EnKhCorp1.0: An English–Khasi Corpus

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji
Darsh Kaushik, Partha Pakray, Sivaji Bandyopadhyay
Department of Computer Science and Engineering
National Institute of Technology Silchar
Assam, India
{sahinur_rs, abdullah_ug, darsh_ug, partha}@cse.nits.ac.in,
sivaji.cse.ju@gmail.com

Abstract

In machine translation, corpus preparation is one of the crucial tasks, particularly for low-resource pairs. In multilingual countries like India, machine translation plays a vital role in communication among people with various linguistic backgrounds. There are available online automatic translation systems by Google and Microsoft which include various languages which lack support for the Khasi language, which can hence be considered low-resource. This paper overviews the development of EnKhCorp1.0, a corpus for English–Khasi pair, and implemented baseline systems for English-to-Khasi and Khasi-to-English translation based on the neural machine translation approach.

1 Introduction

The Khasi language (also spelled Khasia, Khassee, Cossyah, or Kyi) is primarily spoken by people living in the region surrounding the Khasi and Jaintia Hills of Meghalaya state in India. It is a member of the Mon-Khmer linguistic branch of the Austroasiatic language family. Khasi is an associate official language[1] in Meghalaya since 2005. According to the 2011 census of India, there are around one million native speakers of Khasi[2]. Khasi has significant dialectal variation, some of them being Sohra Khasi, Mawlai Khasi, Pnar, Nongkrem Khasi, Mylliem Khasi, Bhoi Khasi Nonglung, War and Maram. Khasi has a subject-verb-object (SVO) sentence structure, similar to English but unlike most of the Indian languages Roberts (2005). Khasi contains several words borrowed from Indo-Aryan languages, mainly from Bengali and Hindi. In the past, the Khasi language had no script of its own. The Welsh missionary Thomas Jones[3], in 1841, wrote the language in the Latin script. As a result, the Latin alphabet of the language has a few similarities with the Welsh alphabet. Khasi in Latin script has a 23-letter alphabet.

1.1 Language Preservation

Every language is a unique perspective to comprehend the world by sharing its history, philosophy, and culture. Extinction of a language results in loss of historical, ecological, and cultural information and opinions. It may even affect its speakers' livelihood and existential mentality as they adopt the dominant languages to attain socio-economic benefit, tackle the scarcity of modern documentation and usage of the language, or mitigate the fear of discrimination. Therefore, the need to preserve such low-resource languages, including the Khasi language, is of significance. Machine translation (MT) helps in language preservation Bird and Chiang (2012). We have attempted to create a corpus and introduce it in an MT environment that helps document the Khasi language and preserve such minority languages by encouraging language usage and eliminating the linguistic barrier in communication.

1.2 Low-Resource Translation

In the domain of Natural Language Processing (NLP), MT deals with translation from one natural language to another. Further, Neural Machine Translation (NMT) is a state-of-the-art approach of incorporating artificial neural networks in MT systems. The data or resources for training the translation systems may include corpora from various online sources, native speakers, and

[1] https://www.indiacode.nic.in/bitstream/123456789/5467/1/the_meghalaya_state_language_act,_2005_(act_no._10_of_2005).pdf

[2] https://www.censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf

[3] http://lisindia.ciil.org/Khasi/Khasi_script.html

computational resources. The Natural languages are categorized into three broad categories: high, medium, and low-resource. A language falls under the low-resource category when it has limited online resources Megerdoomian and Parvaz (2008); Probst et al. (2001). This categorization can also be made on the quantity of data required to train an NMT model Gu et al. (2018). According to Kocmi (2020), a language is considered a low-resource language if the number of training instances present in the corpus is below one million. Along with the corpus's size, the diversity of both language and structure is of importance too. Structurally, it must consist of all types of sentences, including short, medium, and long sentences. Different dialects of the same language might give rise to some inconsistency in translations. Therefore, the designation of a language as "low-resource" is not precise and requires consideration of many factors. Most world languages are categorized as low-resource on account of resource availability. In India, the limited MT works are performed on the northeastern region's low-resource languages, including Mizo Lalrempuii et al. (2021), Assamese Laskar et al. (2021b), Manipuri Singh and Singh (2020), and Khasi Thabah and Purkayastha (2021). We can consider the English–Khasi pair as a low-resource pair based on limited resources.

In this work, we have developed an English–Khasi corpus: EnKhCorp1.0 and built baseline systems based on NMT. There is no standard corpus available for the low-resource English–Khasi pair to the best of our knowledge. It is the hope of the authors that this resource fills that gap and leads to the development of more and better resources for the Khasi community.

The structure of the rest of the paper is as follows: Section 2 presents the overview of corpus preparation. Section 3 presents construction and evaluation of the baseline English–Khasi NMT system and conclude the paper in Section 4.

2 EnKhCorp1.0

This section overviews the corpus. Section 2.1 describes the contents of the corpus. Section 2.2 and 2.3 present the data extraction technique and domain coverage.

2.1 Details of Corpus

The available resource options for English–Khasi (En-Kha) parallel and monolingual data are limited. Therefore, we have explored different possible sources to prepare parallel and monolingual corpora. Table 1 presents some examples of sentences that were collected. The sources of data are reported below:

2.1.1 Parallel Corpus

- **Bible:** The Bible[4] is publicly available on the online in multiple languages, including Khasi. We have collected 26,086 parallel sentences from the Bible source using crawling technique.

- **Online Dictionary:** There are online dictionaries, namely, Glosbe[5], available in the multilingual form in which English–Khasi bilingual words and parallel examples are present. From Glosbe, we have collected 2,225 parallel sentences using crawling technique.

- **Learn-Khasi Website:** The website, namely, Learn Khasi online[6], is developed to teach essential words and daily usage sentences in the Khasi language. We have manually collected 120 parallel sentences from this website.

2.1.2 Monolingual Corpus

A standard monolingual corpus of English is available online: WMT16[7]. Therefore, we have focussed on the preparation of only Khasi monolingual corpus. We have collected 157,968 Khasi sentences from different web pages/blogs and added 25,836 Khasi sentences from the parallel data (train set) to increase the size of the monolingual corpus, totalling 183,804.

2.1.3 Corpus Analysis and Statistics

The collected raw data (both parallel and monolingual) are cleaned by removing unwanted symbols, URLs, many special characters (######, ____, ……., $$$$), blank lines, etc. To keep the contextual meaning of the sentences, we did not remove punctuation marks. If we remove the punctuation marks, then it will alter the meaning of the sentence. Table 2 presents an example sentence having punctuation marks. In addition, long sentences

[4] https://www.bible.com/en-GB/bible/1865/GEN.1.KHASICLBSI

[5] https://glosbe.com/en/kha

[6] https://www.languageshome.com/English-Khasi.htm

[7] http://www.statmt.org/wmt16/translation-task.html

(more than 50 words) are split based on the punctuation marks (?.) at the end. It means a maximum sentence length of 50 words, we have considered. If the sentence length is greater than 50 words, then split it based on punctuation marks. Table 3 presents the overall data statistics. We have removed duplicate sentences from both parallel and monolingual corpora. After removing duplicates, the total parallel data reduced to 28,036 sentences. In order to do the experiment for the MT system, the parallel data needs to split into train, validation, and test set data. In Kunchukuttan et al. (2018), the English–Hindi parallel corpus was developed, and for the baseline system, they split the data by considering 99.79% data for the train set and <1% data for validation and test set. They considered most data for the training set and very few data for validation and test set. In our baseline system experiment, we have considered most data (92.15%) for the training set, 7.13% data for the validation and 0.71% for the test set. Table 4 presents the statistics of train, validation, and test set data. We have considered only 200 test data, since it is used for the baseline system. We will consider more test sentences in the future work. During the training process of a model, the train set is used for learning the parameters, and a validation set is required to verify the performance of a model to generate the optimum model. The unseen or test data is required to check the generated model.

Type	Sentences	Tokens	
		En	Kha
Train	25,836	664,385	830,393
Validation	2,000	42,725	67,474
Test	200	5,105	6,241

Table 4: Data statistics for training, validation and testing set

2.2 Data Extraction technique

We utilized Scrapy[8], an open-source framework for web crawling, to scrape the data from various online sources. The xpath of each element has undergone a certain degree of generalization coding to replicate multiple web pages. It helps to crawl numerous web pages and extract essential information. We first provide the URL of the web page. The Khasi raw text in the HTML files was extracted directly. The obtained Khasi monolingual data is kept as it is. However, the parallel data is aligned by separating them into source and target files. The process of alignment and verification took substantial human effort. Additionally, we collected parallel sentences manually from the : Learn-Khasi website.

2.3 Domain Coverage

The proposed corpus, EnKhCorp1.0 encompasses different domains, including religious material (the Bible), literature, daily usage, and common sentences.

3 Baseline System

In the area of MT, the NMT achieves a state-of-the-art approach for both high and low-resource language pairs Bahdanau et al. (2015); Pathak et al. (2018); Pathak and Pakray (2018); Laskar et al. (2019); Laskar et al. (2019, 2020b, 2021a). Therefore, we have chosen NMT to build the baseline system to estimate benchmark translation accuracy for both En-to-Kha and Kha-to-En translation.A sequence-to-sequence (seq2seq) model based encoder-decoder architecture is adopted for this work following the NMT baseline system of Laskar et al. (2020a); Kunchukuttan et al. (2018).

3.1 Experimental Setup

The OpenNMT-py[9] toolkit is employed to build two seq2seq models, namely RNN and BRNN. We have used two-layer long short term memory (LSTM), having 500 units in each layer with attention (Bahdanau et al., 2015). The default learning rate of 0.001 with Adam optimizer and 0.3 drop-outs are used. Moreover, GloVe[10] Pennington et al. (2014) pre-trained word vectors are used by utilizing the monolingual data. For English monolingual data, we have used 3 million sentences collected from WMT16.

3.2 Results

To evaluate baseline systems, we used the automatic evaluation metrics, namely, bilingual evaluation understudy (BLEU) Papineni et al. (2002), rank-based intuitive bilingual evaluation score (RIBES) Isozaki et al. (2010), translation edit rate (TER) Snover et al. (2006), word error rate (WER) Morris et al. (2004), metric for evaluation of translation with explicit ordering (METEOR)

[8] https://scrapy.org/

[9] https://github.com/OpenNMT/OpenNMT-py

[10] https://github.com/stanfordnlp/GloVe

Corpus	English	Khasi	Source
Parallel	After Jesus died, God restored him to life as a spirit person.	Hadien ba u Jisu u la ïap, U Blei u la ai biang ha u ïa ka jingim ba kynja mynsiem.	Bible
	We'll go to any length to send our child to a good university.	Ngin leh katba lah ban phah ia i khun jongngi sha ka iuniversity ba bha.	Glosbe
	How are you?	Kumno phi long?	Learn-Khasi website
Monolingual	-	Hynrei u wei na ki ba mynsaw u la khlad noh na ka daw ka jingmynsaw jur ha shwa ban poi sha Civil Hospital Shillong.	Web pages/Blogs
	-	Ha ka janmiet sngi nyngkong, ka kyhun ki pulit ka la hiar sha katei ka thain bad kem ïa kiba suba donkti ha ane ka jingjia shoh paidbhur ïa ki samla.	Web pages/Blogs

Table 1: Example sentences from various sources

English	Khasi
Jesus Christ himself said: "Do not marvel at this, because the hour is coming in which all those in the memorial tombs will hear his voice and come out."	U Jisu Krist da lade hi u la ong: "Wat sngew kyndit ia kane, namar ka por ka la jia, ha kaba kito kiba don ha ki jing tep kin ioh sngew ia ka jingpyrta jong u bad kin ia mih noh."

Table 2: Example sentence having punctuation marks

Corpus	Source	Sentences	Tokens	
			En	Kha
Parallel	Bible	26,086	684,090	866,326
	Glosbe	2,225	28,172	37,184
	Learn-Khasi	120	396	472
	Total Number of Sentences	28,431	712,658	903,982
Monolingual	Web Pages/Blogs/Bible/Glosbe	183,804	-	20,575,074

Table 3: Overall data statistics

Lavie and Denkowski (2009) and F-measure. For BLEU score evaluation, we have considered average scores up to trigram Laskar et al. (2020a). Table 5, 6 and 7 present the results of baseline systems.

Translation	BLEU		TER (%)	
	RNN	BRNN	RNN	BRNN
En-to-Kha	14.87	14.88	83.90	83.42
Kha-to-En	11.28	12.77	86.78	86.43

Table 5: BLEU and TER scores of baseline systems (higher the value of BLEU indicates better accuracy and lower the value of TER denotes better accuracy)

3.3 Analysis

From Table 5, 6 and 7, it is noticed that En-to-Kha translation accuracy is higher than Kha-to-En. It is because parallel data contains more Kha tokens than En, and thus, the model encodes a larger amount of token information and the decoder can produce a better translation in the case of En-to-Kha. Also, it is observed that the BRNN model outperforms the RNN model in both directions of translation. The BRNN model achieves BLEU, RIBES, METEOR and F-measure scores:(14.88, 12.77), (0.499622, 0.457185), (0.183745, 0.170957) and (0.433188, 0.392752) for En-to-Kha and Kha-to-En translation respectively. Also, the BRNN model attains better TER and WER scores: (83.42%, 86.43%), (83.59%, 90.30%) for both directions of translation. In the case of TER and WER, lower values indicate higher accuracy.

4 Conclusion and Future Work

This paper presents EnKhCorp1.0, where we have developed a parallel corpus of English–Khasi parallel and Khasi monolingual data. It can be used in various NLP tasks, including MT. The dataset will be publicly available here: `https://github.com/cnlp-nits/EnKhCorp1.0` along with necessary licence agreement. By utilizing this corpus, we have built NMT baseline systems for translation to and from Khasi. We will increase the corpus size in the future and perform more experiments with advanced deep learning techniques to improve translation accuracy.

Acknowledgement

We would like to thank Center for Natural Language Processing (CNLP) and Department of Computer Science and Engineering at National Institute of Technology, Silchar, India for providing the requisite support and infrastructure to execute this work.

Translation	RIBES		WER (%)	
	RNN	BRNN	RNN	BRNN
En-to-Kha	0.426957	0.499622	83.96	83.59
Kha-to-En	0.414650	0.457185	90.34	90.30

Table 6: RIBES and WER scores of baseline systems (higher the value of RIBES indicates better accuracy and lower the value of WER denotes better accuracy)

Translation	METEOR		F-measure	
	RNN	BRNN	RNN	BRNN
En-to-Kha	0.183013	0.183745	0.432489	0.433188
Kha-to-En	0.154466	0.170957	0.366319	0.392752

Table 7: METEOR and F-measure scores of baseline systems (higher the value of METEOR and F-measure indicate better accuracy

References

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*.

Steven Bird and David Chiang. 2012. Machine translation for language preservation. In *Proceedings of COLING 2012: Posters*, pages 125–134, Mumbai, India. The COLING 2012 Organizing Committee.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952.

Tom Kocmi. 2020. Exploring benefits of transfer learning in neural machine translation.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Candy Lalrempuii, Badal Soni, and Partha Pakray. 2021. An improved english-to-mizo neural machine translation. *Transactions on Asian and Low-Resource Language Information Processing*, 20(4):1–21.

S. R. Laskar, A. Dutta, P. Pakray, and S. Bandyopadhyay. 2019. Neural machine translation: English to hindi. In *2019 IEEE Conference on Information and Communication Technology*, pages 1–6.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020a. Enascorp1. 0: English-assamese corpus. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 62–68.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020b. Hindi-Marathi cross lingual model. In *Proceedings of the Fifth Conference on Machine Translation*, pages 396–401, Online. Association for Computational Linguistics.

Sahinur Rahman Laskar, Partha Pakray, and Sivaji Bandyopadhyay. 2019. Neural machine translation: Hindi-Nepali. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 202–207, Florence, Italy. Association for Computational Linguistics.

Sahinur Rahman Laskar, Partha Pakray, and Sivaji Bandyopadhyay. 2021a. Neural machine translation: Assamese–bengali. In *Modeling, Simulation and Optimization*, pages 571–579, Singapore. Springer Singapore.

Sahinur Rahman Laskar, Partha Pakray, and Sivaji Bandyopadhyay. 2021b. Neural machine translation for low resource assamese–english. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India*, volume 170, page 35. Springer.

Alon Lavie and Michael J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2–3):105–115.

Karine Megerdoomian and Dan Parvaz. 2008. Low-density language bootstrapping: the case of tajiki Persian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*

(LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543, Doha, Qatar. ACL.

Katharina Probst, R. Brown, J. Carbonell, A. Lavie, Lori S. Levin, and Erik Peterson. 2001. Design and implementation of controlled elicitation for machine translation of low-density languages.

Hugh Roberts. 2005. *A grammar of the Khasi language*. Mittal publications.

Salam Michael Singh and Thoudam Doren Singh. 2020. Unsupervised neural machine translation for english and manipuri. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 69–78.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

N Donald Jefferson Thabah and Bipul Syam Purkayastha. 2021. Low resource neural machine translation from english to khasi: A transformer-based approach. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India*, volume 170, page 3. Springer.