Statistical Machine Translation Systems of English-Pnar Language Pair : Some Insights of the Emperical Study

Edawanbiang Dhar[1], Surmila Thokchom[2], and Thoudam Doren Singh[1]

[1] Human Language Technology Lab, Department of Computer Science and Engineering,
National Institute of Technology Meghalaya, Sohra, Meghalaya-793108, India
p24cs005@nitm.ac.in, doren.thoudam@nitm.ac.in
[2] Department of Computer Science and Engineering,
National Institute of Technology Meghalaya, Sohra, Meghalaya-793108, India
surmila.thokchom@nitm.ac.in

Abstract. Pnar, an Austroasiatic language spoken by approximately 0.4 million people in the Jaintia Hills of Meghalaya, lacks the digital corpora and natural language processing (NLP) resources. This paper presents the first machine translation study for the English–Pnar language pair. Using articles collected from the *Wyrta* newspaper, we built a parallel corpus comprising of 10,234 sentences and trained phrase-based statistical machine translation (SMT) systems the models using 9,563 parallel corpora under three configurations for each direction using Moses, GIZA++, KenLM, varying lexicalized reordering and minimum error rate training (MERT) tuning. The models are evaluated on a held-out test set of 371 sentences, the best-performing system achieves a BLEU score of 14.97 (chrF2: 33.42, TER: 77.60) for Pnar→English and 11.16 (chrF2: 31.38, TER: 93.51) for English→Pnar, establishing the first quantitative benchmark for this language pair. Lexicalized reordering improves translation quality by 3.73 BLEU points for Pnar→English, reflecting the structural shift from the source language's SOV word order to the target language's SVO order, whereas MERT tuning degrades BLEU performance under low-resource conditions. Finally, we analyze the remaining translation errors, including morphological out-of-vocabulary (OOV) words, long-distance reordering, and Khasi code-mixing and discuss future directions toward neural and multilingual machine translation for Pnar.

Keywords: Moses SMT Toolkit, KenLM, Pnar, BLEU, Astro-Asiatic Languages

1 Introduction

Pnar is a severely under-resourced language spoken by the Jaintia community in Meghalaya in the Northeast India. Despite its linguistic proximity to Khasi

which has received limited attention in natural language processing, Pnar remains largely unexplored in computational research. It is written in the Roman script and belongs to the Austroasiatic language family characterized by distinct morphological and syntactic features that distinguish it from more widely studied Indo-European languages. The scarcity of computational resources for Pnar reflects broader challenges faced by low-resource languages in natural language processing [1]. Developing language technologies for such languages is crucial for their preservation and continued use and particularly by enabling access to digital content, education and communication tools for native speakers [3]. In this context, establishing a Pnar-English machine translation system represents an important step toward advancing research in low-resource NLP as well as supporting linguistic sustainability. Pnar, also known as Jaintia (Spencer 1967) or earlier as Synteng (Grierson 1928), is a language spoken in the East and West Jaintia Hill District of Meghalaya and in a few pockets of Cachar Hills and North Cachar Hills districts of Assam [2]. According to the People's Linguistic Survey of India (2014), the population of Pnar stands at 3,92,853. Pnar is also subject to regional variations which can be noticeable even within a distance of 3-4 kilometres. The most noticeable variants of Pnar can be classified according to their regional locations; in most cases, the regional variants take their names from village names [2]. To address these challenges, we document this low-resource language, Pnar, by building a corpus using local newspapers (Wyrta). Machine Translation (MT) has witnessed remarkable progress in recent years, driven by large-scale parallel corpora and powerful neural architectures such as the Transformer [14]. The vast majority of the world's approximately 7,000 languages remain severely under-resourced, lacking both parallel data and computational tools necessary for developing effective MT systems [1]. This paper makes three main contributions. First, we built the first sizeable parallel corpus for Pnar of a local newspaper. Second, we train and compare phrase-based statistical machine translation (SMT) systems under six configurations (three per translation direction), quantifying the individual effects of lexicalized reordering and minimum error rate training (MERT) tuning in low-resource settings. Third, we establish the first published baseline for English–Pnar machine translation using BLEU, chrF2, and TER evaluation metrics, together with an error analysis that highlights challenges arising from Pnar morphology and its code-mixing with Khasi. Unlike prior machine translation studies on other Northeast Indian languages, such as Manipuri, Mizo, Bodo, and Khasi, no computational study has previously targeted Pnar. This work therefore addresses an important gap in the literature.

2 Related Work

Low-resource machine translation has attracted considerable research attention. Araabi and Monz [1] showed that Transformer architecture optimization, including careful regularization and reduced model capacity benefits low-resource neural MT settings. Bird [3] discusses the broader decolonization imperative in

speech and language technology arguing that community-centered approaches are essential for endangered language documentation. For Northeast Indian languages, SMT and neural MT systems have been reported for Manipuri, Mizo, Bodo, and Khasi. These prior efforts share a common approach namely, phrase-based or low-resource neural machine translation (MT) systems trained on small, domain-specific parallel corpora. However, none has been extended to Pnar, whose closest documented linguistic relative, Khasi, has only recently begun to receive attention from the computational linguistics community. This gap is significant because the linguistic characteristics of Pnar including its SOV constituent order, agglutinative verbal morphology and frequent code-mixing with Khasi, make it substantially different from previously studied low-resource languages. In addition to these structural differences from English, Pnar also suffers from an acute scarcity of digital resources, including even monolingual text. Consequently, the present study is not merely a replication of existing SMT pipelines developed for Khasi or Manipuri; rather, it represents the first systematic investigation of how effectively a standard phrase-based SMT framework performs under these combined linguistic and resource constraints while identifying the specific challenges that limit its performance. Among Austroasiatic languages, there is also report on the initiative of developing Santali–English translation system. However, to the best of our knowledge, no prior computational MT work has addressed Pnar specifically making this study the first of its kind. The Moses toolkit [7] used in this study has been widely applied in low-resource scenarios due to its robustness with small training sets and its well-understood behavior. KenLM [5] provides an efficient language model backend that makes higher-order n-gram models practical even in resource-constrained environments. MGIZA++ [4] enables parallel word alignment using IBM Models substantially reducing training time.

3 Dataset Preparation

3.1 Data Sources

Given the scarcity of digital Pnar text, corpus construction required identification and digitization from multiple sources. We collected Pnar dataset from a single primary source. First, contents of the Wyrta newspaper are collected, providing contemporaneous general-domain text covering community news, local governance, cultural events, sports and health topics.

3.2 Data Preprocessing

Data preprocessing follows standard SMT pipeline conventions. Text is lowercased and tokenised using the Moses tokeniser adapted for Roman-script Pnar text. Punctuation normalisation is applied to handle inconsistencies across newspaper scans and OCR output. Since Pnar lacks a standard spell-checker, manual lexicon-based correction was performed for the 500 most frequent tokens.

Table 1. Sentence counts, token counts and average sentence length for the train, development and test splits of the English-Pnar parallel corpus statistics

Metric	Train	Dev	Test
Sentences	9,563	300	371
Pnar tokens	∼234,800	∼7,350	∼5,777
English tokens	∼229,400	∼7,200	∼5,865
Average Pnar sentence length	∼24.6	∼24.5	∼15.6
Average English sentence length	∼24.0	∼24.0	∼15.9

Sentence pairs with either side exceeding 80 tokens were removed to improve alignment quality. Trailing comma artefacts introduced during corpus splitting are cleaned using regular expression filters. The final cleaned corpus is split into training (9,563 sentences), development (300 sentences, lines 9,563–9,863, strictly non overlapping with training), and test sets (371 sentences from a held-out evaluation set).

4 Experimental Setup

A phrase-based model is trained using the Moses toolkit [7]. Moses is a widely used open-source framework that implements phrase-based translation models, supporting modular components such as translation models, reordering models and decoding strategies. The SMT pipeline includes word alignment using GIZA++, phrase extraction and a 5-gram language model trained with KenLM [5] on the target-side training data. Standard Minimum Error Rate Training (MERT) tuning is applied on the validation set to finetune translation quality. Fig 1. shows the Phrase-based SMT model architecture of the proposed translation system.

Moses SMT Framework We train phrase-based Statistical Machine Translation (SMT) systems using the Moses toolkit [7], a widely used open-source framework that implements phrase-based translation models with modular components, including translation models, reordering models, and decoding strategies. The SMT approach formulates translation as the problem of finding the most probable target sentence e^* given a source sentence f, defined as 1:

$$e^* = \arg\max_e \; p(f \mid e)\, p(e) \tag{1}$$

where $p(f \mid e)$ represents the translation model probability and $p(e)$ denotes the language model probability. The decoder performs a search for the optimal hypothesis using beam search over the combined model score.

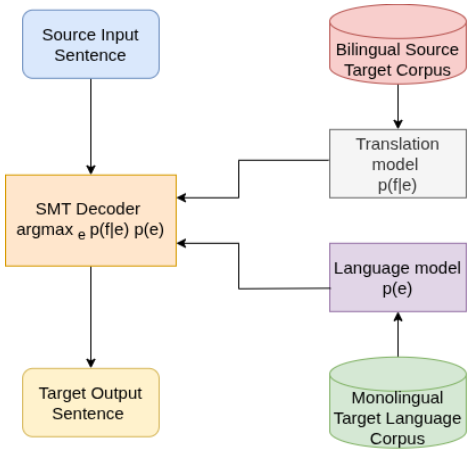

Fig. 1. Phrase-based SMT model architecture : the decoder combines translation-model and language-model scores over the bilingual and monolingual training corpora to produce the target sentence.

4.1 Word Alignment and Phrase Extraction

Phrase alignment is performed using GIZA++ [10] implementing IBM Models in both source-to-target and target-to-source directions. The resulting bidirectional alignments are symmetrized using the *grow-diag-final-and* heuristic [8], which has shown to produce high-quality phrase alignment and improved phrase coverage. Phrase pairs are extracted from the symmetrized word alignments with a maximum phrase length of 7 words. Phrase translation probabilities are estimated using relative frequency over extracted phrase pair counts. Lexical weighting features are computed from word-level translation tables to provide more fine-grained translation scoring. Having established the corpus and evaluation protocol (Section 4), we now describe the three main components of the statistical machine translation (SMT) pipeline, word alignment, the language model and the reordering model followed by the tuning procedure that integrates them.

4.2 Language Model

For Pnar→English systems, the language model is trained on the 8,05,946 English target-side training sentences. For English→Pnar systems, the language model is trained on a monolingual Pnar corpus of 23,429 sentences drawn from *Wyrta* newspaper issues. Two 5-gram KenLM language models are built on the target-side data using the 8,05,946 English target monolingual sentences for Pnar→English translation and a 23,429-sentence monolingual Pnar corpus for English→Pnar translation respectively. In both cases, a 5-gram language model with modified Kneser–Ney smoothing is trained using KenLM [5] which utilises

probing hash tables and trie-based data structures to achieve fast query performance and reduced memory usage. The language model plays a critical role in ensuring fluent and grammatically coherent translations by assigning higher probabilities to well-formed target sentences as given by equation 2 :

$$p(e) = \prod_{i=1}^{n} p(e_i \mid e_1^{i-1}) \tag{2}$$

4.3 Reordering Model

Given the SOV (Subject–Object–Verb) versus SVO (Subject–Verb–Object) word order divergence between Pnar and English, lexicalized reordering models play a crucial role in improving translation quality for this language pair. We employ a lexicalized reordering model based on phrase-pair orientation types including *monotone*, *swap*, *discontinuous-left* and *discontinuous-right* configurations [13]. Reordering model weights are estimated from the training data jointly with phrase translation probabilities within the phrase-based SMT framework [7]. This component is particularly important for handling Pnar verb-final constructions which require reordering when translated into English as well as for managing differences in nominal modifier ordering between the two languages.

4.4 Tuning

Minimum Error Rate Training (MERT) [9] is applied on the validation set to optimize the log-linear combination of model feature weights with respect to the BLEU score [11]. The SMT system is modeled as a log-linear combination of features, including phrase translation probabilities, language model scores, reordering model features, and a word penalty term. The feature weights are tuned on the development set using MERT for up to 25 iterations or until convergence is reached. This optimization procedure directly maximizes translation quality as measured by automatic evaluation metrics such as BLEU. Three system configurations are trained and evaluated in each translation direction (Table 2).

Table 2. The three SMT configurations in each translation direction varying lexicalized reordering and MERT tuning

System	**Reordering**	**Tuning**	**LM**
Config-1	✓ msd-bidirectional	None (untuned)	5-gram KenLM
Config-2	✓ msd-bidirectional	MERT (25 iter.)	5-gram KenLM
Config-3	× None	None (untuned)	5-gram KenLM

5 Experiments and Results

5.1 Results

Tables 3(a) and 3(b) present evaluation results for all configurations on the Pnar→English and English→Pnar test sets respectively.

Table 3. Translation performance on the 371-sentence test set.

(a) Pnar → English

Configuration	BLEU ↑	chrF2 ↑	TER ↓
Config-3	11.24	27.44	91.70
Config-1	**14.97**	33.42	**77.60**
Config-2	12.43	**36.86**	95.04

(b) English → Pnar

Configuration	BLEU ↑	chrF2 ↑	TER↓
Config-3	**11.16**	31.38	**93.51**
Config-1	10.14	30.98	86.53
Config-2	6.91	**32.82**	124.93

5.2 Impact of Lexicalized Reordering

Table 4 summarises the impact of the lexicalized reordering model on translation performance.

Table 4. BLEU, chrF, TER scores with vs. without lexicalized reordering, isolating the reordering effect from MERT tuning (both rows untuned).

Direction	Config	BLEU ↑	chrF2↑	TER↓
Pn→En	No reorder	11.24	27.44	91.70
	Reorder	**14.97**	**33.42**	**77.60**
	Δ	+3.73	+5.98	-14.10
En→Pn	No reorder	**11.16**	**31.38**	93.51
	Reorder	10.14	30.98	**86.53**
	Δ	-1.02	-0.40	-6.98

The reordering model provides a substantial gain of +3.73 BLEU and +5.98 chrF2 for Pnar→English reflecting the significant SOV→SVO structural transformation required in this direction. The TER improvement of −14.10 is particularly noteworthy indicating that reordered hypotheses require substantially fewer post-editing operations. For English→Pnar, the reordering model produces a marginal BLEU decline (−1.02). Qualitative analysis reveals that without reordering the En→Pn model largely copies source English tokens with minor shuffling yielding artificially inflated BLEU scores through named entity. The

with-reordering system (Config-1, BLEU 10.14) produces genuine Pnar output and constitutes the better baseline. This asymmetry where lexicalized reordering improves Pnar→English translation but not the reverse direction is consistent with the behavior of reordering models under low-resource conditions. Specifically, the model is able to learn the transformation from the Pnar SOV word order to the English SVO structure thereby reducing structural divergence during decoding. In contrast, the reverse transformation from English SVO to Pnar SOV is more difficult to capture using the same limited phrase table as English provides comparatively fewer local reordering cues for generating the target-side Pnar word order. Consequently, the benefits of lexicalized reordering are more pronounced for Pnar→English translation than for English→Pnar translation.

5.3 Effect of MERT Tuning

MERT tuning (Config-2) degrades BLEU in both translation directions (-2.50 for Pnar→English and -3.23 for English→Pnar) while improving chrF2 marginally. The tuned models exhibit output length ratios exceeding 1.0 (over-generation) suggesting the MERT weight optimisation converges to degenerate solutions under the small corpus setting. This is consistent with known limitations of MERT under low-resource conditions [1], where the development set is insufficient to reliably estimate 14 feature weights. Despite degraded BLEU, the higher chrF2 for MERT-tuned systems suggests better character-level coverage, warranting further investigation. The chrF2/BLEU divergence under MERT tuning is itself informative: chrF2 rewards character-n-gram overlap and is comparatively robust to the length and word-order errors that BLEU penalizes heavily, so an over-generating system can gain chrF2 while losing BLEU. This suggests that MERT, optimizing directly against BLEU on a 300-sentence dev set with 14 free feature weights, overfits to length statistics of that small set rather than to genuine translation adequacy. In practice, this means BLEU-tuned weights are not a reliable model-selection criterion at this corpus size; a held-out chrF2 or human-adequacy check would be a safer stopping criterion for future low-resource SMT work.

5.4 Statistical Significance

To assess whether the differences reported in Table 3 reflect genuine system differences rather than test-set sampling variance, we apply paired bootstrap resampling [6] with 1,000 resamples over the full 371-sentence test set in each translation direction, using SacreBLEU (detokenized, `tok:13a`) to compute corpus-level BLEU on each resample. For Pnar→English, the reordering gain (Config-1 vs. Config-3) is +3.72 BLEU (95% CI $[3.16, 4.30]$, $p < 0.001$), confirming that lexicalized reordering yields a genuine, statistically significant improvement rather than an artefact of test-set sampling. MERT tuning (Config-2 vs. Config-1) produces a significant *decrease* of -2.50 BLEU (95% CI $[-3.24, -1.75]$, $p < 0.001$), confirming that the degradation reported in Section 5.3 is a robust effect rather than sampling noise.

Table 5. Paired bootstrap significance test (B=1,000) on the 371-sentence test set. † = significant at $p < 0.05$.

Direction	Comparison	Δ BLEU	95% CI	p-value	Sig.?
Pn→En	Config-1 vs. Config-3 (reordering)	+3.72	[3.16, 4.30]	< 0.001	Yes†
Pn→En	Config-1 vs. Config-2 (MERT)	−2.50	[−3.24, −1.75]	< 0.001	Yes†
En→Pn	Config-1 vs. Config-3 (reordering)	−1.02	[−2.19, 0.11]	0.084	No
En→Pn	Config-1 vs. Config-2 (MERT)	−3.22	[−3.96, −2.43]	< 0.001	Yes†

For English→Pnar, the small BLEU decline attributed to reordering (Config-1 vs. Config-3, −1.02 BLEU) is *not* statistically significant (95% CI $[-2.19, 0.11]$, $p = 0.084$): the confidence interval crosses zero, so we cannot reject the possibility of no true difference at this test-set size. This supports the qualitative explanation in Section 4.3 that the apparent advantage of the no-reorder system is likely an artefact of named-entity copying rather than a genuine translation-quality difference. In contrast, MERT tuning's degradation (Config-2 vs. Config-1, −3.22 BLEU) is highly significant (95% CI $[-3.96, -2.43]$, $p < 0.001$), showing that MERT's harm under this low-resource setting is a consistent effect across both translation directions.

5.5 Sample Translations and Error Analysis

Tables 6 and 7 present representative translation examples from the best-performing configurations.

Error analysis identifies four primary failure modes. First, **out-of-vocabulary (OOV) tokens**: Pnar's agglutinative morphology generates many surface forms unseen in the 9,563 sentence training corpus; OOV words are passed through untranslated, accounting for residual Pnar tokens visible in the system output. Second, **long-distance reordering**: although the lexicalized reordering model captures local phrase-level reordering, complex verb-final Pnar constructions requiring long-distance movement remain challenging [7]. Third, **code-mixing with Khasi**: Pnar newspaper text frequently borrows Khasi lexical items not present in the English translation, confusing the phrase extractor a phenomenon also observed in Khasi-language MT by Singh and Hujon [12], who report similar OOV and domain-mismatch challenges for the closely related Khasian language Khasi under comparably constrained parallel data. Fourth, **morphological complexity**: Pnar verbal morphology encodes tense, aspect, and modality through affixation, generating paradigm forms that exceed phrase-table coverage [12].

5.6 Comparison with Related Systems

Comparison with SMT baselines reported for similarly low-resource Austroasiatic language pairs is instructive. Work on Khasi–English SMT which benefits

Table 6. Sample translations from best Pnar→English system (Config-1, BLEU 14.97). System output is from the untuned, reordering-enabled configuration (Config-1); word-for-word artefacts are discussed in 5.5

Type	**Sentence**
Source 1	Sdang ka JAC yei yakhih wow khang yei lai i wan ki Truck ka Star Cement wa Dalmia.
Reference	The JAC began a protest to stop the movement of trucks belonging to Star Cement and Dalmia Cement.
System	The JAC yakhih to the khang the on the wan Truck Star Cement along with the Dalmia.
Source 2	Mynsoo joor ki ar ngut hadien wa noh ka Truck chapoh riat ha Ri-Bhoi.
Reference	Two people were seriously injured after a truck fell into a gorge in Ri-Bhoi.
System	Two the ar ngut after the noh Truck chapoh riat in Ri-Bhoi.
Source 3	Daw pynman ka WJDBA ya ka Inter School Badminton Competition 2026.
Reference	WJDBA will organize the Inter-School Badminton Competition 2026.
System	The WJDBA of the Inter School Badminton Competition 2026.

Table 7. Sample translations from best English→Pnar system (Config-1, BLEU 10.14).System output is from the untuned, reordering-enabled configuration (Config-1); word-for-word artefacts are discussed in 5.5

Type	**Sentence**
Source 1	The JAC began a protest to stop the movement of trucks belonging to Star Cement and Dalmia Cement.
Reference	Sdang ka JAC yei yakhih wow khang yei lai i wan ki Truck ka Star Cement wa Dalmia.
System	Ya ka JAC began protest toh ka movement trucks belonging Star Cement wa u Dalmia Cement.
Source 2	Two people were seriously injured after a truck fell into a gorge in Ri-Bhoi.
Reference	Mynsoo joor ki ar ngut hadien wa noh ka Truck chapoh riat ha Ri-Bhoi.
System	Yap seriously injured hadien ka truck fell into gorge wa Ri-Bhoi.
Source 3	WJDBA will organize the Inter-School Badminton Competition 2026.
Reference	Daw pynman ka WJDBA ya ka Inter School Badminton Competition 2026.
System	WJDBA da organize ka Inter-School Badminton Competition 2026.

from a slightly larger NLP research community has reported BLEU scores in the range of 8–16 using comparable corpus sizes. Our best Pnar→English result of BLEU score 14.97 falls within this range suggesting that the corpus quality and SMT configuration are reasonable for this data scale. The lower En→Pn score (10.14) is consistent with the general difficulty of translating into morphologically richer target languages with limited parallel data.

6 Conclusion and Future Work

We present the first machine translation study for the English–Pnar language pair, an Austroasiatic language for which no prior computational MT research has been reported. The main contributions of this work are threefold. First, we construct a parallel corpus comprising 9,563 sentence pairs using the pnar sentences from *Wyrta* local newspaper archive. Second, we perform a systematic comparison of six phrase-based statistical machine translation (SMT) configurations, isolating the effects of lexicalized reordering and minimum error rate training (MERT). Third, we establish the first published baseline for English–Pnar translation using BLEU, chrF2, and TER, with the best-performing systems achieving BLEU scores of 14.97 for Pnar→English and 11.16 for English→Pnar. Our experimental results demonstrate that lexicalized reordering provides a substantial improvement of 3.73 BLEU points for Pnar→English translation, highlighting the importance of modeling the structural transformation from Pnar's SOV word order to English SVO. In contrast, MERT tuning consistently degrades BLEU performance under the current data scale suggesting that it is prone to overfitting when only a small development set are available. This finding provides an important practical guideline for future low-resource SMT research. The error analysis further identifies four major sources of translation errors: morphological out-of-vocabulary (OOV) words, long-distance reordering, and Khasi code-mixing. As immediate future work, we plan to (i) develop a rule-based Pnar affix segmentation system to reduce the OOV rate with less dataset collection; (ii) expand the parallel corpus through community-assisted translation of the existing *Wyrta* newspaper archive; and (iii) investigate neural machine translation models and the pre-trained multilingual models such as mBART and mT5. The SMT baseline established in this study provides a meaningful benchmark against which future neural and multilingual approaches can be evaluated.

References

1. Araabi, A., Monz, C.: Optimizing transformer for low-resource neural machine translation. In: Proceedings of the 28th International Conference on Computational Linguistics (COLING). pp. 3429–3435. International Committee on Computational Linguistics, Barcelona, Spain (Online) (2020). https://doi.org/10.18653/v1/2020.coling-main.304
2. Bareh, C.: Phonological correspondences between jowai and narwan-pnar. Mon-Khmer Studies: The Journal of Austroasiatic Languages and Cultures **45**, 1–13

(2026), http://mksjournal.org, copyright for this paper vested in the author. Released under Creative Commons Attribution License.
3. Bird, S.: Decolonising speech and language technology. Proceedings of the 28th International Conference on Computational Linguistics pp. 3504–3519 (2020). https://doi.org/10.18653/v1/2020.coling-main.313
4. Gao, Q., Vogel, S.: Parallel implementations of word alignment tool. In: Cohen, K.B., Carpenter, B. (eds.) Software Engineering, Testing, and Quality Assurance for Natural Language Processing. pp. 49–57. Association for Computational Linguistics, Columbus, Ohio (Jun 2008), https://aclanthology.org/W08-0509/
5. Heafield, K.: KenLM: Faster and smaller language model queries. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 187–197. Association for Computational Linguistics, Edinburgh, Scotland (Jul 2011), https://www.aclweb.org/anthology/W11-2123
6. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT Summit (2005)
7. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: Ananiadou, S. (ed.) Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. pp. 177–180. Association for Computational Linguistics, Prague, Czech Republic (Jun 2007), https://aclanthology.org/P07-2045/
8. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. pp. 127–133 (2003), https://aclanthology.org/N03-1017/
9. Och, F.J.: Minimum error rate training in statistical machine translation. In: ACL (2003)
10. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1), 19–51 (2003). https://doi.org/10.1162/089120103321337421, https://aclanthology.org/J03-1002/
11. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (2002). https://doi.org/10.3115/1073083.1073135
12. Singh, T.D., Vellintihun Hujon, A.: Low resource and domain specific english to khasi smt and nmt systems. In: 2020 International Conference on Computational Performance Evaluation (ComPE). pp. 733–737 (2020). https://doi.org/10.1109/ComPE49325.2020.9200059
13. Tillmann, C.: A unigram orientation model for statistical machine translation. In: HLT-NAACL (2004)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 5998–6008. Curran Associates, Inc. (2017)