Why should only High-Resource-Languages have all the fun? Pivot Based Evaluation in Low Resource Setting

Ananya Mukherjee*, **Saumitra Yadav***, **Manish Shrivastava**
MT-NLP Lab, LTRC, KCIS, IIIT Hyderabad, India
ananya.mukherjee@research.iiit.ac.in
saumitra.yadav@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

Evaluating machine translation (MT) systems for low-resource languages has long been a challenge due to the limited availability of evaluation metrics and resources. As a result, researchers in this space have relied primarily on lexical-based metrics like BLEU, TER, and ChrF, which lack semantic evaluation. In this first-of-its-kind work, we propose a novel pivot-based evaluation framework that addresses these limitations; after translating low-resource language outputs into a related high-resource language, we leverage advanced neural and embedding-based metrics for more meaningful evaluation. Through a series of experiments using five low-resource languages: Assamese, Manipuri, Kannada, Bhojpuri, and Nepali, we demonstrate how this method extends the coverage of both lexical-based and embedding-based metrics, even for languages not directly supported by advanced metrics. Our results show that the differences between direct and pivot-based evaluation scores are minimal, demonstrating that this approach is a viable and effective solution for evaluating translations in endangered and low-resource languages. This work paves the way for more inclusive, accurate, and scalable MT evaluation for under-represented languages, marking a significant step forward in this under-explored area of research. The code and data will be made available at https://github.com/AnanyaCoder/PivotBasedEvaluation.

1 Introduction

Machine translation (MT) has made significant progress in recent years, particularly for high-resource languages, where abundant data and sophisticated evaluation metrics have driven improvements in translation quality. Neural-based metrics (Rei et al., 2020) and embedding-based (Mukherjee et al., 2020; Mukherjee and Shrivastava, 2023; Zhang et al., 2019; Feng et al., 2022; Artetxe and Schwenk, 2019; Kakwani et al., 2020; Khanuja et al., 2021) methods have enabled more accurate assessments, capturing meaning, fluency, and context beyond simple word matches. These advanced metrics offer a deeper understanding of translation quality, helping refine MT models for high-resource languages.

Language	Language Family	Resource Presence	Morphological Complexity	No. of Speakers
Kannada (kn)	Dravidian	Low	High	44M
Bhojpuri (bj)	Indo-Aryan	Very low	Moderate	50M
Nepali (np)	Indo-Aryan	Low	Moderate	32M
Manipuri (mn)	Sino-Tibetan	Very low	High	1.7M
Assamese (as)	Indo-Aryan	Low	Moderate	15M

Table 1: Details of the low-resource languages considered in our experiment with different levels of resources and morphological complexity. The number of speakers is obtained from Ethnologue (eth, 2023).

However, low-resource languages face substantial challenges (Haddow et al., 2022). With limited parallel data, sparse high-quality references, and inadequate embeddings, the evaluation of MT systems for these languages continues to rely heavily on lexical-based metrics such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006; Post, 2018), and ChrF (Popović, 2017). While these metrics provide a basic measure of translation accuracy, they fall short in capturing the full complexity of language, often overlooking essential aspects like semantics, syntactic structure, and overall fluency. As a result, there remains a significant gap in the ability to accurately evaluate and improve MT systems for low-resource languages. Despite significant progress in low-resource machine translation, as seen in efforts like WAT (Workshop on Asian Translation) (Nakazawa et al., 2023) and LoResMT (Pal et al., 2023), evaluation methods have not advanced at the same pace, resulting in limited capabilities to assess translation quality effectively.

This lack of resources in evaluation methods cre-

*Authors contributed equally

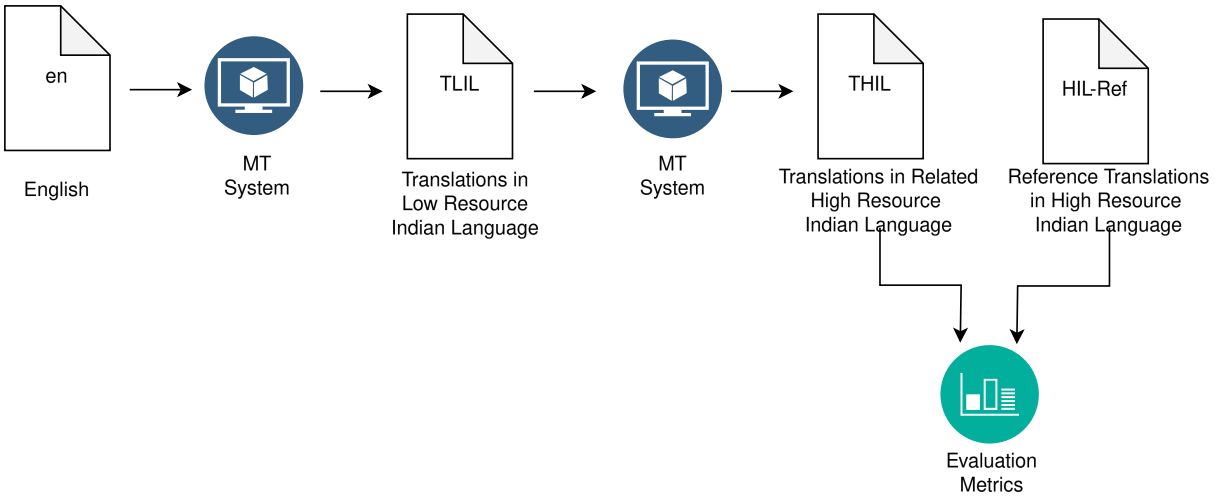


Figure 1: Pivot Based Evaluation Approach. IL:Indian Language, TLIL: Translated Low-resource IL, THIL:Translated High-resource IL, HIL-Ref: High-resource IL References.

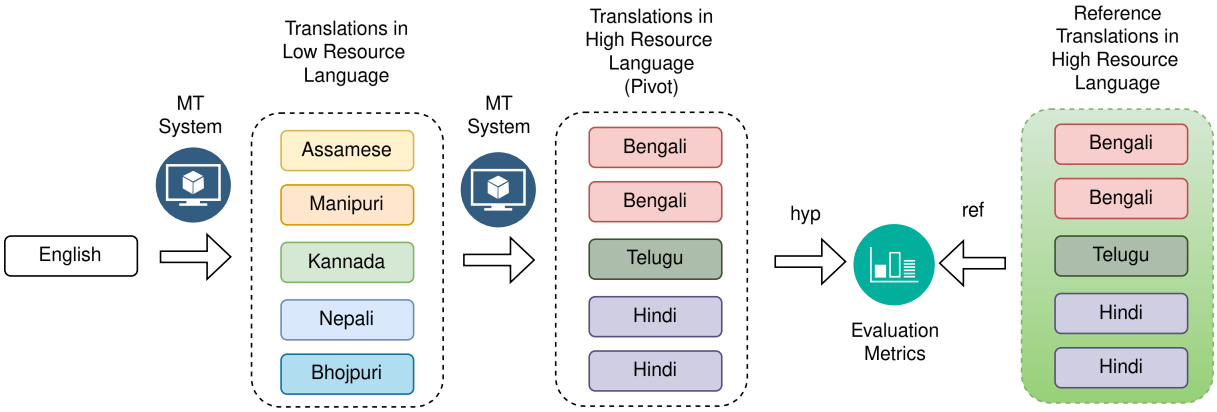


Figure 2: Pivot Based Evaluation Approach with languages used in our experiments

ates a major bottleneck in assessing the actual performance of low-resource MT systems. To address this gap, we propose a novel pivot-based evaluation framework, inspired by the success of pivot-based translation methods, which use an intermediary language as a pivot to improve translation performance (Kim et al., 2019; Mhaskar and Bhattacharyya, 2021). In the context of machine translation systems pivoting refers to a set of techniques where a pivot language is used to facilitate translation between a source and target language (Mhaskar and Bhattacharyya, 2024). By leveraging the linguistic resources of the pivot language, the performance of the source-to-target machine translation model has been significantly improved (Mhaskar and Bhattacharyya, 2022; Gadugoila et al., 2022; Kunchukuttan and Bhattacharyya, 2020). We build on this concept by using a high-resource language as a pivot to evaluate related low-resource languages. This pivoting approach leverages high-resource language resources for a more thorough and precise evaluation of low-resource language translation quality.

Our approach involves translating outputs from these low-resource languages into closely related high-resource languages. And our approach enables the use of advanced neural and embedding-based metrics to provide a more accurate and nuanced assessment of translation quality. In our work, we focus on evaluating translations of languages across three diverse families: **Dravidian** (Kannada), **Indo-Aryan** (Assamese, Nepali, Bhojpuri), and **Sino-Tibetan** (Manipuri). Table 1 details the level of resource availability, morphological complexity (Grierson, 1903-1928; Steever, 1998) and number of speakers for these low-resource languages (eth, 2023).

Our work is the first to introduce pivot-based evaluation in this context, offering a new pathway for research and development in low-resource machine translation. We demonstrate that this approach not only bridges the evaluation gap but also

has the potential to significantly improve the quality of machine translation systems for low-resource languages and endangered languages.

2 Motivation and Objective

In recent years, significant progress has been made in machine translation (MT) for low-resource languages, primarily due to word segmentation (Chang et al., 2008; Sennrich et al., 2016b; Shao et al., 2018), data augmentation (Sennrich et al., 2016a; Li et al., 2019), pivot translation (Mhaskar and Bhattacharyya, 2021), transfer learning (Zoph et al., 2016; Lakew et al., 2018), multilingual models (Devlin et al., 2018; Khanuja et al., 2021; Kakwani et al., 2020), and pre-training techniques (He et al., 2023; Baziotis et al., 2021) with transformer-based architectures (Conneau et al., 2020; Vaswani, 2017). These innovations have allowed for improved translation quality, even when only limited parallel data is available. Low-resource MT systems have also benefited from techniques like zero-shot learning (Romera-Paredes and Torr, 2015) and cross-lingual transfer (Chen et al., 2018), enabling models to leverage data from high-resource languages to improve performance (Mhaskar and Bhattacharyya, 2022; Gadugoila et al., 2022; Kunchukuttan and Bhattacharyya, 2020). Despite these advances, however, depending on resources available for target languages the comprehensive evaluation of low-resource MT is still lagging. Current evaluation methods in these situations remain highly dependent on lexical-based metrics which, although widely used, offer a shallow view of translation quality, focusing primarily on word or character overlaps rather than semantic meaning or fluency. There is an urgent need for innovative methods that can bridge this gap and enable more sophisticated evaluation frameworks for these languages.

The goal of our work is to demonstrate that pivot-based evaluation can be an effective solution for assessing the quality of machine translations for low-resource languages, even in the absence of direct metric support or robust references. By leveraging a related high-resource language as a pivot, low-resource language translations can be evaluated by advanced metrics such as BERTScore (Zhang et al., 2019), LaBSE (Feng et al., 2022), MEE4 (Mukherjee and Shrivastava, 2023), IndicBERT (Kakwani et al., 2020), COMET (Rei et al., 2020) etc., which may not be available for the low-resource languages.

This approach also has broader implications: **it can be extended to endangered and other low-resource languages, ensuring that they are not left behind in the development and evaluation of machine translation systems**. By using a linguistically or geographically related pivot language[1], we can evaluate translations more comprehensively, facilitating the preservation and support of endangered languages through more accurate translation systems.

3 Method

Figure 1 illustrates the process of our proposed approach i.e., '*the translated sentences in low resource language*' (TLIL) are translated further to '*related high resource language*' (THIL) which are then evaluated using reference sentences of high resource language (HIL-Ref). To provide a clearer picture, figure 2 depicts the low-resource languages considered in our experiment and the corresponding high-resource languages used as a pivot. To select the related high-resource languages for pivot-based evaluation, we referred to linguistic proximity and geographic adjacency (Steever, 1998; Grierson, 1903-1928). Languages within the same language family typically share a close relationship, as they have evolved from a common ancestor. Table 2 provides a detailed overview of the relationships between low-resource languages and their respective high-resource counterparts considered in our experiments. The linguistic proximity between Kannada and Telugu, Assamese and Manipuri with Bengali, acts as a supporting evidence for our language-pair selections.

4 Experimental Set up

We conducted three distinct experiments to evaluate the effectiveness of pivot-based evaluation for low-resource languages, using a variety of language pairs and translation directions. Below is a detailed description of each experiment:

- **Experiment 1: English to Low-Resource Indic Languages (En → TLIL)**:
  In this experiment, we translated English source sentences into five low-resource Indic languages: Assamese, Manipuri, Kannada, Bhojpuri, and Nepali. The goal was to assess translations in low-resource languages. The

[1] high-resource pivot language

Language	Kannada	Telugu	Bhojpuri	Hindi	Nepali	Hindi	Assamese	Bengali	Manipuri	Bengali
Language Family	Dravidian	Dravidian	Indo-Aryan	Indo-Aryan	Indo-Aryan	Indo-Aryan	Indo-Aryan	Indo-Aryan	Sino-Tibetan	Indo-Aryan
Script	Kannada Script	Telugu Script	Devanagari	Devanagari	Devanagari	Devanagari	Bengali Script	Bengali Script	Meitei Mayek	Bengali script
Vocabulary Similarity	High, due to shared Dravidian roots		Very high, Bhojpuri is a Hindi-dialect		Medium to High, both languages share Sanskrit roots		High, lexical borrowings, shared Sanskritic origin		Medium, due to geographical proximity and shared vocabulary	
Phonological Features	Similar vowel and consonant systems, intonation patterns		Almost identical phonology, mutual intelligibility		Similar phonological structure (retroflex consonants, vowels), almost mutually intelligible		Similar phonetic systems, nasalization of vowels		Some shared features due to contact, but distinct phonologies	
Geographical Proximity	Southern India, neighboring states (Karnataka and Andhra Pradesh)		North-Central India (Bihar and UP)		Neighboring countries (Nepal and India), continuous historical and cultural exchange		Eastern India (Assam and Bengal are neighboring states)		Northeast India (Manipuri and Bengali speakers live in close proximity in the region)	

Table 2: Proximity of Languages

evaluation of these translations was carried out using both lexical and embedding-based metrics. However, it is important to note that not all five languages are supported by advanced metrics. For instance, BERTScore does not support Assamese and Bhojpuri, while COMET is unavailable for Manipuri and Bhojpuri. In Table 3, unsupported languages are indicated with dashes.

- **Experiment 2: Translate Low-Resource Translated Outputs to High-Resource Indic Languages (TLIL → THIL_1):**
  This experiment involved translating the outputs from the five low-resource Indic languages generated in Experiment 1 into high-resource Indic languages, specifically Bengali, Telugu, and Hindi. By translating into these pivot languages, we aimed to evaluate the quality of translations from low-resource languages using metrics that are more readily applicable to high-resource languages.

- **Experiment 3: Translate Low-Resource FLORES Sentences to High-Resource Languages (LIL → THIL_2):**
  In the final experiment, we translated the sentences released by FLORES devtest (see section 4.1) of the same five low-resource Indic languages[2] into their respective related high-resource languages. This experiment aimed to further evaluate the performance of machine translation systems when translating from low-resource to high-resource languages, providing insights into the effectiveness of pivot-based evaluation in these language directions.

[2]Assamese, Manipuri, Kannada, Bhojpuri, and Nepali

4.1 Test Data

For our experiments, we used the 'devtest' set from the latest FLORES-200 dataset[3] (NLLB Team et al., 2022), a multilingual benchmark designed to evaluate translation quality across diverse languages. This devtest contains 1,012 standardized parallel sentences, enabling consistent evaluation of translation quality across multiple language pairs.

4.2 MT System

We conducted three translation experiments using IndicTrans2 (Gala et al., 2023), an open-source transformer-based multilingual NMT model that supports high-quality translations across all the 22 scheduled Indic languages. In the first experiment, we translated English source sentences into five low-resource languages: Assamese, Manipuri, Kannada, Bhojpuri, and Nepali. Next, we translated these outputs into three high-resource pivot languages—Bengali, Telugu, and Hindi. To further assess the performance of IndicTrans2 in inter-language (IL-IL) translation directions, we also translated FLORES data from these five low-resource languages into their respective related high-resource languages. The evaluation scores for these experiments are presented in Table 3.

4.3 Automatic Evaluation Metrics

For the automatic evaluation of translation quality, we employed a combination of lexical-based, embedding-based, and supervised neural metrics:

- **Lexical Based Metrics:** We used traditional lexical overlap metrics such as BLEU[4] (Papineni et al., 2002), ChrF++[5] (Popović, 2017), and TER (Snover et al., 2006; Post, 2018).

[3]https://github.com/facebookresearch/flores/blob/main/flores200/README.md.

[4]Signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2

[5]Signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2

		Direct Evaluation					Pivot-Based Evaluation					Direct Evaluation				
		Flores English → TLIL					**TLIL → THIL_1**					**Flores LIL → THIL_2**				
	Metrics	**en-as**	**en-mn**	**en-kn**	**en-np**	**en-bj**	**as-bn**	**mn-bn**	**kn-te**	**np-hi**	**bj-hi**	**as-bn**	**mn-bn**	**kn-te**	**np-hi**	**bj-hi**
Lexical	BLEU	0.099	0.071	0.227	0.232	0.081	0.173	0.113	0.205	0.304	0.286	0.119	0.084	0.158	0.256	0.140
-based	ChrF++	0.446	0.424	0.601	0.607	0.367	0.523	0.434	0.579	0.565	0.567	0.448	0.395	0.517	0.524	0.423
Metrics	TER	0.795	0.920	0.656	0.598	0.088	0.696	0.833	0.652	0.555	0.614	0.784	0.859	0.736	0.617	0.423
	BERTScore	-	0.825	0.875	0.877	-	0.858	0.820	0.863	0.876	0.869	0.829	0.805	0.840	0.861	0.814
	LaBSE	0.842	-	0.920	0.936	-	0.909	0.852	0.917	0.931	0.905	0.866	0.825	0.889	0.912	0.821
Embedding	LASER	-	-	-	-	-	0.881	0.854	0.866	0.887	0.832	0.908	0.871	0.885	0.902	0.889
-based	MEE4	0.705	-	0.797	0.814	-	0.784	0.723	0.791	0.827	0.808	0.737	0.695	0.758	0.806	0.720
Metrics	mBERT	-	-	0.872	0.925	-	0.859	0.838	0.860	0.874	0.867	0.844	0.823	0.845	0.864	0.797
	IndicBERT	0.954	-	0.958	-	-	0.957	0.941	0.953	0.956	0.942	0.950	0.944	0.947	0.952	0.935
	MURIL	0.999	-	1.000	1.000	-	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
Supervised Metric	COMET	0.837	-	0.873	0.845	-	0.862	0.802	0.865	0.798	0.786	0.829	0.791	0.849	0.778	0.672

Table 3: Translation Quality scores of the MT outputs for 1) English devtest to Low Resource Languages (TLIL) 2) Low Resource Translations (TLIL) to High Resource Languages (THIL_1) and 3) Low Resource Flores devtest (LIL) to High Resource Translations (THIL_2). Dash (-) indicates that the metric does not support the corresponding language. Metric scores are normalized between 0-1.

These metrics assess translation accuracy by comparing the predicted translations to reference translations based on word-level and char-level matches.

- **Embedding Based Metrics:** To capture semantic similarities beyond lexical overlap, we utilized several embedding-based metrics, including BERTScore (Zhang et al., 2019), LaBSE (Feng et al., 2022), LASER (Artetxe and Schwenk, 2019), MEE4 (Mukherjee and Shrivastava, 2023), mBERT (Devlin et al., 2018), IndicBERT (Kakwani et al., 2020), and MURIL (Khanuja et al., 2021). These metrics compute sentence-level embeddings and evaluate translation quality by measuring the closeness of the embeddings between the hypothesis and reference sentences.

- **Supervised Neural Metrics:** We also employed COMET (Rei et al., 2020), a state-of-the-art supervised neural metric. COMET leverages pre-trained neural models and fine-tunes a human-annotated data, providing a more robust evaluation by predicting human judgment scores directly.

5 Results and Analysis

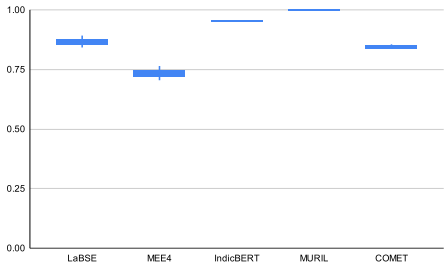
Table 3 presents the results of different evaluation metrics applied to three experiments (see section 4). The metrics are categorized into lexical-based (e.g., BLEU, chrF++, TER), embedding-based (e.g., BERTScore, LaBSE, LASER, MEE4, mBERT, IndicBERT, MURIL), and a supervised neural metric (COMET). These evaluations cover translations from English to five low-resource Indic languages (TLIL): Assamese (as), Manipuri (mn), Kannada (kn), Nepali (np), and Bhojpuri (bj), and translations between these low-resource languages and their high-resource counterparts (Bengali, Telugu and Hindi).

The key analyses of the results, highlighting important observations, are presented below:

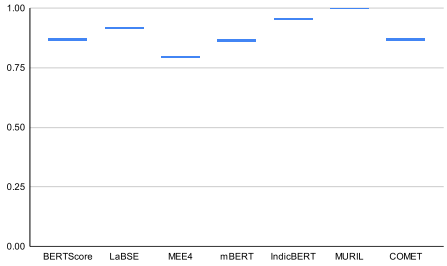
- **Limited Metric Support for Low-Resource Languages by Advanced Metrics**
  The first section of the table, focusing on the English → TLIL translation task, highlights a key challenge: *many advanced metrics do not support all five low-resource languages*. For instance, BERTScore does not cover Assamese and Bhojpuri, while COMET lacks support for Manipuri and Bhojpuri, IndicBERT does not support Manipuri, Nepali and Bhojpuri, LASER doesn't support any of the five languages.

- **Ensuring No Language Was Left Behind Through Pivot-Based Evaluation**
  Despite certain languages lacking support for some advanced metrics, they are not excluded from evaluation. The second section of the table (TLIL → THIL_1) demonstrates how pivot-based evaluation mitigates this issue. By translating the low-resource languages into higher-resource pivot languages (e.g., Bengali, Telugu, Hindi), we ensured comprehensive evaluation across all metrics. *This method provided complete metric coverage, allowing for a fair assessment of translations even when direct metric support was unavailable for certain low-resource languages.*

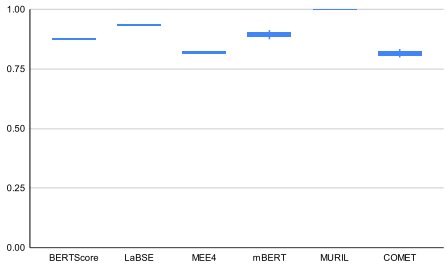
- **IndicTrans2 Performance in LIL → THIL Translation**



(a) Performance differences between en-as and as-bn



(b) Performance differences between en-kn and kn-te



(c) Performance differences between en-np and np-hi

Figure 3: Plots highlighting subtle performance variations between En-TLIL and TLIL-THIL_1

The third section (FLORES LIL → THIL_2) highlights the effectiveness of IndicTrans2 as a robust machine translation system. The high scores across various metrics underscore its ability to produce high-quality translations for low-resource languages. *This reliable performance instils confidence in using the IndicTrans2 system for demonstrating its ability to handle complex translation tasks and supporting its ongoing use and development for low-resource languages.*

- **Minimal Difference Between Direct and Pivot-Based Evaluations**
  Another key finding from the experiments is that the difference between direct evaluation scores (en → TLIL) and those obtained after pivoting (TLIL → THIL_1) is minimal for the low-resource languages that are supported by advanced metrics. Figure 3 presents a candlestick graph, showing the height variations in evaluation scores for Assamese, Kannada, and Nepali when comparing their direct translations and the pivot-based evaluations with Bengali, Telugu, and Hindi. Across all the metrics, the height of each candlestick represents the range of scores for both direct translations and pivot-based evaluations. For Assamese, the short height of the candle indicate a minor deviation between direct translation scores and those obtained through Bengali as the pivot language. Similarly, Kannada shows very minute differences when evaluated directly or through Telugu as the pivot. Nepali translations, evaluated both directly and via Hindi as the pivot, also exhibit comparable score ranges, as indicated by the subtle shifts in candle heights. *This consistent pattern exhibited by all the advanced metrics across these languages reinforces the reliability of pivot-based evaluation, further supporting the robustness of the approach in low-resource language settings.*

Figure 4 clearly depicts our approach where Telugu, a high-resource language, acts as a pivot in assessing translations from English to Kannada, a low-resource language. The translation from English to Kannada has an error in which the word "mice" is mistranslated into "dogs". This error is also seen in the Telugu translation (see pivot example) when Kannada is used as the source language for the pivot assessment. By analyzing the Telugu translation, metrics equipped to handle Telugu can indirectly gauge the quality of the Kannada translation, uncovering the error.

This highlights that if there is any error when translating from English to low-resource, that error would be propagated when translating from low to pivot language, thus showing the utility of using pivot-based evaluation to detect translation errors in low-resource languages.

Overall, our experiments demonstrate that pivot-based evaluation substantially improves the ability to assess translations, even for languages not directly supported by certain metrics.

6 Challenges

A key challenge is **identifying an appropriate pivot language**—a high-resource language that is linguistically or geographically related to the target low-resource language. **Access to comprehen-**

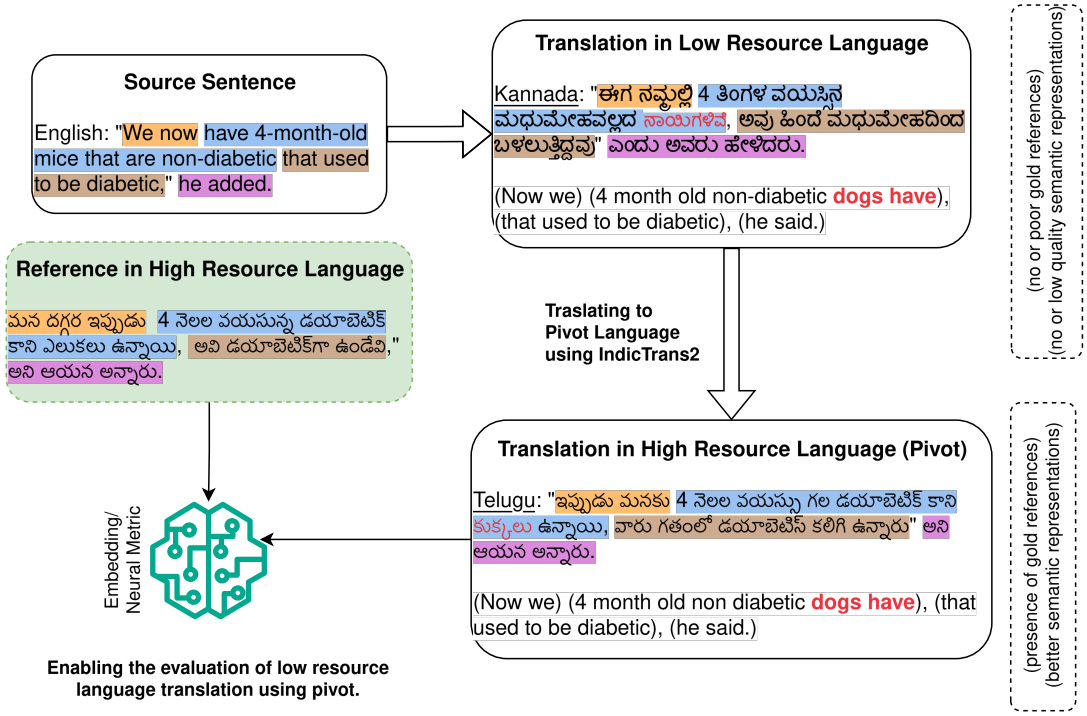


Figure 4: Example of using High Resource Language (Telugu) as pivot for English to Low Resource Language (Kannada) Translation Evaluation. The Kannada translation contains an error (highlighted in red), substituting '**dogs**' for '**mice**'. It is evident that this error is also carried in translations from the low-resource source language to the pivot translation.

sive and well-curated datasets for low-resource languages was another challenge. The availability of parallel corpora and high-quality reference translations in these languages is limited, making thorough evaluations difficult to conduct across a broad spectrum of languages and domains.

7 Future Work

Future work will focus on several key areas to advance the pivot-based evaluation approach. Expanding the methodology to **include a wider array of low-resource and endangered languages** will be crucial, providing a more comprehensive evaluation across diverse linguistic contexts. **Refining the process for selecting pivot languages**, possibly through the development of algorithms that account for linguistic and contextual factors, could enhance the precision and applicability of the evaluations. **Real-world testing** in various practical settings will be essential to validate the approach and demonstrate its effectiveness in operational environments. Lastly, exploring alternative evaluation strategies or **hybrid models that combine pivot-based evaluation with other methods** may yield new insights and improvements in translation quality assessment.

8 Conclusion

In an era where machine translation (MT) systems are becoming increasingly essential for global communication, particularly for underrepresented languages, this paper introduces a groundbreaking solution—pivot-based evaluation—to tackle the long-standing challenge of assessing translations in low-resource languages. In this paper, we explored pivot-based evaluation as a novel method for assessing translations in low-resource languages, where direct evaluation metrics are often unavailable. Our goal was to extend the reach of advanced translation evaluation methods by leveraging linguistically or geographically related high-resource languages as pivots.

We conducted a series of experiments across three translation scenarios using the FLORES devtest dataset. First, we evaluated translations from English to five low-resource Indic languages (Assamese, Manipuri, Kannada, Bhojpuri, Nepali). Second, we applied pivot-based translations from these low-resource languages into related high-resource languages (Bengali, Telugu, Hindi), which allowed us to assess the quality of translations using all available metrics. Finally, we compared the

performance of translations between low-resource languages and high-resource ones to further validate the robustness of the method.

Our results demonstrate that the pivot-based evaluation technique provides reliable assessments with minimal discrepancies between direct and pivot-based translation evaluations. The high performance of IndicTrans2 across multiple language pairs confirmed its effectiveness as an MT system, particularly in low-resource settings. Furthermore, the minor differences in evaluation scores when using pivot languages underscore that this method can be effectively **extended to other endangered and underrepresented languages**, supporting ongoing efforts to **enhance translation evaluation in resource-constrained settings**.

In conclusion, our study proves that pivot-based evaluation can fill a critical gap in translation quality assessment for low-resource languages, making it an impactful tool for future research and practical applications in multilingual and underrepresented language contexts.

9 Limitations

While pivot-based evaluation provides a valuable approach for assessing translations in low-resource languages, it does come with several limitations. The most significant is the **performance of low-resource to high-resource (pivot) MT systems**: if the translations into the pivot language are inaccurate, the resulting evaluation scores can be misleading. Another limitation is the **additional step in the evaluation process**, which increases computational resources and time due to the extra layer of translation into the pivot language. Lastly, there is the potential for **loss of linguistic nuance**: unique elements of the source language may be altered or lost during the pivot translation, impacting the accuracy of the final evaluation.

Despite the limitations, pivot-based evaluation offers a meaningful alternative where direct evaluation methods fall short. It allows for a more informed assessment than lexical-based metrics alone, making it an important stepping stone toward more comprehensive and inclusive translation evaluation practices.

10 Ethical Considerations

This research exclusively utilizes publicly available resources, including the IndicTrans2 model and the FLORES dataset, ensuring transparency and ethical compliance throughout the study. Additionally, the evaluation of translations was performed using publicly available metrics, including both lexical-based and embedding-based methods. The goal of this work is to advance the evaluation of low-resource language translations without infringing on privacy or data ownership rights.

Acknowledgement

We sincerely thank Ashok Urlana for his iterative feedback, which was crucial to this work's progress. We also acknowledge AI4Bharat[6] and Meta AI[7] for the IndicTrans2 machine translation systems and FLORES parallel data, which greatly enriched our research quality and reliability.

References

2023. Ethnologue: Languages of the world.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Christos Baziotis, Ivan Titov, Alexandra Birch, and Barry Haddow. 2021. Exploring unsupervised pretraining objectives for machine translation. *Preprint*, arXiv:2106.05634.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing Chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232, Columbus, Ohio. Association for Computational Linguistics.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2018. Multi-source cross-lingual model transfer: Learning what to share. *arXiv preprint arXiv:1810.03552*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[6] https://ai4bharat.iitm.ac.in/

[7] https://ai.meta.com/research/no-language-left-behind/

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Hemalatha Gadugoila, Shailashree K Sheshadri, Priyanka C Nair, and Deepa Gupta. 2022. Unsupervised pivot-based neural machine translation for english to kannada. In *2022 IEEE 19th India Council International Conference (INDICON)*, pages 1–6.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

George A. Grierson. 1903-1928. *Linguistic Survey of India*. Government of India.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Zexue He, Graeme Blackwood, Rameswar Panda, Julian McAuley, and Rogerio Feris. 2023. Synthetic pre-training tasks for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8080–8098, Toronto, Canada. Association for Computational Linguistics.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Yunsu Kim, Petre M. Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-english languages. In *Conference on Empirical Methods in Natural Language Processing*.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2020. Utilizing language relatedness to improve machine translation: A case study on languages of the indian subcontinent. *Preprint*, arXiv:2003.08925.

Surafel M. Lakew, Aliia Erofeeva, Matteo Negri, Marcello Federico, and Marco Turchi. 2018. Transfer learning in multilingual neural machine translation with dynamic vocabulary. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 54–61, Brussels. International Conference on Spoken Language Translation.

Guanlin Li, Lemao Liu, Guoping Huang, Conghui Zhu, and Tiejun Zhao. 2019. Understanding data augmentation in neural machine translation: Two perspectives towards generalization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5689–5695, Hong Kong, China. Association for Computational Linguistics.

Shivam Mhaskar and Pushpak Bhattacharyya. 2021. Pivot based transfer learning for neural machine translation: CFILT IITB @ WMT 2021 triangular MT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 336–340, Online. Association for Computational Linguistics.

Shivam Mhaskar and Pushpak Bhattacharyya. 2022. Multiple pivot languages and strategic decoder initialization helps neural machine translation. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 9–14, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Shivam Mhaskar and Pushpak Bhattacharyya. 2024. Pivot based neural machine translation: A survey.

Ananya Mukherjee, Hema Ala, Manish Shrivastava, and Dipti Misra Sharma. 2020. Mee : An automatic metric for evaluation using embeddings for machine translation. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 292–299.

Ananya Mukherjee and Manish Shrivastava. 2023. MEE4 and XLsim : IIIT HYD's submissions' for WMT23 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 800–805, Singapore. Association for Computational Linguistics.

Toshiaki Nakazawa, Kazutaka Kinugawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Makoto Morishita, Ondrej Bojar, Akiko Eriguchi, Yusuke Oda, Akiko Eriguchi, Chenhui Chu, and Sadao Kurohashi, editors. 2023. *Proceedings of the 10th Workshop on Asian Translation*. Asia-Pacific Association for Machine Translation, Macau SAR, China.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti,

John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

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.

Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Yan Shao, Christian Hardmeier, and Joakim Nivre. 2018. Universal word segmentation: Implementation and interpretation. *Transactions of the Association for Computational Linguistics*, 6:421–435.

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

Sanford B. Steever, editor. 1998. *The Dravidian Languages*. Routledge.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.