Findings of the WMT 2023 Shared Task on Low-Resource Indic Language Translation

Santanu Pal
Wipro AI Lab45

Partha Pakray
NIT Silchar

Sahinur Rahman Laskar
UPES, Dehradun

Lenin Laitonjam
NIT Mizoram

Vanlalmuansangi Khenglawt
Mizoram University

Sunita Warjri
MIT-WPU, Pune

Pankaj Kundan Dadure
UPES, Dehradun

Sandeep Kumar Dash
NIT Mizoram

Abstract

This paper presents the results of the low-resource Indic language translation task organized alongside the Eighth Conference on Machine Translation (WMT) 2023. In this task, participants were asked to build machine translation systems for any of four language pairs, namely, English-Assamese, English-Mizo, English-Khasi, and English-Manipuri. For this task, the IndicNE-Corp1.0 dataset is released, which consists of parallel and monolingual corpora for northeastern Indic languages such as Assamese, Mizo, Khasi, and Manipuri. The evaluation will be carried out using automatic evaluation metrics (BLEU, TER, RIBES, COMET, ChrF) and human evaluation.

1 Introduction

Low-resource Indic languages refer to the vast array of languages spoken in India, which, unfortunately, possess limited linguistic resources available for their study and development. These languages typically suffer from a combination of factors that set them apart from the more prominent and widely supported languages spoken in the country. The challenges these languages face include having a smaller number of speakers, a relative lack of governmental support, inadequate documentation, and limited access to technological resources.

India is renowned for its linguistic diversity, with a rich tapestry of languages spoken across the subcontinent. The Eighth Schedule of the Indian Constitution officially recognizes 22 languages, which receive significant government backing and protection. However, beyond these major languages, numerous smaller languages and dialects are spoken by various indigenous and minority communities throughout the country.

These low-resource Indic languages face a series of interconnected challenges that make their preservation and promotion difficult: lack of written scripts, limited vocabulary resources, inadequate linguistic research, and insufficient digital content. The collective impact of these factors makes it challenging to preserve and promote low-resource Indic languages. As a consequence, they are at risk of falling into disuse, with their speakers shifting to more widely recognized languages. Efforts to document, revitalize, and support these languages are crucial not only for linguistic diversity but also for the preservation of cultural heritage and the rights of minority language communities in India.

Efforts are being made by various organizations, researchers, and language enthusiasts to address the issues faced by low-resource Indic languages (Pal et al., 2013a,b; Pal, 2018). These initiatives involve language documentation, the development of writing systems, the creation of linguistic resources such as parallel corpora (Ramesh et al., 2022), parallel fragment extraction from comparable corpora (Gupta et al., 2013; Pal et al., 2014), dictionaries and grammars, and the promotion of language use through educational programs and digital platforms.

Technology indeed plays a pivotal role in supporting low-resource Indic languages. In recent years, machine learning and natural language processing techniques have been harnessed to create innovative solutions for these languages, including speech recognition, machine translation, and text-to-speech systems. These technological advancements offer a transformative potential in addressing the linguistic challenges faced by these languages and can have a profound impact on their preservation and revitalization.

To work towards the goal of supporting low-resource Indic languages, we organized the "Indic MT Shared Task" focusing on several less-popular languages that belong to different language families. These languages include As-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 682–694
December 6–7, 2023. ©2023 Association for Computational Linguistics

samese (Indo-Aryan), Mizo (Sino-Tibetan), Khasi (Austroasiatic), and Manipuri (Sino-Tibetan). In this shared task, we present IndicNE-Corp1.0 in which parallel (English-Assamese (en–as), English–Mizo (en–lus), English–Khasi (en–kha), English–Manipuri (en–mni)) and monolingual (Assamese, Mizo, Khasi, Manipuri) corpora for northeastern Indic languages available.

2 Shared Task: Low-Resource Indic Language Translation

In recent years, there has been significant improvement in the performance of machine translation (MT) systems. This progress can be attributed to the development of new techniques, such as multilingual translation and transfer learning. As a result, the benefits of MT are no longer restricted to users of widely spoken languages. This advancement has led to a growing interest within the research community in expanding MT coverage to encompass a wider range of languages, each with its unique geographical presence, degree of diffusion, and level of digitalization.

However, despite the enthusiasm for extending MT to more languages and users, there remains a substantial challenge. The challenge stems from the fact that MT methods typically require large volumes of parallel data for training high-quality translation systems. This requirement has proven to be a major hurdle, particularly when dealing with low-resource languages where obtaining such extensive parallel data can be exceedingly difficult. Consequently, there is a pressing need to develop MT systems that can perform well even when trained on relatively small parallel datasets. The ability to achieve effective machine translation with limited resources is of paramount importance for increasing accessibility and usability across a wide spectrum of languages and linguistic communities. In this translation task, our focus was on the following language pairs (both directions for each):

- Subtask-1 : English $\leftrightarrow$ Assamese
- Subtask-2 : English $\leftrightarrow$ Mizo
- Subtask-3 : English $\leftrightarrow$ Khasi
- Subtask-4 : English $\leftrightarrow$ Manipuri

In this translation task, participants had the opportunity to submit up to 1 PRIMARY system for each language pair/translation direction, where no additional parallel data was permitted for training. Additionally, participants could submit up to 2 CONTRASTIVE systems for each language pair/translation direction. This structure allowed participants to showcase their translation systems under various conditions and constraints, including the absence of additional parallel data in the case of PRIMARY systems.

3 Dataset: IndicNE-Corp1.0

In the creation of IndicNE-Corp1.0, we compiled datasets from our prior research projects, including contributions from Laskar et al. (2020, 2022); Khenglawt et al. (2022); Laskar et al. (2021); Laitonjam and Ranbir Singh (2021). These datasets served as the foundation for constructing both parallel and monolingual corpora. In our earlier works, we undertook the development of English-Assamese (eng-asm) (Laskar et al., 2020, 2022), English-Mizo (eng-lus) (Khenglawt et al., 2022), English-khasi (eng-kha) (Laskar et al., 2021), English-Manipuri (eng-mni) (Laitonjam and Ranbir Singh, 2021) parallel and monolingual corpora for Assamese, Mizo, Khasi and Manipuri languages. The different online sources were explored that include Bible, multilingual online dictionary (Xobdo and Glosbe), multilingual question paper, PMIndia[1] (Haddow and Kirefu, 2020), web pages, blogs and online news papers. The collected data statistics for parallel (train, validation and test set) and monolingual corpora are presented in subsequent sections below. For primary investigation in this shared task, we have not included very complex sentences in the test set.

3.1 Assamese

Assamese exhibits a subject-object-verb (SOV) word order, in contrast to the subject-verb-object (SVO) word order found in English. Additionally, it is characterized as an agglutinative language, as discussed by Sarma et al. (2017) and Baruah et al. (2021), signifying its propensity to incorporate suffixes and prefixes into words to convey diverse grammatical meanings. This intricacy poses a notable challenge for machine translation systems, as they must accurately analyze and generate these intricate word forms.

Furthermore, Assamese boasts a complex verb conjugation system encompassing tense, aspect, mood, and agreement markers. These markers hold

[1] http://data.statmt.org/pmindia/v1/ parallel/

the power to significantly alter the meaning of a verb, making it a formidable task for translation systems to capture these subtleties with precision. The data statistics for the English-Assamese parallel data are presented in Table 1.

Type	Sentences	Tokens	
		eng	asm
Train	50,000	969,623	825,063
Validation	2,000	31,503	25,929
Test	2,000	32,466	27,483

Table 1: English-Assamese parallel data statistics for train, valid, and test set

3.2 Mizo

Mizo follows the object-subject-verb order when the object is considered. Mizo is a tonal language (Lalrempuii et al., 2021; Khenglawt et al., 2022), which means that differences in pitch or tone can represent different meanings. The vowels (a, aw, e, i, o, u) primarily indicate intonation. In the Mizo language, the main tones are rising, falling, high, and low. For example, depending on the tone, the word "ban" in Mizo can mean a pillar, the arm, to stretch, arrive at, sticky, or dismiss. A circumflex (ˆ) is frequently used to indicate long intonations (primarily to distinguish them from short intonations). Mizo is an agglutinative and highly inflected language with declension of nouns and pronouns. It also has many monosyllables and decomposable polysyllables, with meaning derived from each syllable. A sentence's tense can be changed by including particles such as "ang," "dawn," "mek," "tawh," and so on. The data statistics for the English-Mizo parallel data are presented in Table 2.

Type	Sentences	Tokens	
		eng	lus
Train	50,000	981,468	1,06,2414
Validation	1,500	38,525	40,983
Test	2,000	21,905	25,098

Table 2: English-Mizo parallel data statistics for train, valid and test set

3.3 Khasi

Khasi follows the subject-verb-object word order. It orthography has 23 alphabets and has six vowels, the vowels are "a e i ï o u". In Khasi orthography the alphabets "c f q v x z" are not present and instead the letters "ï ñ ng" are present which makes the orthography to be different from English or other orthographies (Warjri et al., 2021). Khasi is rich in subject agreement markers. Subject agreement is indicated by verbs, adjectives, and adverbs. Nouns have their own grammatical number and gender. In morphology, Khasi is mostly isolating; while some words are derived through specific morphological processing, others are found standing alone with no morphology indicated. As a result, (a) word categories such as Nouns, Verbs, Adjectives, and so on are invariant, and (b) words are mostly mono-morphemic in nature, so it is common to encounter only isolating words in a single long sentence or discourse. The data statistics for the English-Khasi parallel data are presented in Table 3.

Type	Sentences	Tokens	
		eng	kha
Train	24,000	7,29,930	8,75,545
Validation	1,000	24,609	37,407
Test	1,000	24,150	35,901

Table 3: English-Khasi parallel data statistics for train, valid and test set

3.4 Manipuri

Manipuri language uses Bengali script[2] and Meetei mayek[3] in written form. In this dataset, we use the Bengali script. Manipuri language also has an extensive suffix with limited prefixation and verb-final word order in a sentence, i.e., subject-object-verb order (Huidrom et al., 2021). Linguistic features of this language include agglutinative verb morphology, tone, the absence of grammatical person, number, gender, and a prevalence of aspect over tense. The data statistics for the English-Manipuri parallel data are presented in Table 4.

Type	Sentences	Tokens	
		eng	mni
Train	21,687	390,730	330,319
Validation	1,000	16,905	14,469
Test	1,000	14,886	12,775

Table 4: English-Manipuri parallel data statistics for train, valid, and test set

Table 5 presents the statistics for parallel data length differences among the four language pairs. In Figure 1, we illustrate the overlapping tokens between the test set and the training and validation sets for these same four language pairs. In addition to the parallel data, we have also made available monolingual corpora for Assamese, Mizo, Khasi,

[2]http://unicode.org/charts/PDF/U0980.pdf
[3]http://unicode.org/charts/PDF/UABC0.pdf

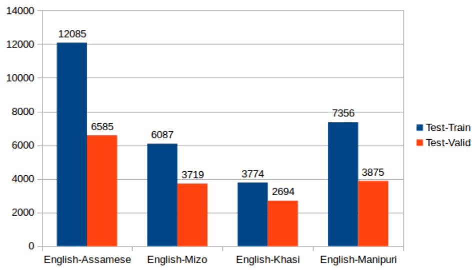

Figure 1: Overlapping tokens among test-train and test-validation data set of English-Assamese, English-Mizo, English-Khasi and English-Manipuri

and Manipuri. The monolingual data statistics per languages are presented in Table 6.

Language	Size (MB)	Sentences	Tokens
asm	805	2,624,715	49,232,154
lus	145	1,909,823	27,936,225
kha	104	182,737	22,140,361
mni	716	2,144,897	36,514,693

Table 6: Monolingual data statistics for Assamese, Mizo, Khasi and Manipuri languages

4 Participants and System Descriptions

In this shared task, a total of 31 teams registered and contributed, as indicated in Table 8, the released the dataset have been distributed among participants. In Table 7, we have compiled the system outputs submitted by participants, encompassing both primary and contrastive submission types.

Language Pair	Number of Participants
English-Assamese	13 (Primary), 11 (contrastive)
English-Mizo	10 (Primary), 8 (contrastive)
English-Khasi	11 (Primary), 8 (contrastive)
English-Manipuri	14 (Primary), 11 (contrastive)

Table 7: Number of participants in the low-resource Indic language translation task at WMT23

However, we have received system description papers from 9 teams and included concise system details for those teams where the authors provided such information.

CFILT-IITB (Gaikwad et al., 2023): The participant utilized phrase-pair injection (Sen et al., 2021), back-translation (Sennrich et al., 2016), and transfer learning with the help of large pre-trained multilingual IndicTrans2 model (Gala et al., 2023) to build NMT systems for the English-Assamese and English-Manipuri language pairs.

IOL Research (Zhang et al., 2023): The contributor used monolingual data to train two denoising language models similar to T5 (Raffel et al., 2020) and BART (Lewis et al., 2019), and then used parallel data to fine-tune the pre-trained language models to obtain two multilingual machine translation models. Besides, the multilingual machine translation models were used to translate English monolingual data into other multilingual data, forming multilingual parallel data as augmented data (Sennrich et al., 2016) to build NMT systems for English-Assamese, English-Mizo, English-Khasi, and English-Manipuri language pairs.

IACS-LRILT (Suman et al., 2023): The team IACS-LRILT used IndicBART (Dabre et al., 2022) pre-trained language model for fine-tuning the training data to build NMT systems for English-Assamese, and English-Manipuri language pairs.

GUIT-NLP (Ahmed et al., 2023): Team GUIT-NLP used back-translation (Sennrich et al., 2016) strategy and explored NMT systems by leveraging subword tokenization (Sennrich et al., 2015; Kudo and Richardson, 2018) and hyperparameters tuning for English-Assamese, English-Mizo, and English-Khasi language pairs.

NITS-CNLP (Singh et al., 2023): The NITS-CNLP team used the OpenNMT toolkit (Klein et al., 2017) and built a transformer-based (Vaswani et al., 2017) NMT model with hyperparameters tuning for the English-Manipuri language pair.

NICT-AI4B (Dabre et al., 2023): The group explored NMT systems by leveraging back-translation strategy (Sennrich et al., 2016) with denoising techniques (Lewis et al., 2020; Dabre et al., 2022) and fine-tuned IndicTrans2 model (Gala et al., 2023) for the English-Assamese, English-Mizo, English-Khasi, and English-Manipuri language pairs.

MUNI-NLP (Signoroni and Rychly, 2023): The participant explored transformer-based (Vaswani et al., 2017) NMT systems by investigating different hyperparameters tuning for English-Assamese, English-Mizo, English-Khasi, and English-Manipuri language pairs.

CUNI (Kvapilíková and Bojar, 2023): The CUNI team used back-translation (Sennrich et al., 2016) for data augmentation, denoising, leveraging multilingual masked language modelling, and

Data	Length	Number of Sentences			
		eng-asm	eng-lus	eng-kha	eng-mni
Test	1-10	435	1071	61	327
	11-20	1013	804	315	462
	21-30	481	120	381	164
	31-40	71	5	194	43
	41-50			49	4
Train	1-10	560	148	32	339
	11-20	910	437	341	335
	21-30	468	433	385	216
	31-40	62	292	194	86
	41-50		190	48	24
Valid	1-10	6895	10940	488	6351
	11-20	21032	16264	5559	7681
	21-30	18679	14316	7320	4764
	31-40	3245	8007	6656	1947
	41-50	149	473	3977	944

Table 5: Length-wise sentence group distribution for the test, train, and validation parallel data of English-Assamese, English-Mizo, English-Khasi, and English-Manipuri

built NMT systems for English-Assamese, English-Mizo, English-Khasi, and English-Manipuri language pairs.

ATULYA-NITS (Agrawal et al., 2023): This group used Google Colab, and trained the transformer model (Vaswani et al., 2017) using a T4 GPU for the English-Assamese, and English-Manipuri language pairs.

Organizer: The shared task organizer used the OpenNMT toolkit (Klein et al., 2017) and built biLSTM-based NMT systems with hyper-parameters tuning only on parallel data for the English-Assamese, English-Mizo, English-Khasi, and English-Manipuri language pairs.

5 Results and Discussion

We present results[4] for both directions of the four language pairs, namely, English-Assamese in Table 9, English-Mizo in Table 10, English-Khasi in Table 11, and English-Manipuri in Table 12. Here, we have reported the evaluation scores of those teams who submitted system output and their associated papers. To evaluate quantitative results, standard evaluation metrics (Papineni et al., 2002), namely, BLEU (bilingual evaluation under study), TER (translation error rate) (Snover et al., 2006), RIBES (rank-based intuitive bilingual evaluation score) (Isozaki et al., 2010), ChrF (character n-gram F-score) (Popović, 2015) and COMET (Rei et al., 2020). Moreover, we have hired linguistic experts who possess linguistic knowledge of the concerned language pair and randomly selected 20 sample sentences of primary submission type for manual evaluation (reported in Table 13 to 16). The human evaluator evaluates the candidate translations based on adequacy, fluency, and overall rating. Adequacy of translation measures the amount of meaning of reference translation, which is contained in a candidate translation. Furthermore, a translation is considered fluent if it is a well-formed sentence of the target language, irrespective of its correspondence with the reference translation. For example, given the reference translation to be "He wakes up early in the morning," the candidate translation "He is flying to Delhi" is inadequate, as it contains no content of the reference translation. However, the translation is fluent because the sentence has a proper meaning, and it is a well-formed sentence in the English language. The overall rating takes into account adequacy as well as fluency of candidate translation. An adequate and fluent translation is considered excellent and assigned a high overall rating. The human evaluation parameters have been rated on a scale of 0–5, with larger values signifying the better. Final adequacy, fluency, and overall rating scores are the average scores of individual test sentences.

[4]http://www2.statmt.org/wmt23/indic-mt-task.html

Team Name	Organization
BITS-P	Birla Institute of Technology and Science, Pilani, India
NITS-CNLP	**National Institute Of Technology, Silchar, India**
OneMT	IIIT-Hyderabad, India
SML lab	IISc, Bangalore, India
NICT-AI4B	**NICT Japan**
ANVITA	Centre for AI and Robotics (CAIR), India
MUNI-NLP	**Masaryk University, Czechia**
HV-NITS	National Institute Of Technology, Silchar, India
IREL-IIITH	IIIT HYDERABAD India
NVIDIA-India	NVIDIA, India
AIMLNLP-IITI	Indian Institute of Technology, Indore, India
NLP_NITH	NIT Hamirpur, India
TRANSSION MT	TRANNSION, China
CNLP-IISC	IISc, Bangalore, India
CUNI	**Charles University, Czechia**
A3-108	LTRC, IIIT Hyderabad, India
IOL Research	**Transn, China**
SLP-BV	Banasthali Vidyapith, India
IACS-LRILT	**Indian Assosciation for the Cultivation of Science, India**
NITR	NIT Rourkela, India
IIT-NLP lab	IIT dharwad, India
Team SiggyMorph	University of British Columbia, Canada
Lexical wizards	Kalinga Institute of Industrial Technology, India
JUNLP	Jadavpur University, India
ATULYA-NITS	**National Institute of Technology, Silchar, India**
CFILT-IITB	**Indian Institute of Technology, Bombay, India**
COGNITIVE LAB-IIITM	Indian Institute of Information Technology, Manipur, India
LRNMT-IIITH	IIIT Hyderabad, India
GUIT-NLP	**Gauhati University, India**
TRDDC	TCS Research, India
HW-TSC	Huawei Translate Center, China

Table 8: Registered participants in the low-resource Indic language translation task at WMT23 and dataset released to them. Not all the teams participated in all language pairs. **Bold marks** are those who submitted system outputs and system description papers

Discussion:

- For both directions of English-Assamese, Team: IACS-LRILT attains the best BLEU score (as shown in Table 9). They utilized the IndicBART language model in fine-tuning the training model. Also, Assamese-to-English translation attains higher scores than English-to-Assamese translation. It is due to the fact that Assamese is a highly inflectional, morphologically rich, and agglutinative language.
- For both directions of English-Mizo, Team: NICT-AI4B attains the best BLEU score (as shown in Table 10). They utilized IndicTrans2 model in fine-tuning the training model. It is observed that encountering tonal words for English-to-Mizo translation is a challenging task.
- For both directions of English-Khasi, Team: IOL Research attains the best BLEU score (as shown in Table 11). They used denoising language models (T5 / BART) and data augmentation techniques.
- For English-to-Manipuri translation, Team: CUNI attains the best BLEU score (as shown in Table 12). They used data augmentation, denoising, leveraging multilingual, and masked language modelling techniques. And, Manipuri-to-English translation, Team: IACS-LRILT attains the best BLEU score (as shown in Table 12) by utilizing the IndicBART language model in fine-tuning the training model. Also, it is observed that Manipuri-to-English translation attains higher scores than English-to-Manipuri translation. This is due to the fact that Manipuri is a morphologically rich and highly agglutinative language.
- In human evaluation, it is noticed that fluency scores are better than adequacy scores for all language pairs submission. The reason behind this is that NMT systems are well known for producing fluent translations (Koehn and Knowles, 2017).

6 Conclusion

We presented the results of the participating teams in the four language pairs translation task in terms of automatic and human evaluation metrics. We released a dataset, namely, IndicNE-Corp1.0 in the shared task on low-resource Indic language translation at the eighth conference on machine translation (WMT) 2023. The dataset comprises four low-resource languages, namely, Assamese, Mizo, Khasi, and Manipuri which belong to the northeastern region of India. In the future, we will include more northeastern Indic language datasets in addition to increasing the existing dataset size.

Comments

A few teams, namely, TRANSSION MT (TRANNSION, China), HW-TSC (Huawei Translate Center, China), ANVITA (Centre for AI and Robotics (CAIR), India), COGNITIVE LAB-IIITM (Indian Institute of Information Technology, Manipur, India) and NITR (NIT Rourkela, India) submitted system results but unfortunately did not submit the associated system description paper. Therefore, we have not reported their results in this paper.

Acknowledgements

We would like to thank all the participants for their active participation in this shared task. Also, thankful to the organizers and reviewers of the Eighth Conference on Machine Translation (WMT) 2023.

References

Goutam Agrawal, Rituraj Das, Anupam Biswas, and Dalton Meitei Thounaojam. 2023. Neural machine translation for english - manipuri and english - assamese.

Mazida Ahmed, Kuwali Talukdar, Parvez Boruah, Shikhar Kumar Sarma, and Kishore Kashyap. 2023. Guit-nlp's submission to shared task: Low resource indic language translation.

Rupjyoti Baruah, Rajesh Kumar Mundotiya, and Anil Kumar Singh. 2021. Low resource neural machine translation: Assamese to/from other indo-aryan (indic) languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(1).

Raj Dabre, Jay Gala, and Pranjal Chitale. 2023. Nict-ai4b's submission to the indic mt shared task in wmt 2023.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Team Name	Translation Type	BLEU	ChrF	RIBES	TER	COMET
IACS-LRILT	English-To-Assamese (Primary)	**34.82**	56.58	0.87	55.10	0.77
	English-To-Assamese (Contrastive-1)	34.71	56.59	0.87	54.75	0.78
	English-To-Assamese (Contrastive-2)	6.57	39.71	0.45	86.26	0.79
	Assamese-To-English (Primary)	**66.36**	75.88	0.93	37.44	0.84
	Assamese-To-English (Contrastive-1)	66.33	75.88	0.93	37.38	0.84
	Assamese-To-English (Contrastive-2)	23.19	48.42	0.61	71.79	0.75
CFILT-IITB	English-To-Assamese (Primary)	18.15	50.16	0.53	75.53	0.80
	English-To-Assamese (Contrastive)	18.15	50.16	0.53	75.53	0.80
	Assamese-To-English (Primary)	35.24	57.73	0.70	60.85	0.80
	Assamese-To-English (Contrastive)	35.24	57.73	0.70	60.85	0.80
NICT-AI4B	English-To-Assamese (Primary)	17.03	45.31	0.58	76.57	0.78
	English-To-Assamese (Contrastive-2)	21.07	51.71	0.58	73.03	0.81
	English-To-Assamese (Contrastive-1)	18.09	51.98	0.57	73.41	0.82
	Assamese-To-English (Primary)	27.02	50.71	0.71	62.46	0.76
	Assamese-To-English (Contrastive-1)	37.28	59.97	0.72	58.81	0.81
	Assamese-To-English (Contrastive-2)	36.97	59.82	0.72	58.53	0.81
IOL Research	English-To-Assamese (Primary)	14.35	43.87	0.63	73.37	0.78
	English-To-Assamese (Contrastive)	14.10	43.66	0.63	72.77	0.78
	Assamese-To-English (Primary)	28.73	51.99	0.76	57.06	0.78
	Assamese-To-English (Contrastive)	27.83	51.45	0.76	57.44	0.78
CUNI	English-To-Assamese (Primary)	13.92	41.66	0.60	80.26	0.76
	English-To-Assamese (Contrastive-1)	3.98	41.57	0.59	78.91	0.75
	English-To-Assamese (Contrastive-2)	3.88	39.68	0.59	80.34	0.75
	Assamese-To-English (Primary)	20.71	44.94	0.69	73.56	0.72
	Assamese-To-English (Contrastive-1)	17.49	42.21	0.65	80.90	0.70
	Assamese-To-English (Contrastive-2)	16.85	41.55	0.65	85.24	0.70
Organizer	English-To-Assamese (Primary)	8.57	25.24	0.44	86.14	0.59
	Assamese-To-English (Primary)	11.28	28.70	0.53	83.10	0.56
MUNI-NLP	English-To-Assamese (Primary)	7.96	27.31	0.31	91.38	0.59
	Assamese-To-English (Primary)	11.29	30.13	0.64	73.39	0.64
ATULYA-NITS	English-To-Assamese (Primary)	5.47	21.66	0.32	96.76	0.57
	Assamese-To-English (Primary)	8.50	24.36	0.45	89.53	0.53
GUIT-NLP	English-To-Assamese (Primary)	4.89	25.16	0.46	87.21	0.61
	English-To-Assamese (Contrastive-1)	4.27	24.59	0.43	90.13	0.59
	English-To-Assamese (Contrastive-2)	3.75	22.65	0.42	93.57	0.58
	Assamese-To-English (Primary)	5.50	25.81	0.56	80.10	0.57
	Assamese-To-English (Contrastive-1)	4.70	24.96	0.55	81.53	0.56
	Assamese-To-English (Contrastive-2)	4.14	23.73	0.53	83.41	0.55

Table 9: Automatic evaluation scores of participated teams for English-Assamese language pair

Team Name	Translation Type	BLEU	ChrF	RIBES	TER	COMET
NICT-AI4B	English-To-Mizo (Primary)	33.18	56.73	0.73	55.68	0.70
	English-To-Mizo (Contrastive-2)	**33.64**	56.88	0.72	57.71	0.71
	English-To-Mizo (Contrastive-1)	26.47	50.60	0.66	65.97	0.69
	Mizo-To-English (Primary)	32.47	51.33	0.69	60.56	0.67
	Mizo-To-English (Contrastive-2)	**33.30**	52.74	0.70	60.87	0.68
	Mizo-To-English (Contrastive-1)	28.47	47.93	0.61	67.54	0.69
CUNI	English-To-Mizo (Primary)	31.20	54.56	0.76	54.54	0.70
	English-To-Mizo (Contrastive-1)	31.28	54.58	0.76	54.20	0.70
	English-To-Mizo (Contrastive-2)	30.66	54.48	0.76	54.98	0.69
	Mizo-To-English (Primary)	29.47	49.98	0.73	60.44	0.66
	Mizo-To-English (Contrastive-1)	28.63	48.58	0.72	62.21	0.65
	Mizo-To-English (Contrastive-2)	28.53	49.51	0.73	62.55	0.66
IOL Research	English-To-Mizo (Primary)	28.24	54.02	0.78	53.04	0.70
	English-To-Mizo (Contrastive-1)	27.74	53.71	0.78	53.40	0.70
	Mizo-To-English (Primary)	32.54	51.83	0.78	53.48	0.71
	Mizo-To-English (Contrastive-1)	31.37	50.94	0.77	55.37	0.70
Organizer	English-To-Mizo (Primary)	23.67	45.1	0.71	62.29	0.67
	Mizo-To-English (Primary)	22.59	39.53	0.66	68.83	0.57
GUIT-NLP	English-To-Mizo (Primary)	23.29	46.72	0.75	59.93	0.68
	English-To-Mizo (Contrastive-1)	23.78	48.06	0.75	58.07	0.69
	Mizo-To-English (Primary)	18.81	40.33	0.66	73.65	0.57
	Mizo-To-English (Contrastive-1)	18.51	41.32	0.67	73.70	0.60
MUNI-NLP	English-To-Mizo (Primary)	20.48	45.60	0.73	61.22	0.68
	Mizo-To-English (Primary)	23.16	43.02	0.72	62.31	0.63

Table 10: Automatic evaluation scores of participated teams for English-Mizo language pair

Team Name	Translation Type	BLEU	ChrF	RIBES	TER	COMET
IOL Research	English-To-Khasi (Primary)	**21.63**	44.47	0.72	62.10	0.68
	English-To-Khasi (Contrastive)	21.48	44.30	0.65	62.55	0.68
	Khasi-To-English (Primary)	**20.72**	43.34	0.72	71.78	0.63
	Khasi-To-English (Contrastive)	20.60	43.09	0.58	71.35	0.63
NICT-AI4B	English-To-Khasi (Primary)	19.95	43.30	0.68	66.47	0.67
	English-To-Khasi (Contrastive-2)	21.05	46.06	0.65	73.80	0.68
	English-To-Khasi (Contrastive-1)	20.77	43.82	0.65	69.51	0.68
	Khasi-To-English (Primary)	17.80	39.22	0.66	74.10	0.60
	Khasi-To-English (Contrastive-1)	20.06	40.33	0.58	78.44	0.60
	Khasi-To-English (Contrastive-2)	20.02	39.82	0.59	77.50	0.59
CUNI	English-To-Khasi (Primary)	16.64	39.92	0.65	70.69	0.67
	English-To-Khasi (Contrastive-1)	16.49	40.00	0.65	69.92	0.67
	English-To-Khasi (Contrastive-2)	15.79	38.79	0.65	71.29	0.66
	Khasi-To-English (Primary)	13.84	37.05	0.65	79.73	0.58
	Khasi-To-English (Contrastive-1)	12.71	36.32	0.66	81.37	0.57
	Khasi-To-English (Contrastive-2)	11.55	35.62	0.64	87.54	0.56
MUNI-NLP	English-To-Khasi (Primary)	13.90	37.31	0.61	73.99	0.65
	Khasi-To-English (Primary)	12.71	34.55	0.65	78.15	0.56
GUIT-NLP	English-To-Khasi (Primary)	10.41	33.31	0.63	71.67	0.64
	Khasi-To-English (Primary)	8.74	30.54	0.63	79.64	0.52
Organizer	English-To-Khasi (Primary)	10.08	31.13	0.59	75.57	0.62
	Khasi-To-English (Primary)	8.02	28.04	0.56	86.94	0.49

Table 11: Automatic evaluation scores of participated teams for English-Khasi language pair

Team Name	Translation Type	BLEU	ChrF	RIBES	TER	COMET
CUNI	English-To-Manipuri (Primary)	**29.50**	59.85	0.73	60.60	0.74
	English-To-Manipuri (Contrastive-1)	5.96	60.96	0.75	58.97	0.75
	English-To-Manipuri (Contrastive-2)	5.86	60.13	0.73	60.25	0.74
	Manipuri-To-English (Primary)	36.08	62.29	0.76	61.19	0.76
	Manipuri-To-English (Contrastive-1)	33.62	60.29	0.75	65.96	0.75
	Manipuri-To-English (Contrastive-2)	31.03	59.08	0.74	77.42	0.74
NICT-AI4B	English-To-Manipuri (Primary)	27.36	61.60	0.74	58.28	0.76
	English-To-Manipuri (Contrastive-2)	27.40	61.55	0.74	58.16	0.76
	English-To-Manipuri (Contrastive-1)	24.17	62.95	0.70	62.85	0.76
	Manipuri-To-English (Primary)	39.40	64.70	0.77	51.27	0.79
	Manipuri-To-English (Contrastive-1)	46.06	69.96	0.80	47.44	0.83
	Manipuri-To-English (Contrastive-2)	43.35	69.27	0.80	47.43	0.82
CFILT-IITB	English-To-Manipuri (Primary)	26.36	63.48	0.70	62.04	0.76
	English-To-Manipuri (Contrastive-1)	26.36	63.48	0.70	62.04	0.76
	Manipuri-To-English (Primary)	47.54	70.41	0.81	47.17	0.83
	Manipuri-To-English (Contrastive-1)	47.54	70.41	0.81	47.17	0.83
IACS-LRILT	English-To-Manipuri (Primary)	25.78	49.94	0.84	60.43	0.71
	English-To-Manipuri (Contrastive-1)	25.82	49.93	0.84	60.57	0.71
	English-To-Manipuri (Contrastive-2)	9.69	40.45	0.54	81.18	0.67
	Manipuri-To-English (Primary)	**69.75**	78.16	0.94	32.08	0.84
	Manipuri-To-English (Contrastive-1)	69.75	78.16	0.94	32.10	0.84
	Manipuri-To-English (Contrastive-2)	22.10	48.03	0.63	72.19	0.70
IOL Research	English-To-Manipuri (Primary)	23.51	60.03	0.74	60.68	0.75
	English-To-Manipuri (Contrastive)	23.05	59.85	0.70	61.04	0.75
	Manipuri-To-English (Primary)	42.68	67.55	0.83	46.27	0.82
	Manipuri-To-English (Contrastive)	42.48	67.51	0.80	46.31	0.82
NITS-CNLP	English-To-Manipuri (Primary)	22.75	48.35	0.61	70.02	0.70
	Manipuri-To-English (Primary)	26.92	48.64	0.65	67.62	0.66
Organizer	English-To-Manipuri (Primary)	21.58	45.97	0.61	69.76	0.69
	Manipuri-To-English (Primary)	24.86	46.37	0.64	70.26	0.63
MUNI-NLP	English-To-Manipuri (Primary)	19.65	53.26	0.66	69.70	0.72
	Manipuri-To-English (Primary)	32.18	58.71	0.76	56.35	0.74
	Manipuri-To-English (Contrastive)	32.18	58.71	0.74	67.86	0.74
ATULYA-NITS	English-To-Manipuri (Primary)	15.02	35.96	0.46	85.96	0.65
	Manipuri-To-English (Primary)	18.70	38.49	0.54	81.02	0.59

Table 12: Automatic evaluation scores of participated teams for English-Manipuri language pair

Team Name	Translation Type	Adequacy	Fluency	Overall Rating
NICT-AI4B	English-To-Assamese (Primary)	3.60	4.35	3.98
	Assamese-To-English (Primary)	3.75	4.30	4.03
CFILT-IITB	English-To-Assamese (Primary)	2.80	3.85	3.33
	Assamese-To-English (Primary)	3.50	4.35	3.93
IACS-LRILT	English-To-Assamese (Primary)	2.55	3.20	2.88
	Assamese-To-English (Primary)	3.20	3.35	3.28
IOL Research	English-To-Assamese (Primary)	3.10	4.20	3.65
	Assamese-To-English (Primary)	3.70	4.60	4.15
CUNI	English-To-Assamese (Primary)	3.60	4.05	3.82
	Assamese-To-English (Primary)	2.85	3.80	3.32
Organizer	English-To-Assamese (Primary)	1.60	3.05	4.65
	Assamese-To-English (Primary)	1.50	2.55	2.02
MUNI-NLP	English-To-Assamese (Primary)	1.35	3.35	2.35
	Assamese-To-English (Primary)	1.50	2.45	1.97
ATULYA-NITS	English-To-Assamese (Primary)	1.50	2.95	2.22
	Assamese-To-English (Primary)	1.30	2.60	3.90
GUIT-NLP	English-To-Assamese (Primary)	1.35	3.05	2.20
	Assamese-To-English (Primary)	1.00	2.45	3.45

Table 13: Human evaluation score of English-Assamese

Team Name	Translation Type	Adequacy	Fluency	Overall Rating
NICT-AI4B	English-To-Mizo (Primary)	3.60	4.25	3.92
	Mizo-To-English (Primary)	3.10	4.50	3.80
CUNI	English-To-Mizo (Primary)	2.85	4.35	3.60
	Mizo-To-English (Primary)	3.30	4.40	3.85
IOL Research	English-To-Mizo (Primary)	3.95	4.45	4.20
	Mizo-To-English (Primary)	3.75	4.55	4.15
Organizer	English-To-Mizo (Primary)	2.05	3.55	2.80
	Mizo-To-English (Primary)	1.60	3.35	2.47
MUNI-NLP	English-To-Mizo (Primary)	3.05	3.85	3.45
	Mizo-To-English (Primary)	2.50	4.20	3.35
GUIT-NLP	English-To-Mizo (Primary)	3.25	4.15	3.70
	Mizo-To-English (Primary)	2.00	3.75	2.87

Table 14: Human evaluation score of English-Mizo

Team Name	Translation Type	Adequacy	Fluency	Overall Rating
IOL Research	English-To-Khasi (Primary)	4.45	4.75	4.60
	Khasi-To-English (Primary)	4.30	4.70	4.50
NICT-AI4B	English-To-Khasi (Primary)	4.20	4.60	4.40
	Khasi-To-English (Primary)	3.70	4.40	4.05
CUNI	English-To-Khasi (Primary)	3.30	4.20	3.75
	Khasi-To-English (Primary)	3.40	4.40	3.90
MUNI-NLP	English-To-Khasi (Primary)	2.70	4.50	3.60
	Khasi-To-English (Primary)	2.65	4.05	3.35
GUIT-NLP	English-To-Khasi (Primary)	2.80	4.60	3.70
	Khasi-To-English (Primary)	2.45	3.80	3.12
Organizer	English-To-Khasi (Primary)	1.95	4.05	3.00
	Khasi-To-English (Primary)	1.80	3.45	2.62

Table 15: Human evaluation score of English-Khasi

Team Name	Translation Type	Adequacy	Fluency	Overall Rating
CUNI	English-To-Manipuri (Primary)	3.25	3.55	3.45
	Manipuri-To-English (Primary)	3.05	3.15	3.00
NICT-AI4B	English-To-Manipuri (Primary)	2.95	4.10	3.50
	Manipuri-To-English (Primary)	3.50	3.50	3.45
CFILT-IITB	English-To-Manipuri (Primary)	4.25	4.50	4.35
	Manipuri-To-English (Primary)	4.80	4.75	4.75
IACS-LRILT	English-To-Manipuri (Primary)	2.45	2.65	2.45
	Manipuri-To-English (Primary)	3.45	3.45	3.45
IOL Research	English-To-Manipuri (Primary)	2.80	4.60	3.70
	Manipuri-To-English (Primary)	3.95	4.00	3.95
NITS-CNLP	English-To-Manipuri (Primary)	2.45	3.05	2.70
	Manipuri-To-English (Primary)	2.15	2.50	2.20
Organizer	English-To-Manipuri (Primary)	2.50	3.50	2.95
	Manipuri-To-English (Primary)	2.05	2.10	2.05
MUNI-NLP	English-To-Manipuri (Primary)	3.00	3.50	3.15
	Manipuri-To-English (Primary)	3.20	3.30	3.20
ATULYA-NITS	English-To-Manipuri (Primary)	1.75	2.15	1.95
	Manipuri-To-English (Primary)	1.80	1.85	1.80

Table 16: Human evaluation score of English-Manipuri

Pranav Gaikwad, Meet Doshi, Sourabh Deoghare, and Pushpak Bhattacharyya. 2023. Machine translation advancements for low-resource indian languages in wmt23: Cfilt-iitb's effort for bridging the gap.

Jay Gala, Pranjal A Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Rajdeep Gupta, Santanu Pal, and Sivaji Bandyopadhyay. 2013. Improving mt system using extracted parallel fragments of text from comparable corpora. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 69–76.

Barry Haddow and Faheem Kirefu. 2020. Pmindia - A collection of parallel corpora of languages of india. *CoRR*, abs/2001.09907.

Rudali Huidrom, Yves Lepage, and Khogendra Khomdram. 2021. EM corpus: a comparable corpus for a less-resourced language pair Manipuri-English. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 60–67, Online (Virtual Mode). INCOMA Ltd.

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, Cambridge, MA. Association for Computational Linguistics.

Vanlalmuansangi Khenglawt, Sahinur Rahman Laskar, Santanu Pal, Partha Pakray, and Ajoy Kumar Khan. 2022. Language resource building and English-to-mizo neural machine translation encountering tonal words. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 48–54, Marseille, France. European Language Resources Association.

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

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.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Ivana Kvapilíková and Ondřej Bojar. 2023. Low-resource machine translation systems for indic languages.

Lenin Laitonjam and Sanasam Ranbir Singh. 2021. Manipuri-English machine translation using comparable corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 78–88, Virtual. Association for Machine Translation in the Americas.

Candy Lalrempuii, Badal Soni, and Partha Pakray. 2021. An improved english-to-mizo neural machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(4).

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji Darsh Kaushik, Partha Pakray, and Sivaji Bandyopadhyay. 2021. EnKhCorp1.0: An English–Khasi corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages*

(*LoResMT2021*), pages 89–95, Virtual. Association for Machine Translation in the Americas.

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, Riyanka Manna, Partha Pakray, and Sivaji Bandyopadhyay. 2022. A domain specific parallel corpus and enhanced english-assamese neural machine translation. *Computación y Sistemas*, 26(4):1669—-1687.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Santanu Pal. 2018. A hybrid machine translation framework for an improved translation workflow.

Santanu Pal, Mahammed Hasanuzzaman, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2013a. Impact of linguistically motivated shallow phrases in pb-smt. In *ICON 2013*. https://www.researchgate.net/publication ....

Santanu Pal, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2013b. Mwe alignment in phrase based statistical machine translation. In *Proceedings of the XIV Machine Translation Summit*, pages 61–68.

Santanu Pal, Partha Pakray, and Sudip Kumar Naskar. 2014. Automatic building and using parallel resources for smt from comparable corpora. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 48–57.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Himangshu Sarma, Navanath Saharia, and Utpal Sharma. 2017. Development and analysis of speech recognition systems for assamese language using htk. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(1).

Sukanta Sen, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, and Andy Way. 2021. Neural machine translation of low-resource languages using SMT phrase pair injection. *Nat. Lang. Eng.*, 27(3):271–292.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

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)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Edoardo Signoroni and Pavel Rychly. 2023. Muni-nlp systems for low-resource indic machine translation.

Kshetrimayum Boynao Singh, Avichandra Singh Ningthoujam, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023. Nits-cnlp low-resource neural machine translation systems of english-manipuri language pair.

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.

Dhairya Suman, Atanu Mandal, Santanu Pal, and Sudip Naskar. 2023. Iacs-lrilt: Machine translation for low-resource indic languages.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Sunita Warjri, Partha Pakray, Saralin A. Lyngdoh, and Arnab Kumar Maji. 2021. Part-of-speech (pos) tagging using deep learning-based approaches on the designed khasi pos corpus. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(3).

Wenbo Zhang, Zeyu Yan, Qiaobo Deng, Jie Cai, and Hongbao Mao. 2023. Iol research machine translation systems for wmt23 low-resource indic language translation shared task.