A Baseline System for Khasi and Assamese Bidirectional NMT with Zero available Parallel Data : Dataset Creation and System Development

Kishore Kashyap, Kuwali Talukdar, Mazida Akhtara Ahmed, Parvez Aziz Borauh
Department of Information Technology
Gauhati University
Guwahati, Assam, India
{kb.guwahati, kuwalitalukdar,14mazida.ahmed, parvezaziz70}@gmail.com

Abstract

In this work we have tried to build a baseline Neural Machine Translation system for Khasi and Assamese in both directions. Both the languages are considered as low-resourced Indic languages. As per the language family in concerned, Assamese is a language from Indo-Aryan family and Khasi belongs to the Mon-Khmer branch of the Austroasiatic language family. No prior work is done which investigate the performance of Neural Machine Translation for these two diverse low-resourced languages. It is also worth mentioning that no parallel corpus and test data is available for these two languages. The main contribution of this work is the creation of Khasi-Assamese parallel corpus and test set. Apart from this, we also created baseline systems in both directions for the said language pair. We got best bilingual evaluation understudy (BLEU) score of 2.78 for Khasi to Assamese translation direction and 5.51 for Assamese to Khasi translation direction. We then applied phrase table injection (phrase augmentation) technique and got new higher BLEU score of 5.01 and 7.28 for Khasi to Assamese and Assamese to Khasi translation direction respectively.

1 Introduction

While Neural Machine Translation (NMT) has achieved State-Of-The-Art in Machine Translation domain, the inherent nature of huge data requirement of neural systems makes it extremely difficult to get higher score in tasks where the data is scarce. To achieve high BLEU (Papineni et al., 2002) score in the translation task, the languages under study must have a good amount of parallel data. For languages where such parallel data is very low or not available at all, it becomes challenging to create a decent NMT system for these languages. In this work, we have engaged our research in studying how one can proceed to develop an NMT system for languages where no parallel data is available and the languages are completely unrelated. We will briefly describe about the languages in Section 2. In Section 3 we explored existing works. Section 4 describe the process of data collection and generation. Section 5, 6 and 7 shows the experiments, results and finally conclusions respectively.

2 About the languages

Assamese is an official language in India among 22 languages as declared in the Eighth schedule of Indian Constitution[1]. Assamese belongs to the Indo Aryan language family and is mainly used in the North Eastern part of India with more than 16 million speakers (Ahmed et al., 2023) and is considered as a lingua franca in that region. But the language has been recognised as a low-resourced language for machine translation task (Ahmed et al., 2023). Assamese is written using the Bengali-Assamese Script[2] or sometimes also called Assamese-Bengali script (Mahanta, 2012).

Another language, 'Khasi', is mainly spoken in the North Eastern state of India, Meghalaya. Khasi is not yet included as a scheduled language in India. It is a resource poor language. It belongs to the Mon-Khmer branch of Austroasiatic language (Hujon et al., 2023b), which is a language family found mainly in Southeast Asia and the eastern part of the Indian subcontinent. The other languages in the Khasic group of the Shillong Plateau, such as Pnar, Lyngngam, and War, are also closely related to Khasi[3]. Khasi is written using the Latin script.

3 Review of existing work

While studying the existing work done with respect to machine translation system development involv-

[1] https://rajbhasha.gov.in/en/languages-included-eighth-schedule-indian-constitution
[2] https://en.wikipedia.org/wiki/Bengali-Assamese_script
[3] https://en.wikipedia.org/wiki/Khasi_language

ing Assamese-xx or xx-Assamese languages we got higher number of works as compared to Khasi-xx or xx-Khasi languages.

The works in MT involving Assamese language were developed using both statistical and neural approaches, where the later became the contemporary technology to further the research in translation task. In Laskar et al. (2020), the authors have reported about developing Statistical Machine Translation (SMT) and NMT baseline works for Assamese and English bilingual translation system in both directions. They have used two-layer network consists of long short term memory (LSTM) cell with 500 nodes in each layer and attention mechanism (Bahdanau et al., 2015). Before this baseline work on Assamese NMT, there are few works (Baruah et al., 2014), (Das and Baruah, 2014), (Singh et al., 2014), (Barman et al., 2014), (Kalita and Islam, 2015), (Hannan et al., 2019) on MT development for Assamese using SMT. Das et al. (2014) developed a rule based machine translation system for Assamese and English using Apertium (Forcada et al., 2011). In the neural domain, there are works on Assamese NMT development such as (Baruah et al., 2021), (Dutt et al., 2022), (Nath et al., 2022), (Laskar et al., 2022a), (Laskar et al., 2021), (Laskar et al., 2022b), (Laskar et al., 2023).

For MT systems involving Khasi languages, we got one work (Singh and Vellintihun Hujon, 2020) using SMT (the authors also developed an NMT system in the same work). Hujon et al. (2023a) used convolutional sequence to sequence learning for English-Khasi Neural Machine Translation. Use of both supervised and unsupervised technique is reported in Thabah et al. (2019) for Khasi to English NMT. In Nonghuloo et al. (2020), the authors have used Attention mechanism to get good BLEU score for English to Khasi NMT system.

From this study of existing work, we conclude that almost all work on Khasi MT system development, emphasis is given to translating Khasi to/from English and no attempt is made to couple the language with other non-English languages. To fill this gap, we attempt to develop a baseline system involving Khasi and non-English language, Assamese, in our case.

4 Data collection and generation

The languages, for which we are attempting to create NMT system, do not have any parallel data. What we had at our disposal are two resources: 1) English-Khasi Parallel data with 24000 sentences from WMT23 (IndicMT subtask) and 2) English to Assamese NMT system (developed in house) with 19.48 BLEU score for our own test set and 15.16 BLEU score for FLORES-200 test set (Team et al., 2022).

We analysed the English-Khasi data to check the sentence length for source and target language. The Figure 1 shows the density histogram of the sentence length variation for English and Khasi. We got extremely unbalanced data in the parallel corpus (maximum English sentence length of 103 and maximum Khasi length of 1723). So, we filtered the data in two ways. Firstly, we have selected sentences from the original corpus where sentence lengths were less than or equal to 200 (Set 1 Figure 2a) and secondly, we made the threshold criteria to be sentence length less than or equal to 50 (Set 2 Figure 2b). A few sentences that are filtered from Set 1 are filled from the first available sentences of the validation set (provided with the English-Khasi IndicMT subtask of WMT23) to keep the total number at 24000. The English side of both the two sets are then translated to Assamese using our in-house English-Assamese NMT system (as mentioned above).

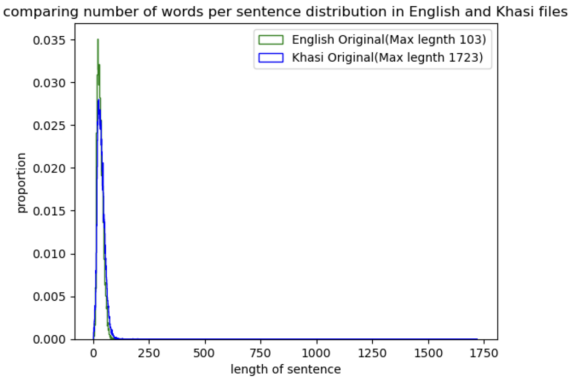


Figure 1: Original English-Khasi sentence length variation

Now, we have two sets of synthetic parallel corpus for Assamese and Khasi. Set 1 has 24000 sentences (Figure 2c) and Set 2 has 18508 sentences (Figure 2d). We split the data into training, testing and validation sets. The details of train-test-valid split is shown in Table 1

	#train	#test	#validation	#total
Set 1	22000	1000	1000	24000
Set 2	17008	1000	500	18508

Table 1: Details of train, test and valid set

Again, to generate the test data, we have translated the English side of the test set provided by English-Khasi IndicMT sub task of WMT23, to Assamese with our NMT system, and then manually post edited and validated the same. The test set contains 1000 sentences.

5 Experimental setup

For the purpose of evaluating the influence of sentence length, we have performed four experiments in Assamese to Khasi and Khasi to Assamese language directions (two in each directions). Moreover, we have also performed Phrase Table Injection experiment on all the sets, again having four different models. The details of the experiments along with the machine details are shown in the following subsections.

5.1 Machine details

For all the experiments we have used a single GPU machine. The details of the hardware available are as follows:

NVIDIA Quadro P1000 GPU with 4096MB of GPU memory and 640 CUDA Cores, Graphics Clock speed (min: 136MHz, max:1544MHz), Memory Transfer Rate (min:810MHz, max:5010MHz) is the only system used for the purpose of this work. The machine has a RAM of 16GB and 64 bit Intel Xeon CPU.

5.2 System development

We adopted same training and testing setup to build the NMT models from Set 1 and Set 2.

For both the sets, tokenization of Khasi data is done with moses tokenizer (Koehn et al., 2007) and tokenization of Assamese data is done with tokenizer from IndicNLP Library (Kunchukuttan, 2020). We have used IndicNLP tokenizer for Assamese as it can preserve some language specific aspects of the Assamese language as shown in (Ahmed et al., 2023).

Byte Pair Encoding (BPE, Sennrich et al. (2016)) is used for the purpose of subword tokenization of Assamese and Khasi data separately. BPE tackles the OOV problem for rare words. The number of BPE symbols is set to 8k for both language.

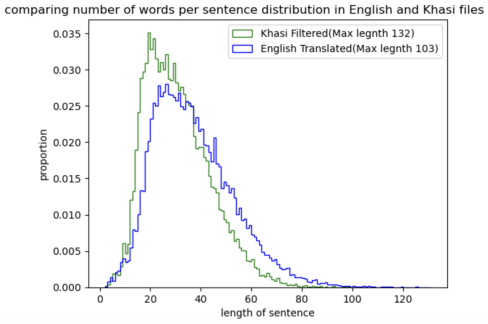
We also trained a PBSMT system (Koehn et al., 2007) on our dataset and extracted phrases as explained in Batheja and Bhattacharyya (2022). We then augmented our extracted phrases to the training corpus of Set 1 ans Set 2.

All the models are trained with **Fairseq**[4] (Ott et al., 2019). It is a sequence-to-sequence learning toolkit which facilitates experimenting with NMT development. Fairseq has various implementations of standard NMT architectures. We employ transformer architecture as described in Vaswani et al. (2017). All the models are trained with a batch size of 512 tokens, an initial learning rate of 0.0005 with inverse square root scheduler. Along with this, we apply a dropout 0.3 (as the data is very low), label smoothing 0.1 and weight-decay 0.0001. Adam optimizer with adam betas (0.9, 0.98) and warm-up updates of 4000 is used. We trained the models for maximum epochs of 150. This is a fairly standard setup for NMT training. As the main aim of this study is to develop a translation system for two low-resource Indic languages which, we hope, will serve as a baseline work to further the development, we have not experimented much with model architecture and hyper parameter tuning and keep it as a pointer to future studies. In the next section we describe the result of the experiments.

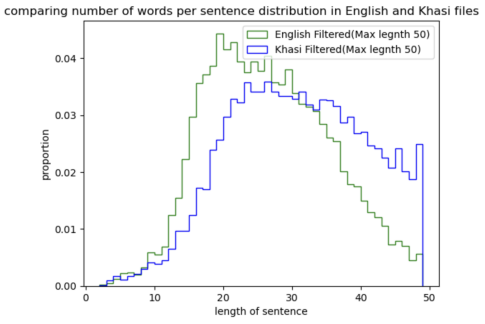
6 Result

We use BLEU, ChrF2++(Popović, 2017), TER(Snover et al., 2006) metrics to evaluate the translation score of **Assamese→Khasi** and **Khasi→Assamese** directions. We use SacreBLEU[5] (Post, 2018) library for the purpose. The results are shown in the Table 2. In both the translation directions, Set 2, which has less number of parallel sentences, has shown best score in all the evaluation metrics, viz., BLUE, chrF2++ and TER. We conjecture that the sentence length affected the performance of Set 1 as explained in (Koehn and Knowles, 2017) and (Cho et al., 2014). It is also interesting to note that, **Assamese→Khasi** direction got better translation score in both the sets as compared to **Khasi→Assamese** direction.

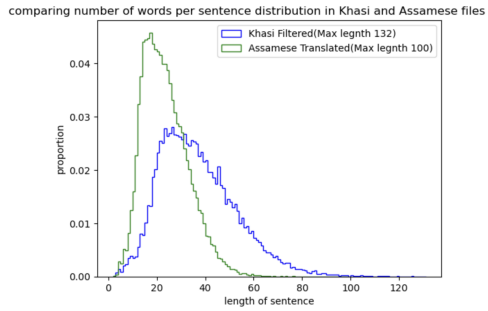
[4] https://github.com/facebookresearch/fairseq
[5] https://github.com/mjpost/sacrebleu



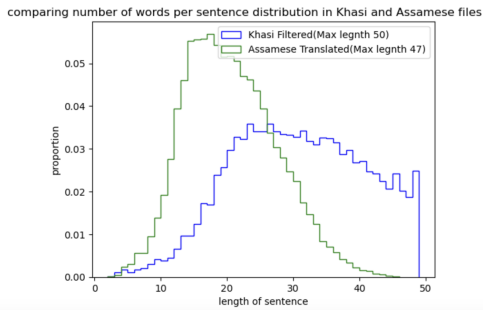
(a) Filtered <= 200



(b) Filtered <= 50



(c) Khasi-Assamese (Synthetic) <=200



(d) Khasi-Assamese (Synthetic) <=50

Figure 2: Sentence Length variation

		BLEU↑	chrF2++↑	TER↓
As-Kh	Set 1	3.80	24.6168	87.9395
	Set 2	**5.51**	**26.5478**	**87.8170**
Kh-As	Set 1	2.66	20.0046	102.2310
	Set 2	**2.78**	**20.6976**	**100.9910**

Table 2: Evaluation score for Set 1 and Set 2

After augmenting the data with phrases obtained from PBSMT model, we got the test result as shown in Table 3

		BLEU↑	chrF2++↑	TER↓
As-Kh	Set 1	5.12	27.5043	84.4632
	Set 2	**7.23**	**28.3312**	**83.4672**
Kh-As	Set 1	4.11	22.1122	99.1010
	Set 2	**5.01**	**22.9867**	**99.0020**

Table 3: Evaluation score for Set 1 and Set 2 with *phrase augmentation*

The same consistency in terms of better score is observed for all three evaluation metrics holds in the case of phrase augmentation too. **Assamese→Khasi** direction is performing well as compared to **Khasi→Assamese** direction. From both Table 2 and Table 3, it can be noted that we achieved +1.72 increase in BLEU in **Assamese→Khasi** direction and +2.23 increase in BLEU in **Khasi→Assamese** direction using phrase injection. We got best scores in all the experiments for Set 2.

7 Conclusion

Firstly, we conclude that, with only synthetic data, it is possible to develop a baseline system for language pairs where no parallel corpus is available previously. Secondly, it is observed that sentence length plays a crucial role in getting good translation score. To establish the fact we would like to direct our attention in the future to formulate optimal sentence length for other language pairs. And lastly, when the size of training corpus is very small, extracting phrase pair from an SMT model trained on the same data set and augmenting the phrases to the training set can further enhance the translation quality in terms of automatic evaluation metrics such as BLEU, chrF and TER.

In this work, we have created a synthetic Khasi-Assamese corpus and test set for the same pair. Along with that we also created a baseline NMT system for Khasi-Assamese language pair. This is the very first work which demonstrates the use of NMT for Khasi and non-English language. We encourage other researcher to pair diverse languages to create NMT systems.

Limitations

We observe that the system fail to translate long sentences correctly. Moreover, extraction of good quality phrases using language agnostic models is a computing intensive work. Again, generation of good quality synthetic parallel corpus is a huge challenge.

Acknowledgements

We thank the Ministry of Electronics and Information Technology (MeitY), Government of India for their assistance through the Project ISHAAN.

References

Mazida Akhtara Ahmed, Kishore Kashyap, and Shikhar Kumar Sarma. 2023. Pre-processing and resource modelling for english-assamese nmt system. In *2023 4th International Conference on Computing and Communication Systems (I3CS)*, pages 1–6.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Anup Barman, Jumi Sarmah, and Shikhar Kr Sarma. 2014. Assamese wordnet based quality enhancement of bilingual machine translation system. In *Proceedings of the Seventh Global Wordnet Conference*, pages 256–261.

Kalyanee Kanchan Baruah, Pranjal Das, Abdul Hannan, and Shikhar Kumar Sarma. 2014. Assamese-english bilingual machine translation. *CoRR*, abs/1407.2019.

Rupjyoti Baruah, Rajesh Kumar Mundotiya, and Anil Kumar Singh. 2021. Low resource neural machine translation: Assamese to/from other indo-aryan (indic) languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–32.

Akshay Batheja and Pushpak Bhattacharyya. 2022. Improving machine translation with phrase pair injection and corpus filtering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5395–5400, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches.

P Das, KK Baruah, A Hannan, and SK Sarma. 2014. Rule based machine translation for assamese-english using apertium. *International Journal of Emerging Technologies in Computational and Applied Sciences*, 8(5):401–406.

Pranjal Das and Kalyanee K Baruah. 2014. Assamese to english statistical machine translation integrated with a transliteration module. *International Journal of Computer Applications*, 100(5).

Rudra Dutt, Tarun Aditya Kusupati, Akshat Srivastava, and Basab Nath. 2022. Neural machine translation for english-assamese language pair using transformer. In *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, pages 1–5. IEEE.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.

Abdul Hannan, Shikhar Kr Sarma, and Zakir Hussain. 2019. Marie: a statistical approach to build a machine translation system for english assamese language pair. *International Journal of Computer Sciences and Engineering, Available: https://doi.org/10.26438/ijcse/v7i3*, 774779.

Aiusha Vellintihun Hujon, Khwairakpam Amitab, and Thoudam Doren Singh. 2023a. Convolutional sequence to sequence learning for english-khasi neural machine translation. In *2023 4th International Conference on Computing and Communication Systems (I3CS)*, pages 1–4.

Aiusha Vellintihun Hujon, Thoudam Doren Singh, and Khwairakpam Amitab. 2023b. Neural machine translation systems for english to khasi: A case study of an austroasiatic language. *Expert Systems with Applications*, 238:121813.

Nayan Jyoti Kalita and Baharul Islam. 2015. Bengali to assamese statistical machine translation using moses (corpus based). *arXiv preprint arXiv:1504.01182*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. EnAsCorp1.0: English-Assamese corpus. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 62–68, Suzhou, China. Association for Computational Linguistics.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2022a. Improved neural machine translation for low-resource english–assamese pair. *Journal of Intelligent & Fuzzy Systems*, 42(5):4727–4738.

Sahinur Rahman Laskar, Bishwaraj Paul, Pankaj Dadure, Riyanka Manna, Partha Pakray, and Sivaji Bandyopadhyay. 2023. English–assamese neural machine translation using prior alignment and pre-trained language model. *Computer Speech & Language*, 82:101524.

Sahinur Rahman Laskar, Bishwaraj Paul, Partha Pakray, and Sivaji Bandyopadhyay. 2022b. Improving english-assamese neural machine translation using transliteration-based approach. In *International Conference on Frontiers of Intelligent Computing: Theory and Applications*, pages 223–231. Springer.

Sahinur Rahman Laskar, Bishwaraj Paul, Siddharth Paudwal, Pranjit Gautam, Nirmita Biswas, and Partha Pakray. 2021. Multimodal neural machine translation for english–assamese pair. In *2021 International Conference on Computational Performance Evaluation (ComPE)*, pages 387–392. IEEE.

Shakuntala Mahanta. 2012. Assamese. *Journal of the International Phonetic Association*, 42(2):217–224.

Basab Nath, Sunita Sarkar, and Surajeet Das. 2022. Development of neural machine translator for english-assamese language pair. In *Advanced Techniques for IoT Applications: Proceedings of EAIT 2020*, pages 279–288. Springer.

Mark S Nonghuloo, Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India., Nagaraja Rao A., and Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India. 2020. Analyses and modeling of neural machine translation for English-to-Khasi. *International Journal of Recent Technology and Engineering (IJRTE)*, 9(2):115–118.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. 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. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Moirangthem Tiken Singh, Rajdeep Borgohain, and Sourav Gohain. 2014. An english-assamese machine translation system. *International Journal of Computer Applications*, 93(4).

Thoudam Doren Singh and Aiusha Vellintihun Hujon. 2020. Low resource and domain specific english to khasi smt and nmt systems. In *2020 International Conference on Computational Performance Evaluation (ComPE)*, pages 733–737.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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

Bipul Syam Purkayastha, and Department of Computer Science, Assam University, Silchar, India. 2019. Khasi to english neural machine translation: An implementation perspective. *Int. J. Eng. Adv. Technol.*, 9(2):4330–4336.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.