is able to produce an understandable translation though not accurate.

### 1.1. Nepali Language

Nepali is one of the official languages of Nepal. It is the mother tongue of little more than half of the population of Nepal and a lingua franca for the rest. Nepali is also spoken in the neighboring countries of Nepal like India, Bhutan and Burma. The Nepali language belongs to the Indo-Aryan branch of the Indo-European language family. It is written in the Devanagari script and is a free word order language. There are 11 vowels and 33 consonants in Nepali. It is a highly inflectional and agglutinative language [1].

Machine translation research for English-Nepali language pair is limited to just a few research projects in the past. Nepali is a morphologically rich but at the same time an under resourced language in terms of the availability of language resources like the corpus.

Today, there is abundant information available in the internet but the problem is that the information is largely in English which is a language barrier for many people not knowledgeable in English. This clearly suggests the need of a machine translation system so that there is a wider access to information among the global audience. This applies equally well to Nepali as well.

### 1.2. Statistical Machine Translation (SMT)

SMT is the most researched approach and is still used in present day machine translation systems that have been developed. In statistical machine translation system, the goal of the system is to find a translation *f* given a source sentence *e*, which maximizes

$$p(f|e) \propto p(e|f)p(f) \quad (1)$$

Where, $p(e|f)$ is called the translation model and $p(f)$ the language model [10]. In practice, however, most SMT systems model *log p(f|e)* as a log-linear model with additional features and corresponding weights. There are also different models to SMT including word-based model, phrase-based model and syntax-based model. And currently the phrase-based SMT system produces better translation.

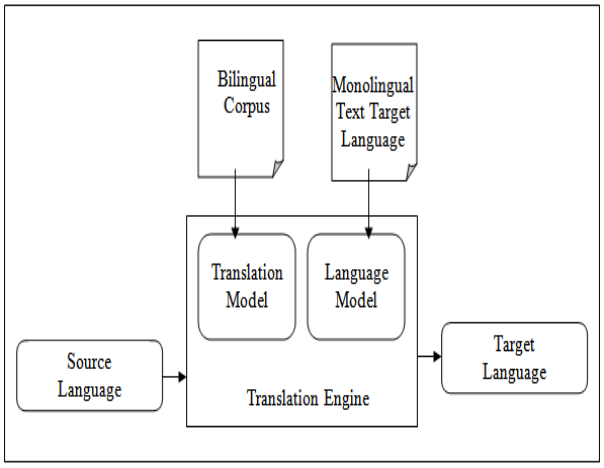


Figure 1: *SMT Framework*


10.21437/SLTU.2018-19

### 1.3. Neural Machine Translation (NMT)

More recently neural network based machine translation system is being researched extensively and has been delivering promising results at least in the case where huge amount of parallel corpus is readily available. Using neural network to learn statistical models was achieved in [2]. Recurrent neural networks have been used successfully for natural language processing tasks and has achieved close to the state-of-the-art accuracy in machine translation [3].

Works in NMT are based on recurrent neural network and convolutional deep belief network. Deep learning approaches often take an encoder-decoder approach to learn translation. In this model, an encoder neural network reads a source sentence and encodes it into a fixed length vector. A translation is then made by the decoder, which decodes the fixed length vector into a sentence of variable length in the target language. The system is trained to maximize the conditional probability of a correct translation given the source sentence.

Until recently research efforts in statistical machine translation have relied on the use of phrase-based approach as described in [4]. However, recently an entire new deep learning approach has been proposed which has produced more promising results compared to the existing phrase-based statistical machine translation [5] [6] [7].

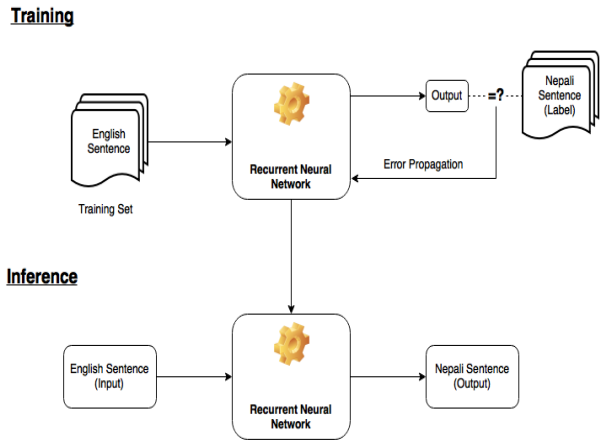


Figure 2: *NMT Framework*

## 2. Previous Works in the English-Nepali Language Pair

The first ever work on English to Nepali machine translation was conducted under the "Dobhase" Project [8]. It was a rule-based MT system. Further works on enhancement of Dobhase was also undertaken [9]. Another MT system that translates between English-Nepali has been developed by Google and is popularly known as Google Translate [10]. It is a free multilingual machine translation service provided by Google.

### 2.1. Dobhase

It is a rule-based MT system which was developed in 2006. The system takes a sentence in the source language, analyzes it, parses it to form a parse tree and then generates syntax of target language and finally applies morphology generation rules to form sentence in target language. The system comprises of modules as shown in Figure 3.

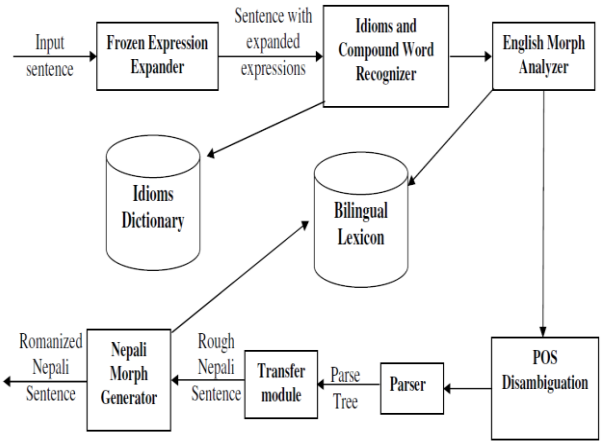


Figure 3: *Dobhase system architecture*

### 2.2. Dobhase Enhancement

Works on the enhancement of Dobhase was undertaken to improve the existing Dobhase system. The enhancements were accomplished by building an independent module on top of the Dobhase system. Even after the improvements, the system was still unable to resolve ambiguous words, handle multiple conjunctions and was also unable to handle single word joiner.

### 2.3. Google translate

Google Translate is a free multilingual machine translation service provided by Google to translate text, speech, images, or real-time video from one language into another.

Nepali language support was added to Google translate in the 36th stage which was launched in December 2013 and made publicly available. Google Translate has been trying to improve the translation quality by involving native speakers in the correction of translated output and its verification in order to improve the translation quality.

## 3. Experiments and Results

### 3.1. Data Set

For conducting the experiments regarding the given work, we borrowed the parallel corpus distributed as part of the Nepali National Corpus (NNC) [11]. This corpus consists of a collection of national development texts in English and Nepali. Apart from the parallel corpus provided under NNC corpus; we have collected additional sentences with the help of linguists. This data set was manually cleaned and a sentence level parallel corpus of 6535 sentences was created. This parallel corpus was then randomly shuffled and divided into 3 parts namely training set, development set and testing set. Each set of the corpus consists of respectively 5724, 358 and 453 English-Nepali parallel sentences. Additionally, 503 English sentences were collected and categorized as the second test set in order to test how the performance of the systems could be generalized.

### 3.2. Data Preprocessing

The parallel corpus consists of two separate text files. One file consists of the English sentences and the corresponding line in the other text file is the translation of that sentence in Nepali.

For the preprocessing step the sentences have been lower cased and tokenized into tokens and then the tokenized sentences are used for the purpose of training both the SMT and NMT systems.

### 3.3. SMT system

We trained Phrase based SMT system with Moses [12]. We used GIZA++ toolkit for word aligning the parallel corpus. We trained 5-gram language models with Kneser-Ney smoothing using KenLM [13]. Both GIZA++ and KenLM is included within MOSES. We used Nepali segments that were used to train the system as the monolingual corpora to train the language model.

### 3.4. NMT system

We used lamtram[14] toolkit for training our NMT system. We trained an encoder-decoder architecture based NMT system with attention [15].The network contains a single hidden RNN layer, containing 100 LSTM unit. The model is trained with a batch size of 256 sentences using Adam optimizer [16] with a learning rate of 0.001. We used early-stopping as well as rate decay of 0.5. We also used regularization by setting dropout rate of 0.5. We ran the experiment for 10 epochs. For decoding we used a beam size of 10 with word penalty of 1.

### 3.5. Evaluation

BLEU [17] is the most commonly and widely used evaluation metric in machine translation. It is one of the first metrics to report high correlation with human judgment of quality and therefore, one of the most popular in the field of MT. BLEU metric calculates scores for an individual segment, generally sentences; the final score is calculated by averaging over the whole corpus. It has been shown to correlate highly with human judgments of quality at the corpus level. No other machine translation metric is yet to significantly outperform BLEU with respect to correlation with human judgment across language pairs [18].

The translated output from the both the systems have been detokenized and then the *multi-bleu.perl* script available in the MOSES Toolkit has been used to get the BLEU score for this purpose. Table 1 shows the results of our experiments.

### 3.6. Results

Table 1: *BLEU Scores.*

| System | TestSet1 | TestSet2 |
|---|---|---|
| SMT | **5.27** | **2.51** |
| NMT | 3.28 | 1.73 |
| Google Translate | 4.68 | 6.28 |

## 4. Discussion

SMT and NMT perform well when we have a large amount of parallel corpus available to train the systems. They do not perform as good when there is a limited amount of parallel corpus as is evident from the results. English-Nepali language pair being under resourced in terms of the availability of parallel corpus, the performance of both the systems is not that significant. But, if we compare both the systems trained using limited corpus we can see that SMT system outperforms the NMT based system in both the test sets. As can be seen from Table 1, it is also worthwhile to note that Google Translate which is a NMT based system outperforms both the SMT and NMT system for TestSet2. This result can be attributed to Google having larger training corpus to build their NMT systems and also, the nature of TestSet2 which was collected exclusively and not as part of an existing parallel corpus dataset. So, given that we have a large dataset that we can use to build our system and we want it to be able to generalize better, NMT is the better option.

Table 2 and Table 3 show some examples of the test sentences and their corresponding translations generated by the SMT and NMT systems. The words in the translated output that are present in the reference Nepali sentence have been highlighted.

Table 2: *Sample Translation result for TestSet2*

| **Example 1** | |
|---|---|
| English Sentence: | We were poor. |
| Nepali Sentence: | हामी गरिब थियौँ। |
| SMT Translation: | हामीले **गरिब** थिए । |
| NMT Translation: | हामी **गरिब** गरिब थिए । |

| **Example 2** | |
|---|---|
| English Sentence: | We'll always be under pressure. |
| Nepali Sentence: | हामीहरू सधैँ दबाबमा हुन्छौँ। |
| SMT Translation: | हामी **सधैँ दबाबमा** साथसाथै हुन सक्छ। |
| NMT Translation: | हामी **सधैँ** सधैँ हुन सक्छ। |

Table 3: *Sample Translation result for TestSet1*

| **Example 1** | |
|---|---|
| English Sentence: | Another major cause of accidents is the poor maintenance of vehicles. |
| Nepali Sentence: | दुर्घटनाको अर्को महत्त्वपूर्ण कारण सवारी साधनहरू उचित मर्मत सम्भार नगर्नु हो। |
| SMT Translation: | **अर्को** मुख्य **कारण** दुर्घटनाका पुअर्स को **सवारी हो।** |
| NMT Translation: | **अर्को** मुख्य **कारण** दुर्घटनाका गरिब छ। |

| **Example 2** | |
|---|---|
| English Sentence: | The security system in the country should be very good. |
| Nepali Sentence: | देशमा सुरक्षा प्रणाली निकै राम्रो हुनु पर्दछ। |
| SMT Translation: | यो **देशमा सुरक्षा प्रणाली निकै राम्रो** छ। |
| NMT Translation: | देशको **सुरक्षा** देशको **निकै राम्रो** हुन्छ। |

## 5. Conclusion

Researchers have been trying to find ways to improve upon the existing machine translation systems with the goal of being able to build a translation system that can efficiently and effectively translate sentences form source to target language. Several approaches have been used in developing such a system and new approaches are being studied so as to achieve the goal. Deep learning approach is the more recent idea that is being researched and the initial finding makes the approach

seem impressive. However, we can see that in the case where the language pairs are under resourced SMT still outperforms the NMT system. Therefore, further investigation into statistical as well as deep learning approaches for under resourced language pair is necessary for building a MT system with better performance for such language pairs.

## 6. Acknowledgements

We would like to thank NVIDIA for providing us with the Titan X (non-pascal) GPU for the research. This work is being conducted at the Information and Language Processing Research Lab, Kathmandu University within the scope of a Masters thesis.

## 7. References

[1] A. R. Dahal, Development of a Nepali-English MT system Using the Apertium MT platform, The Language Technology Kendra, July 2011.

[2] Y. Benigo, et al. "A neural probabilistic language model." The Journal of Machine Learning Research 3 (2003): 1137-1155

[3] M. Auli, et al. "Joint Language and Translational Modeling with Recurrent Neural Networks." EMNLP. 2013

[4] P. Koehn, Statistical machine translation, Cambridge University Press, 2009.

[5] K. Cho, B. V. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Benigo. 2014b. Learning phrase representations using rnn encoder-decoder for statistical machine tanlsation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014), October

[6] I. Sutskever, O.Vinyals, and Q. Le. 2014. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems (NIPS 2014), December.

[7] K. Cho, B. v. Merrienboer, D. Bajdanau and Y. Benigo 2014a. On the properties of neural machine translation: Encoder-Decoder approaches. In Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, October.

[8] S.K. Bista, B. Keshari, J. Bhatta and K.Parajuli, Dobhase: online English to Nepali Machine Translation System, In the proceedings of the 26th Annual conference of the Linguistic Society of Nepal, December 2005.

[9] U. Tandukar, C. Thapa, Project report on Enhancements on Dobhase English-Nepali Machine Translation System, July 2007.

[10] Google Translate, https://en.wikipedia.org/wiki/Google_Translate

[11] Corpora, Volume 3 Issue 2, Page 213-225, ISSN 1749-5032 Available Online Feb 2009, https://doi.org/10.3366/E1749503208000166

[12] P. Koehn et al. Moses: Open source toolkit for statistical machine translation. ACL, 2007

[13] Heafield, K, KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation, 2011

[14] Neubig, G., *lamtram: A Toolkit for Language and Translation Modeling using Neural Networks*, 2015. Retrieved from http://www.github.com/neubig/lamtram

[15] Bahdanau, D., Cho, K. and Bengio, Y., Neural machine translation by jointly learning to align and translate. International Conference on Learning Representations, 2015.

[16] Kingma, D. and Ba, J., Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2014.

[17] Papineni, K., Roukos, S., Ward, T., and Zhu, W. J., "BLEU: a method for automatic evaluation of machine translation" in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318

[18] Graham, Y. and T. Baldwin, "Testing for Significance of Increased Correlation with Human Judgment". Proceedings of EMNLP 2014, Doha, Qatar