International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Khasi to English Neural Machine Translation: an Implementation Perspective

N. Donald Jefferson Thabah, Bipul Syam Purkayastha

Abstract: ***Being able to translate and communicate consistently from one language to another would have been the ultimate goal of an intelligent system. With recent advancement of Neural Machine Translation (NMT), it has shown a promising solution to the problem of machine translation. NMT generally requires large size parallel corpora to obtained a good translation accuracy. In this paper, we would like to explore a Translation system from Khasi to English language using both supervised and unsupervised technique. Unsupervised was inspired to help attaining a better translation accuracy for low resource language. It was influenced by the recent advancement of unsupervised neural machine translation which primarily relies on monolingual corpora. In this work, Supervised NMT technique was also implemented and compared with the standard OpenNMT toolkit. Here, we also use Statistical Machine Translation (SMT) tools like Moses as a standard benchmark to compare the translation accuracy. When considering monolingual corpus, we obtain an accuracy of 0.23%. Given the small size monolingual corpus the result was lacking but showed promising rooms for improvement. We obtain much better accuracy of 35.35% and 41.87% when we use parallel corpus in supervised NMT and OpenNMT respectively. On comparison with SMT system with Blue score of 43.76%, the supervised NMT system was on par in its performance. Lastly, with improvement in corpus size and better adaptation of preprocessing steps on the source language (Khasi) the result can be tune to a better outcome.***

Keywords: ***Khasi to English NMT, machine translation, supervised NMT, unsupervised NMT .***

I. INTRODUCTION

Neural Machine Translation (NMT) has become a very popular mechanism for language Translation during these last few years [1], [6]. NMT is trained in an end to end approach for translating from one language to another [16], [6], [8]. Basically, NMT takes a conditional language model by modelling the probability of target sentence given a source sentence [11]. It has two main components, encoder which computes a hidden vector for each source sentence and a decoder then generates the translation. More precisely, an encoder-decoder architecture consists of a Recurrent Neural Network (RNN) and an attention mechanism that aligns target with source tokens [1], [11]. On recent findings, NMT has shown great improvement compared to Statistical Machine Translation Approach [27].

In this paper, a neural machine translation from Khasi [3] language to English was implemented on different scenario base on the type of corpus. NMT in general requires a large corpora size to train the model [18]. Unfortunately, for a Low Resource language it is slightly difficult to get access to such large corpora size, for example, Khasi language which is spoken in North Eastern State of Meghalaya, India is a low resource language.

The first aspect of this work considered only monolingual corpus to create a NMT system which was inspired by an Unsupervised Neural Machine Translation [12]. The basic aim is to use only monolingual corpus to train the unsupervised translation model. Unsupervised NMT remove the needs of cross lingual information [20] and make use of only monolingual corpus to train the translation model. This Unsupervised Neural Translation has been possible because of recent development in unsupervised cross-lingual embeddings [13], [14]. The second aspect of NMT system used parallel corpus to train a supervised translation model.

We performed a translation from Khasi to English. First, in an unsupervised translation techniques using only monolingual corpus with a seed dictionary. We obtain a Blue score [10] accuracy of 0.23%. Secondly, in a supervised neural translation approach using general domain parallel corpus. We are able to obtain an accuracy of 35.35% using NMT [12] and 41.87% using OpenNMT [29] toolkit. Finally, the model was also train in a phrased based statistical machine translation using Moses toolkit [28] and obtain an accuracy of 43.76%. The objective of using Moses tools for our work was to treat the result obtaining from Moses as a comparison benchmark with regard to our Neural Machine Translation system of Khasi to English language.

II. RELATED WORK

A. Unsupervised Neural Machine Translation

Unsupervised Translation system [12] was implemented to translate French to English and German to English using the WMT 2014 standard corpus and was able to attain a better outcome as compared to the previous existing system. The same work was extended by using a combination of monolingual and parallel corpus and was able to show impressive result.

B. Supervised Neural Machine Translation

NMT using attention model [1] has been a very successful improvement over the classic encoder decoder architecture which is a sequence to sequence recurrent neural network technique. Meaning, attention model is an improvement over the existing encoder decoder architecture.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
N. Donald Jefferson Thabah*, Department of Computer Science, Assam University, Silchar, India. E-mail: jefson08@gmail.com
Bipul Syam Purkayastha, Department of Computer Science, Assam University, Silchar, India. E-mail: bipul_sh@hotmail.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

It solve the problem of overloaded hidden vector encoding of a source sentence and thus help in efficient translation. As an outcome the translation accuracy was drastically improve specially in long source sentences.

OpenNMT [29], an open source tools for Neural Machine Translation. This toolkit was design for easy extensibility of its existing functionality. Due to its high community support, this is one of the current state of the arts open source tool for NMT system.

C. Statistical Machine Translation Moses Toolkit

Phrased based statistical machine translation is one of the most successful machine translation technique. It is still a popular and efficient tools for translation system specially for long source sentences. The paper [28] describe the details working of Moses and its corresponding translation example for translating English to Spanish and vice versa.

III. KHASI LANGUAGE

Khasi is the name of the language, the community, and its traditional religion as well. It stands as a cover term for many variations spoken in Khasi Hills, Jaintia Hills and the Ri Bhoi District of the State of Meghalaya [3]. The Khasi language belong to the Austro- Asiatic group of the Monkhmer language family [9]. Written Khasi literature started with the translation of the Bible and other Christian literature from Welsh and English into Khasi by Christian Missionaries. The latter part of the nineteenth century saw a number of translated works, poems, plays and fiction from other languages, by many Khasi writers. The twentieth century also witnessed a rapid growth in the number of books written in the Khasi language which include translation works from different languages. Dictionaries, book on grammar, poetry, fiction, drama, literary theory and criticism, science, geography and history are found in plenty in the Khasi language. During the latter half of the twentieth century many writers published books on linguistics, culture, folklore, environment, mathematics, science, geography and history in Khasi. This contributed to the introduction of Khasi as an academic subject in schools, then later in colleges, and in the university in the 1980s, few decades after the Roman script was adapted. Currently, The Khasi Language is used as a medium of instruction and a subject in Khasi medium Primary Schools. Standard Khasi has been accorded the status of an Associate Official Language in the State of Meghalaya in 2005.

IV. CORPUS

The Corpus for this work was mainly taken and constructed from the English-Khasi Bible[1]. Few parallel corpus was collected from Tatoeba project[2] which is an open and free collection of sentences and its translations. The remaining corpus were constructed from Books, newspaper, articles and so on. More precisely, monolingual corpus was mainly taken from the Khasi-English Bible and the parallel corpus was a combination of Khasi-English Bible and the remaining sets of resources as stated above.

V. IMPLEMENTATION

Below we describe the necessary details required for running our translation system. Our main focus is on both the unsupervised and supervised learning to create a translation model using the constructed corpus.

A. Cross-lingual embedding

The main building block of Unsupervised Neural Machine translation is the fixed cross-lingual embedding [12]. For the corpus considered in this paper, we obtained the fixed cross-lingual embedding using a method which was influenced by the work on unsupervised cross lingual embedding [13], [14]. Before generating the fixed-cross lingual embedding as depicted in Figure 1, we need to perform text preprocessing to clean up words e.g. stop words, special characters etc. which are not contributing in the embedding outcome. Further, the Bilingual word embedding takes the source and target language word embedding with an optional seed dictionary as input to produce a fixed cross-lingual embedding which are desired for training an unsupervised translation model.

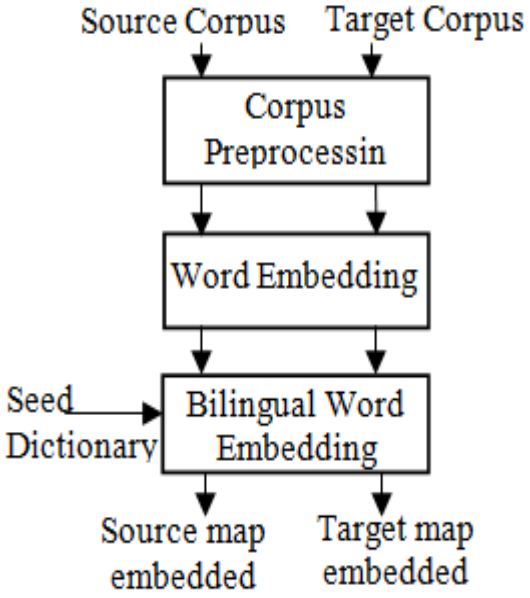

Fig. 1: Fixed Cross-lingual embedding

B. System Architecture for Unsupervised NMT

The architecture for this work was based on the unsupervised neural machine translation [12]. It follows standard encoder-decoder architecture with attention mechanism [1]. It uses a two layer bidirectional Recurrent Neural Net (RNN) in the encoder, another two layer RNN in the decoder and a global attention method proposed by Minh-Thang Luong and his colleague [11], with general alignment functionality. The encoder which is shared by both the source and target language [12] is input with a pre-trained cross-lingual embedding that are fixed during training.

[1] https://www.bible.com/en-GB/bible/296/GEN.INTRO1.GNB?parallel=1865

[2] http://www.manythings.org/anki/

The shared encoder aims at producing a language independent representation of the input text whereas the two separate decoders should then transform into its corresponding language.

The objectives of fixed embeddings in the encoder is to provides language independent word-level representations, and the encoder only requires to learn how to compose them to build representations of larger phrases.

NMT model are generally trained using parallel corpus to translate from source to target language [2]. The modified architecture made it possible to train the system using monolingual corpus only [12], hence, suitable for low resource language. The training is summarized in the following steps:

Denoising: The shared encoder and the dual structure decoder can be trained to recreate its own input [12]. Meaning, the system can be optimized to take an input sentence in a particular language, encode it using the shared encoder and rebuild the original sentence using the decoder of that language. Since the shared encoder is fed with a previously trained cross-lingual embeddings, the encoder should learn to create the embeddings of both languages in a language-independent manners, and each decoder should learn to decompose this representation into their own corresponding language. At inference time, simply replace the decoder with that of the target language, so it generates the translation of the input text from the language independent representation given by the encoder. The training procedure is merely a copying task and it is incapable to capture knowledge of the language involved. In order to learn the compositionality of its input words in a language independent fashion, random noise is introduced in the input sentence [1].

On-the-fly backtranslation. Denoising alone cannot fulfill the translation capability as the training is still a copying task. Hence back-translation approach [12] was used. Given an input in one language and then use inference mode with greedy decoding to translate to other language. This will create pseudo-parallel sentence pair, making use of this generated pseudo-parallel sentence we train the system to predict original sentence.

C. System Architecture for supervised NMT

The encoder decoder architecture is the standard method adopted for most of the neural machine translation [6], [8]. An encoder in a recurrent neural network takes a source sentence as input and transforms it into a fixed length vector. The decoder used this fixed length vector and translate it into a target sentence. This architecture in its base form overload the encoded hidden vector since the input sentence is encoded in just a single fixed length vector for example h_4 in its basic structure (below diagram) is overloaded by all the input words of a source sentence. In fact, all the hidden vector from each word of a input sentence are carrying some useful information for translating into a target sentence.

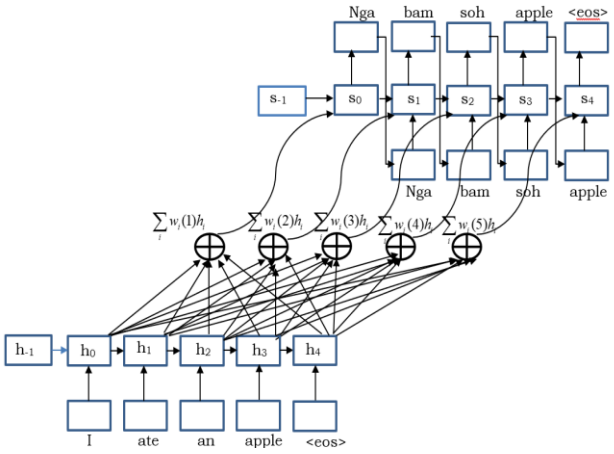

Fig. 2: Example of Attention
Taken from the lecture notes on "Sequence to Sequence models: Attention Models", 2018.[3]

In Fig. 2, h_{-1} is initialize to 0 and it is a learnable parameter, $s_{-1} = h_4$ in its simplest form and $\sum_i w_i(t)h(i)$ is the weighted combination of all the hidden outputs from an encoder which is inputted to the decoder. The weights $w_i(t)$ is a scalar value and must automatically highlight the most important input component during the decoding process so as to obtain the best possible translated sentence. Below equations briefly suggest the method to calculate the weights and its detail can be found from the work by Dzmitry Bahdanau [1].

$$e_i(t) = g(h_i, s_{t-1}) \quad (1)$$

$$[\, g(h_i, s_{t-1}) = h_i^T s_{t-1} \text{ in its simplest form}]$$

$$w_i(t) = \frac{\exp(e_i(t))}{\sum_j \exp(e_j(t))} \quad (2)$$

VI. EXPERIMENT SETUPS

Training and testing data size are summarized in Table-I and Table-II. These corpus are classified into two class. They are General Domain Monolingual corpus constructed from the Khasi to English Bible with a seed dictionary to improve it learning capabilities. The General Domain parallel corpus was extracted from Khasi to English Bible and other resources like books, newspaper, article etc.

Table- I: Training corpus size

	Khasi		English	
Corpus Type	***Size (No. of Lines)***	***Size in MB***	***Size (No. of Lines)***	***Size in MB***
General Domain Monolingual	38363	5	44869	4
General Domain Parallel	38697	3.8	38697	3.1

[3] https://www.youtube.com/watch?v=oiNFCbD_4Tk&t=17s

Table- II: Test corpus size

	Khasi		English	
Corpus Type	***Size (No of Lines)***	***Size in KB***	***Size (No of Lines)***	***Size in KB***
General Domain Monolingual	6302	722	6302	589
General Domain Parallel	12900	1300	12900	1000

VII. RESULTS AND DISCUSSION

To determine the accuracy of our model we performed a Blue score using Perl script from Moses tool. Based on the type of corpus namely General Domain Monolingual and General Domain Parallel corpus, four class of model were trained and the result are illustrated in the Table-III.

Table-III: Blue Score for Khasi to English Monolingual Translation Model

Corpus	Type	Khasi to English Translation		
		Test Data (%)	**Train Data (%)**	**Combine Test and Train Data (%)**
General Domain Monolingual	UNMT	0.23	0.17	0.28
General Domain Parallel	NMT	35.35	73.66	64.48
General Domain Parallel	OpenNMT	41.87	86.43	74.56
General Domain Parallel	SMT(Moses)	43.76	90.59	78.15

In Table-III, we calculate the accuracy of the translation model which were trained using General Domain Monolingual and General Domain Parallel corpus. The model were trained in four platforms. The first model, UNMT (Unsupervised Neural Machine Translation) [12] was implemented and trained using monolingual corpus whereas the remaining three model were based on parallel corpus. The second model i.e. second row of Table-III were implemented using the attention mechanism of NMT which is included as a separate module in Unsupervised Neural Machine Translation [12][4]. The third model was tested on OpenNMT toolkit [29] and the last model used statistical machine learning toolkit Moses [28]. The accuracy were performed by feeding three sets of input to these model. These inputs are testing data, training data and a combination of testing and training data, summarised in Table-II, training data had been used to train the model whereas here we used it to test the accuracy of the translation system and similarly we take a combination of both testing and training data as input in calculating Blue score. In Table-III, the Blue score for model trained using General Domain Monolingual corpus is 0.23% for testing data, 0.17% for training data and 0.28% for combination of testing and training data. The model trained using General Domain Parallel corpus yields the following accuracy. First, for the model train in supervised Neural Machine Translation system (second row of Table-III) produced an accuracy of 35.35 % for testing data, 73.66 % for training data and 64.48% for combination of testing and training data. Secondly, when model trained on OpenNMT toolkit we obtained a Blue Score of 41.87% for testing data, 86.43% for training data and 74.56% for combination of testing and training data. Lastly, model trained in SMT system using Moses yielded an accuracy of 43.76% for testing data, 90.59% for training data and 78.15% for combination of testing and training data.

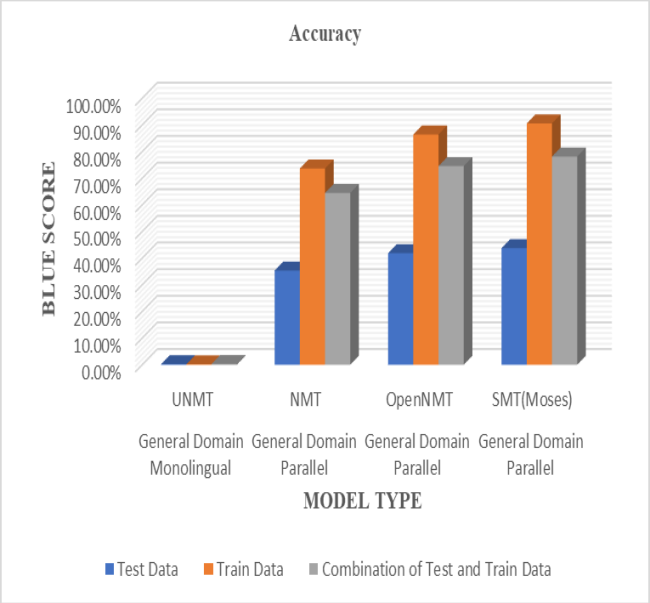

Fig.3: Accuracy Graph

Fig.3, shows the Blue score accuracy of the four different model namely (UNMT, NMT, OpenNMT, Moses). The accuracy of an UNMT was very low as compare to the other three model. This was deeply affected due to the monolingual small corpus size and improper preprocessing steps deploy for the source language (Khasi) but this does not discourage us in carrying out further work on this area as it open many opportunities on exploring NMT for low resource languages. Because of the latest advancement in NMT research we could see that the model train in supervised NMT system has a competitive results to that of the Statistical Machine Translation system.

[4] https://github.com/artetxem/undreamt

Table-VI: Sample Khasi to English Translation for a supervised NMT

Khasi (Source)	Reference (English)	Machine Translated (English)
nangta ki im tymmen shong prah	and live in Peace their Life for long	they also live a long long life
suk kita ki briew kiba U Blei u long u Trai	happy are the people whose God is the LORD	happy are the people whose God is the LORD
haei ka painkhana	where is the toilet	where is the toilet
kum ki ñiut kiba saphriang kylleng ka kper	like weeds that spread all through the garden	like <OOV> who spread up over
nga la thoh shithi sha phi	I wrote a letter for you	I wrote you letter to you
ka jakarieh jong nga dei ha U Blei	my refuge is God	my <OOV> is to God
balei ba ki jaitbynriew kin kylli ïa ngi	why should the nations ask us	why nations would ask us

Table- VII: Sample Khasi to English Translation train in OpenNMT

Khasi (Source)	Reference (English)	Machine Translated (English)
u phlang ban bsa ïa u u mih	grass to feed him grows	the grass to feed it rose
ki shakri jong u ki jubab Ym shym la leh ei ei na ka bynta jong u	his servants answered Nothing has been done for him	his servants answered No matter for him
u iam pangnud wei briew	broken hearted Weeps alone	he called a neighbour
wat ai ba u lurstep un tyngshaiñ	keep the morning star from shining	the dont let the Monster shine
kumba ka long don shibun ki dkhot hynrei tang kawei ka met	as it is there are many parts but one body	its it is there many parts but one body
hynrei pynban kin sngewshyrkhei shikatdei eh	but then they will be terrified	but then they will be terrified
hapdeng ngi ki briew ngi don ka jingsngewthuh ngi don ka jingmut ngi ka jingkyrmen ia kaba shadien	amongst us human beings we have understanding we have our plans and hopes and dreams for the future	we people say we have understanding we have meaning our hope
ka jingrwai ainguh	a Hymn of Thanksgiving	a Song of Thanksgiving

Table- VIII: Sample Khasi to English Translation train in Moses

Khasi (Source)	Reference (English)	Machine Translated (English)
ki shakri jong u ki jubab Ym shym la leh ei ei na ka bynta jong u	his servants answered Nothing has been done for him	his servants they answered We have not done anything for him
u iam pangnud wei briew	broken hearted Weeps alone	one people with him
bad ïaroh ïa ka kyrteng bakhuid jong me	and praise your holy name	and praise your holy name
wat ai ba u lurstep un tyngshaiñ	keep the morning star from shining	dont let him lurstep will shine
kumba ka long don shibun ki dkhot hynrei tang kawei ka met	as it is there are many parts but one body	as it is has many parts but only one body
hynrei pynban kin sngewshyrkhei shikatdei eh	but then they will be terrified	but then they will be terrified
shaei u director	wheres the director	where is director
ka jingrwai ainguh	a Hymn of Thanksgiving	a Song of Thanksgiving
ka buit aiu khun kum kata	what trickery is this my son	Whatre trick son such

Table-VI, Table-VII and Table-VIII shows the sample output for the Khasi to English machine translation. Reference translation are the manually target sentence constructed with human intervention who have language knowledge whereas machine translation is the outcome when translation are performed by a machine. OOV in the tables stand for Out of Vocabulary Words in which the models could not find a corresponding translation for a particular source word.

VIII. CONCLUSIONS AND FUTURE WORK

This paper focused on NMT system implementation perspective with the intension of improving translation accuracy for low resource languages. Unsupervised neural machine translation [12] was highly inspired in this paper.

Moreover, the technique allows to train both unsupervised and supervised learning. The former made use of only monolingual corpus while the later use parallel corpus.

The same parallel corpus was also train in OpenNMT and SMT (Moses toolkit) to have a better performance evaluation.

It was observed that supervised NMT accuracy was very closed to that of the SMT system specially in the case of model trained in OpenNMT. The accuracy of unsupervised translation model is low as compared to supervised model. The main reason of this poor result is due to less corpus size. Secondly, since the monolingual corpus was directly taken and constructed from the internet with only minimal preprocessing hence the accuracy of the translation was affected. On the contrary result for supervised learning was comparatively better which is based on the parallel corpus for training the model. To increase the translation accuracy more corpus should be accumulated into the current work. Since Khasi is a low resource language, more features should be added in the existing module of the preprocessing steps and hence more exploration on the subject is recommended so that word embedding efficiency will be increase and ultimately contribute to better accuracy of the translation system.

ACKNOWLEDGMENT

I express my gratitude to the staff of Bible Society translation, Shillong for guidance and direction in connection with my work which have been of immense help to me.

Further I also extend my gratitude to the various Press House viz namely Ri Khasi Press for their unflinching support in providing data and information of the translated books and articles without any reluctance on their part.

My acknowledgement also goes to the North Eastern Hill University Central library for free access to all translated materials and data which had enriched this work.

REFERENCES

1. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. (2016). *Neural machine translation by jointly learning to align and translate.* arXiv preprint arXiv:1409.0473. https://arxiv.org/pdf/1409.0473.pdf.
2. Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations.* Association for Computational Linguistics, pages 67–72. https://doi.org/10.18653/v1/P17-4012.
3. Gruessner and Karl-Heinz. (2004). Khasi: a minority language of North–East India, From an unwritten to a written language. In *18th European Conference on Modern South Asian Studies*.
4. Ganesh Narendra Devy, and Esther Syiem(eds). (2014). People's Linguistic Survey of India: Volume Nineteen, Part II. *The languages of Meghalaya*. Orient Blackswan Private Limited, New Delhi.
5. G. R. Mawblei. (2017). *The Khasis : Culture and Beliefs as Confronted by the Gospel. Bilingual U Khasi: Ka Dustur bad ka Jingngeit Ba Lyngkhuh ia Ka Gospel*. Syiem Offset Printers, Shillong.
6. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. (2014). *Sequence to sequence learning with neural networks. In Advances in neural information processing systems*, page 3104-3112. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.
7. Jeebon Roy. (2005). *Shaphang U wei U Blei : About One God. A Translation by Bijoy Sawian*. Ri Khasi Press, Shillong.
8. Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. (2014). *On the properties of neural machine translation: Encoder-decoder approaches*. arXiv preprint arXiv:1409.1259. https://arxiv.org/pdf/1409.1259.pdf.
9. K. S. Nagaraja, Paul Sidwell, and Simon. (2013). *A lexicostatistical study of the Khasian languages: Khasi, Pnar, Lyngngam, and War*. Mon-Khmer Studies (Volume 42). pages 1-11. https://pure.mpg.de/rest/items/item_2316456/component/file_2316455/content.
10. 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 (ACL)*, pages 311-318. https://www.aclweb.org/anthology/P02-1040.
11. Minh-Thang Luong, Hieu Pham, and Christopher D Manning. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025. https://arxiv.org/pdf/1508.04025.pdf.
12. Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. (2018). Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*. https://arxiv.org/pdf/1710.11041.pdf.
13. Mikel Artetxe, Gorka Labaka, and Eneko Agirre. (2017). Learning bilingual word embeddings with (almost) no bilingual data. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 451–462. https://www.aclweb.org/anthology/P17-1042.
14. Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. (2017). Adversarial training for unsupervised bilingual lexicon induction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 1959–1970. https://aclweb.org/anthology/P17-1179.
15. Madaline Tham. (2015). *The Golden Duitara: Translation from Soso Tham's Duitara Ksiar Khasi to English*. Print Xpress, Shillong.
16. Nal Kalchbrenner, and Phil Blunsom. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, page 1700-1709. https://www.aclweb.org/anthology/D13-1176.
17. Nissor Singh, and Gurdon, R. R. T. (Eds). (1904). *Khasi English Dictionary*. Mittal Publication, New Delhi.
18. Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. (2017). In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics, pages 28-39. https://www.aclweb.org/anthology/W17-3204.
19. Rico Sennrich, Barry Haddow, and Alexandra Birch. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 86–96. https://www.aclweb.org/anthology/P16-1009.
20. Sebastian Ruder, Ivan Vulić, Anders Søgaard.(1993). A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, Volume 1, pages 1-15. https://arxiv.org/pdf/1706.04902.pdf.
21. Seng Khasi Seng Kmie. (nd). Centenary Souvenir: *Ka Shad Suk Mynsiem (1911-2011)*. Bhabani Print & Publication, Guwahati.
22. Shlur Manik Syiem. (2006). *The Olden Days of the Seven Hut. Translated from Soso Tham's Ki Sngi Ba Rim U Hynniewtrep*. Walvens Computer System. Synod Complex, Shillong.
23. Soso Tham. (1972). *Ka Duitara Ksiar Ne Ki Poetry Khasi*. Dispur Print House, Guwahati.
24. Soso Tham. (1936). *Ki Sngi Barim U Hynniew Trep*. Ri Khsi Press, Shillong.
25. Radhon Singh Berry Kharwaniang. (1978). *Ka Jingsneng Tymmen Khasi*. Book Stall, Shillong.
26. Radhon Singh Berry Kharwaniang. (2016). *The Teaching of Khasi Elders: Ka Jingsneng Tynmen Part I & II Translated by Bijoya Sawian*. Vivekananda Kendra Institute of Culture, Guwahati.
27. 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*. arXiv preprint arXiv:1609.08144. https://arxiv.org/pdf/1609.08144.pdf%20(7.pdf.
28. Déchelotte, D., Schwenk, H., Bonneau-Maynard, H., Allauzen, A., & Adda, G. (2007). *A state-of-the-art statistical machine translation system based on moses*. In MT Summit (pp. 127-133). https://pdfs.semanticscholar.org/1e29/1349bdbe06941f3d9f2e2cccbd00d36ef980.pdf
29. Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. (2017), OpenNMT: Neural Machine Translation Toolkit. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, pages 67–72. Association for Computational Linguistics https://doi.org/10.18653/v1/P17-4012.

AUTHORS PROFILE

N Donald Jefferson Thabah, PhD, Research Scholar, Department of Computer Science, Assam University, Silchar, Assam, India.

Prof. Bipul Syam Purkayastha, Department of Computer Science, Assam University, Silchar, Assam, India.

Specialization: Computational Linear Algebra, Computer Programming, Natural Language Processing, Internet Programming, Website Designing.