

NMT Evaluation Workshop

Minutes of the Workshop-cum-Report

Date: 10th – 12th July 2024

Venue: VMCC, IIT Bombay



From

Prof. Pushpak Bhattacharyya

Department of Computer Science and Engineering, IIT Bombay

Mumbai-400076

pb@cse.iitb.ac.in



Ministry of Electronics &
Information Technology,
Government of India



Contents

Executive Summary	3
Discussion Points	4
Day-1; 10th July, 2024	4
Day-2; 11th July 2024	5
Day-3; 12th July 2024	7
Post Workshop General Feedback:	8
NMT Evaluation Workshop: Key Outcomes/Takeaways	9
Annexure-1.....	10
Participants Feedback:	10
Annexure-II.....	13
Moments from the Workshop	13

MT Evaluation workshop
MINUTES OF MEETING
Venue: Room No. 32, VMCC, IIT Bombay

Executive Summary

This document contains the details of the 'Neural Machine Translation Evaluation' workshop held at IIT Bombay from 10th to 12th July 2024. This pan-MT consortia workshop was a pivotal part of the Mission Bhashini initiative. The workshop convened experts to address the challenges and advancements in Machine Translation (MT) for Indian languages. The event was marked by a collaborative spirit and a shared commitment to breaking language barriers. The event focused on developing a comprehensive evaluation framework for automatic and human assessment of English-Indic and Indic-Indic bidirectional Neural Machine Translation systems. The topics of discussion ranged from the translation of 22 Indian languages into English, to the evaluation of unsupervised metrics for translations like BERTScore, and the relevance of metrics like BLEU even in the era of LLMs. The event also thoroughly looked into situations like cultural identities as factors for translation and also repeated translation errors and similarities between syntax, semantics and morphemes of various languages as highlighted by annotators.

The workshop also discussed issues related to post-deployment hallucinations in MT systems, changes in their performance, and biases present in machine translation datasets. High-level questions like accurate emotion and pragmatics transmission in translation, hypothesis testing for understanding underlying distribution of text data for MT evaluation as well as linguistic phenomenon like multiword expressions in languages like Bengali were the important discussion points.

Over the three days, the event provided valuable insights into MT evaluation techniques and metrics, culminating in practical exercises and technical discussions on system development, data quality, and addressing emerging challenges like model bias and hallucination. Key outcomes emphasized the need for a balanced approach between classic and modern metrics, a standardized human evaluation framework, and a focus on improving data quality and mitigating biases in MT systems.

The minutes of the 3 days of the event contain valuable discussion points on the different techniques and metrics for evaluation of Machine Translation.

Discussion Points

Day-1; 10th July, 2024

The inaugural day of the EIMLT workshop laid the groundwork for the subsequent discussions by establishing the overarching goals and challenges in the field of Machine Translation (MT) for Indian languages.

01. The dignitaries (Shri. Vijay Kumar, Director, MEITY; Prof. Sachin Patwardha, Dean R&D, IIT Bombay; Prof. Varsha Apte, HoD, Dept. of CSE, IIT Bombay; Prof. Pushpak Bhattacharyya, IIT Bombay, Workshop Co-ordinator) welcomed and addressed the participants.
 - a. Mr. Vijay Kumar highlighted that to date, Bhashini's technologies have been adopted by more than 100 government agencies/industry partners, and the goal is to penetrate deep into the market so that the technology reaches common citizens.
02. Prof. Pushpak Bhattacharyya initiated the technical discussions of the workshop with foundational classic MT evaluation metrics (BLEU, ROGUE). He proposed the adoption of hypothesis testing to make any claims about MT performance and robustness.
03. Prof. Manish Shrivastav briefed about the evolution of MT evaluation metrics (CHRF, Bert Score, COMET, MEE4, XLSIM, YISI, INSTRUCTSCORE) and compared automated and human evaluation. He discussed the decision chart to pick a suitable evaluation metric based on resource availability for a language pair.
04. It was decided to use the combination of classic (eg. BLEU) and modern evaluation metrics till modern evaluation metrics mature.
05. Prof. Mitesh Khapra described developing an MT system for all 22 Indian languages. He also detailed parallel corpus collection and creation workflow, highlighting the challenges and the solutions.
06. Prof. Preethi Jyothi discussed the three approaches TOLSTOI, PRISM, and SALSA for enhancing ASR systems by using domain-specific text data and leveraging pre-trained language models to improve performance, especially for low-resource languages.
07. Ms. Ananya Mukherjee explained the potential effectiveness of unsupervised MT evaluation metrics. She also discussed a multilingual BERT-based metric, REUSE, focusing on chunk-level similarity.
08. Prof. Pushpak Bhattacharyya suggested extending our tests to other Indian languages to identify potential language divergences, which could offer valuable insights into the broader applicability of our findings.

Overall, the first day of the MT Evaluation Workshop provided a comprehensive overview of the MT landscape in India, setting the stage for in-depth discussions on evaluation methodologies, data challenges, and technical advancements in the following days.

Day-2; 11th July 2024

Prof. Pushpak Bhattacharyya initiated the day-2 discussions by briefing the background of the MT evaluation and raising the insightful question, “Will NLP be reduced to table lookup?” Furthermore, he emphasized the accurate transmission of emotion in the translation and pragmatics.

01. Prof. Nisheeth Joshi talked about human evaluation and highlighted its importance and challenges. He explained the human evaluation process and discussed Eagle's framework, human clarity score, multidimensional quality metric (MQM), scalar quality metric, etc. Additionally, he detailed HEVAL, which considers 11 translation error types for scoring.
02. Attention was drawn to the inability of automatic evaluation metrics to capture semantic similarity and hence the necessity of human evaluation. Also, the necessity for standardized training and evaluation processes was highlighted.
03. Prof. Pushpak Bhattacharyya remarked:
 - a. No assumptions are to be made about the competence and biases of evaluators.
 - b. A training program for human evaluators to perform granular translation quality assessments should be designed.
 - c. Human evaluation should lead to the findings of frequently occurring error patterns.
04. Prof. Shikhar Sarma recommended equipping the human evaluation framework with a feedback mechanism for human evaluators, informing them about their mistakes/biases to improve evaluation consistency.
05. Mr. Bibekanand Kundu and Prof. Manish Shrivastav highlighted the trustworthiness of evaluation metrics and recommended working towards integrating explainability.
06. Prof. Manish Shrivastav discussed the MIDAS framework and stressed having a uniform human evaluation framework across Bhashini MT consortia.
 - a. Ms. Barnali Pal raised a query about whether Indian English is suitable for the current MiDAS evaluation and proposed the potential inclusion of such data to increase robustness.
07. Ms. Ananya Mukherjee discussed the requirement for a synergy between benchmark data, automatic evaluation metrics, and a human evaluation framework for a foolproof MT assessment.
08. Ms. Ananya Mukherjee described various Direct Assessment-based human evaluation metrics highlighting their pros and cons, highlighting heavy reliance only on adequacy. She further introduced and discussed how Multidimensional Quality Metrics (MQM) and DA-SQM address the limitations of traditional metrics.
 - a. Prof. Manish Shrivastav pointed out that unless warranted by the source language, even ontologically related substitutions in the translations should be considered equally incorrect.
 - b. Prof. Dipti Sharma suggested the degree of an adequacy error should also be considered while evaluating the translations.

- c. Prof. Sudip Naskar pointed out the need for an evaluation metric sensitive to error patterns.
 - d. Prof. Pushpak Bhattacharyya remarked that human evaluators should heavily penalize MT systems that make mistakes in important language phenomena such as collocation, idioms, etc.
09. Prof. Manish Shrivastav discussed the creation of the EILMT benchmark data. He pointed out the lack of high quality benchmark datasets for Indian languages and highlighted the challenges in creating them. The aspects of a good benchmark dataset include diversity in terms of lexical and stylistic variations, coverage of language phenomena/error patterns, and domain-specificity.
10. Prof. Manish Shrivastav recommended having a Red Team to avoid data leakage.
11. Prof. Ganesh Ramakrishnan discussed his work on BharatGPT development and the UDAAN project. He highlighted the necessity of using ML optimization techniques and efficient approaches to ensure high performance even after the deployment. He further pointed out the importance of selecting data instances to ensure adequate representation of language phenomena.

Day-3; 12th July 2024

1. Prof. Manish Shrivastava introduced a hands-on exercise. He introduced the task, which is to recognize and annotate the types of error which a source and target sentence may have in a parallel corpus of Indian languages. Mr. Gajanan Rane detailed error categories for adequacy (e.g., incorrect terminology, orthographic errors) and fluency (e.g., syntactic errors, punctuation).
2. Prof. Soma Paul Discussed Multi-Word Expression (MWE) annotation for NMT, defining MWEs and stressing the need for workshops to address MWE challenges.
3. Various Consortia Members Discussed NMT Approaches:
 - Prof. Kishore Kashyap (Gauhati University): Used IndicNLP to tokenize Assami and explored bidirectional NMT between Assami and Hindi.
 - Shashipal Singh (CDAC-Pune): Discussed the shift from RNNs to Transformers, using OPEN-NMT and NLLB-200, and emphasized the need for diverse datasets.
 - Justwant Singh (NIT Silchar): Focused on a Transformer model, addressing challenges with subword tokenizers and Bengali script.
 - Sourabh (IITB): Described his NMT pipeline, employing LABSE for data filtering and training both bilingual and multilingual models, evaluating performance with BLEU and COMET, with eye-tracking experiments revealing that fixation duration correlates with translation quality, while saccade features are less reliable.
4. Ujjwal Sharma from IITB highlighted that model quantization, particularly using formats like BFloat16, reduces model size and speeds up inference with minimal performance loss. The study revealed that quantized models showed comparable performance to the base model, with BFloat16 providing the most balanced trade-off between speed and accuracy.
5. Prof. Pushpak Bhattacharyya and Dr. Karunesh Arora announced the BhashaKriti Hackathon, which will tackle challenges such as misspelling correction, next-word prediction, parallel NE pair and sentence generation, ASR output correction, and OCR output correction.
6. Prof. Pushpak Bhattacharyya discussed the importance of hypothesis testing in MT evaluation, emphasizing the need to understand probability distributions and confidence intervals to assess model performance accurately.
7. Nihar from IITB discussed how bias in NLP and MT systems often originates from training data, model decoders, and evaluation metrics. He highlighted methods for detecting bias, such as the Word Embedding Association Test (WEAT), and illustrated bias with examples, including gendered translations in MT systems.

8. Prof. Pushpak Bhattacharyya thanked everyone, appreciating the participants and guests for their presence, and ended the session by re-emphasizing the slogan of being NLP research scientists who give equal importance to both probability and linguistics.

Post Workshop General Feedback:

- Participants found the sessions highly informative and learned new topics, including hypothesis testing. Brainstorming ideas in groups was suggested.
- Suggested exploring more use cases for large language models (LLMs) and their evaluation methods. There could have been at least one break-out session (BOS), where: (i) workshop participants are grouped into small groups of 5-10 people, based on either their expertise/interest, or randomly, (ii) a slot of at least half an hour given for the BOS to brainstorm and think wild, (iii) a rapporteur assigned to each group who in some later session could report the proposals made during the BOS. Such break out sessions can/will bring out new and fresh ideas.
- More insights on LLMs and how linguistics inputs help LLM's inference for decoding Translation.
- It would be good if there were more hands-on exercises on the evaluation interface and NE and MWE annotation.

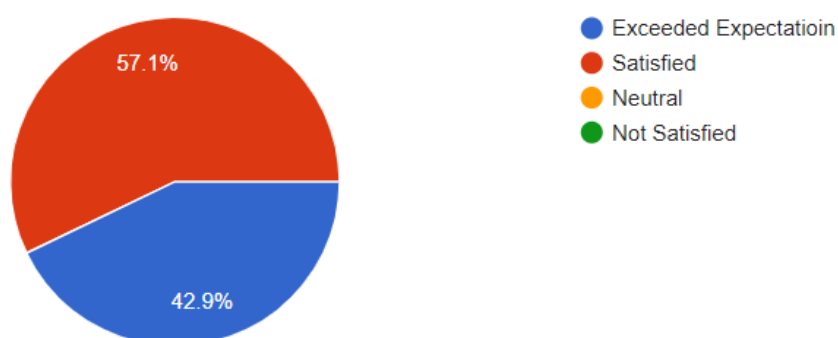
NMT Evaluation Workshop: Key Outcomes/Takeaways

1. NMT evaluation, be it automatic or human, must be backed by hypothesis testing.
2. Benchmark data for evaluating a particular language pair must be generated considering the use case and cultural sensitivity.
 - a. The aspects of a good benchmark dataset include diversity in terms of lexical and stylistic variations, coverage of language phenomena/error patterns, and domain-specificity.
3. Automatic evaluation metrics:
 - a. We cannot switch to embedding-based or supervised metrics for Indic NMT evaluation. A combination of classic and these metrics should be used until they mature.
 - b. The availability of resources should be used to decide on a set of metrics to be used for a particular language pair.
 - c. Chunk-based evaluation approaches should be investigated.
 - d. Integration of explainability with the metrics should be investigated.
4. Human evaluation:
 - a. Mitigating subjectivity in human evaluation is a major challenge.
 - i. Multiple evaluators (Inter-annotator agreement).
 - ii. Consideration for difficulty in identifying expert bilingual evaluators.
 - iii. Training program for human evaluators with a feedback mechanism to identify biased patterns.
 - b. The granularity of human evaluation needs to be deepened. MQM-backed sentence-level assessment needs to be undertaken.
 - c. MQM error types from MIDAS, HEVAL, and ILMT-MQM human evaluation methods have been combined to bring consistency.
 - i. Penalization of mistakes in crucial language phenomena like collocation and idioms must be included while computing the MQM score.
 - d. Hands-on translation evaluation exercise.
5. Language Phenomenon X Gaze-based Feature matrix should be the foundation of gaze-behavior-based translation quality analysis.
6. Deployment of NMT systems demands careful attention to speed-performance trade-offs.
7. The multi-word expression challenge (even at the level of its definition in NMT systems) needs to be resolved for Indian languages like Bengali to enable proper translation.
8. Bias mitigation in each component of the NMT workflow requires attention to each component: Training/Testing Data, Model, Evaluation metric, and Human evaluators.

Annexure-1

Participants Feedback:

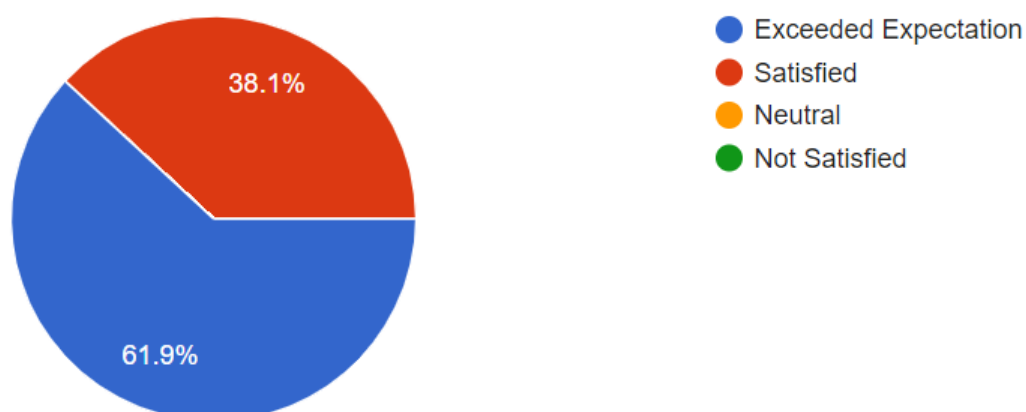
01. Satisfaction level of participants:



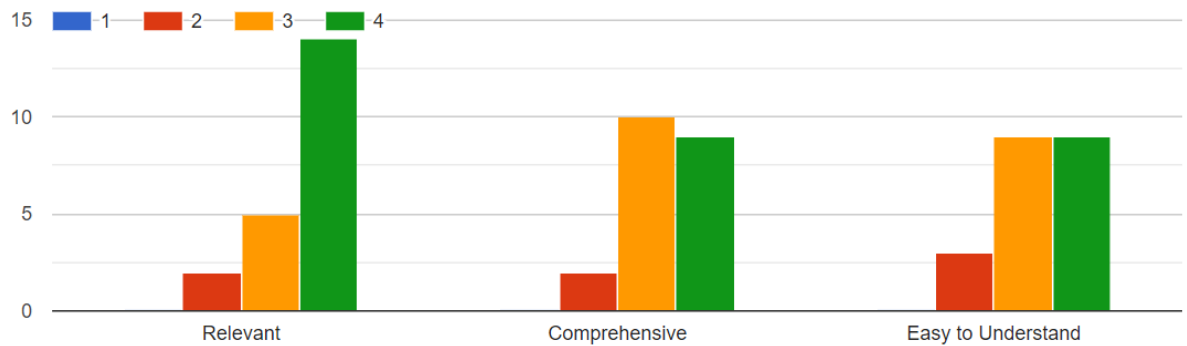
02. New things you learned during the workshop;

- Evaluation metrics useful depending on amount of data, human physical behaviour assessment during translation, MT ranking
- New NMT Approaches, Evaluation Scope, LLMs, etc
- Details of evaluation parameter
- Lots of research insights in a very organized manner
- Evaluation of Post Deployments MT Performance.
- Different evaluation metrics and perspectives to train the machine
- Metrics for MT Evaluation, Hypothesis Testing for MT Evaluation, Human Evaluation and Automatic Evaluation Parameters, MT Approaches, and lots more...
- Hypothesis testing, NMT approaches, human evaluation, evaluation metrics

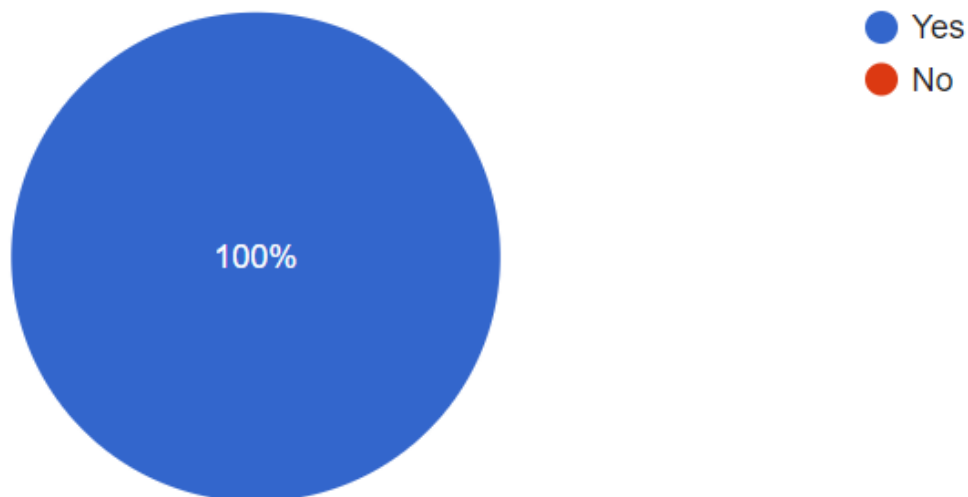
03. Feedback about speakers:



04. About the Content of the workshop:



05. Organization of workshop:



06. Most liked Topics (selected):

- Discussion on LLMs
- Hypothesis Testing
- MWE, NER, POS Annotations for MT and Eye Tracking Presentation + Translation Analysis and Hands on exercise on Human Eval Benchmarking (MQM +DA)
- Speech to Speech translation by Preethi Jyothi ma'am
- Resources (Data, GPU, Annotators , etc.) by IITM
- Pushpak Sir's teaching methods and hypothesis
- Hands on exercise on Human Eval Benchmarking (MQM +DA)
- Each topic was relevant and equally important, but as a linguist, I liked the lectures related to the linguistic analysis of the corpus.
- Human evaluation, Multiword expression, Eye tracking presentation

07. Least liked about workshop:

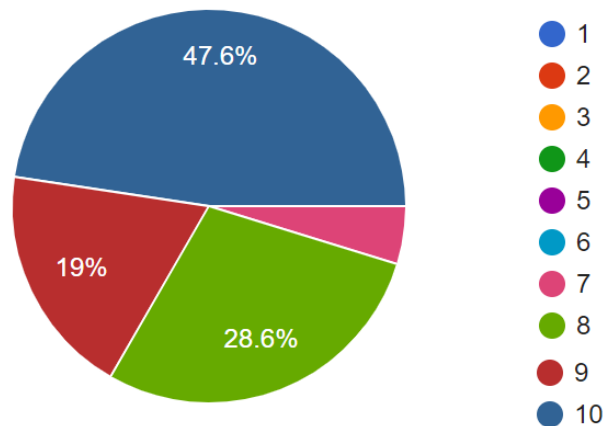
- NMT Approches session was too fast. Needed more time.

- b. Liked everything... only problem i faced was BSNL connectivity... Blessing in Disguise...

08. Topics/activities that can be included in future workshops:

- a. More on LLM use cases and their evaluation
- b. There could have been at least one break out session (BOS), where: (i) workshop participants are grouped into small groups of 5-10 people, based on either their expertise/interest, or randomly, (ii) a slot of at least half an hour given for the BOS to brainstorm and think wild, (iii) a rapporteur assigned to each group who in some later session could report the proposals made during the BOS. Such break out sessions can/will bring out new and fresh ideas.
- c. Yes half a day session on NMT model training approaches in the beginning would have been helpful.
- d. More insights on LLMs and how linguistics inputs help LLM's inference for decoding Translation
- e. It would be good if there were more hands-on exercises on the evaluation interface and NE and MWE annotation. Additionally, our technical team needs step-by-step guidance on dictionary creation from the available parallel corpus so that we can utilize it for better output, as explained by Prof. Sikhar.

09. Ratings of workshop on a scale of 1-10



Annexure-II

Moments from the Workshop







Ministry of
**ELECTRONICS AND
INFORMATION TECHNOLOGY**
Government of India

Neural Machine Translation Evaluation Workshop



SCAN TO LOCATE
@VICTOR_MENEZES_CONVENTION_CENTRE

DATE: 10th TO 12th JULY, 2024.

VENUE: VMCC - ROOM NO.32



@cfiltnlp



@cfiltindia



@cfilt_nlp



@cfilt