



Human Evaluation of Some English-Hindi Machine Translation Systems

Riya Lonchenpa^{1*} , Anju Krishnia² , Mhasilenuo Vizo³ , Nisheeth Joshi⁴

¹Department of English and Modern European Languages, Banasthali Vidyapith. *Corresponding author.

^{2,3}Department of Hindi and Modern Indian Languages, Banasthali Vidyapith.

⁴Centre for Artificial Intelligence, Banasthali Vidyapith.

Abstract

This study examines the human evaluation of machine translation (MT) systems, with particular emphasis on its significance in the contemporary era of neural machine translation. The evaluation employed a scalar quality metric comprising 11 distinct parameters, each rated on a scale from 0 to 4. Three MT engines—Google Translate, Bing Translator, and Anuvad MT Engine—were assessed using this framework. Human evaluators systematically analyzed translations across all parameters and documented recurring error patterns specific to each engine. Among the systems evaluated, the Anuvad MT Engine demonstrated superior performance in terms of overall translation quality.

Keywords: Neural Machine Translation, Human Evaluation, Scalar Quality Metric.

Conflicts of Interest: The author/s declared no conflicts of interest.

Funding: Ministry of Electronics and Information Technology, Government of India.

Article History: Received: 05 April 2025. Revised: 10 September 2025. Accepted: 12 September 2025. First published: 15 September 2025.

Copyright: © 2025 by the *author/s*.

License: License Aesthetix Media Services, India. Distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Published by: [Aesthetix Media Services, India](#)

Citation: Lonchenpa, R., Krishnia, A., Vizo, M., & Joshi, N. (2025). Human Evaluation of Some English-Hindi Machine Translation Systems. *Rupkatha Journal*, 17(3). <https://doi.org/10.21659/rupkatha.v17n3.04g>



1. Introduction

In a world where technology and automation continue to grow, human evaluation remains essential across various fields. Despite the rise of machine learning, artificial intelligence, and automated systems, it is human evaluation that is crucial for assessing subtle aspects such as context, relevance, ethical considerations, cultural sensitivity, and user satisfaction through the insights of human experts, which automated metrics might miss. Therefore, human evaluation involves involving humans or experts to analyse the outputs of machine learning models to measure their performance, usability, accuracy, and relevance in real-world situations. It addresses the limitations of automated metrics, ensuring not only accuracy but also outputs that are contextually appropriate and ethically sound. It helps to address and reduce biases, promote fairness and reliability, and provides human feedback that refines models to better meet user needs, thereby enhancing overall user engagement and satisfaction.

To adapt to the continuous flow of new data and emerging issues, machine learning models undergo ongoing evaluation and refinement, reflecting the perpetual human evaluation process. Some challenges faced by human evaluators include the influence of their own perspectives and experiences, which can lead to inconsistent assessments where the same output might be judged differently by different evaluators, thereby affecting reliability. This process is often labour-intensive and time-consuming, especially when dealing with large datasets or models requiring frequent updates. It can also be less cost-effective and may demand specialised knowledge or expertise in specific domains, which may not always be available, thus limiting its effectiveness and accuracy. Despite these challenges, human evaluation remains a vital component in ensuring that models are developed and deployed to meet real-world needs. By providing technically sound, accurate, fair, and contextual judgement-based, user-centric outputs, human evaluators help create machine learning systems that are trustworthy, effective, and capable of making meaningful societal contributions. Therefore, human evaluation continues to be a cornerstone of decision-making in a technology-driven world. While automated systems offer scalability and efficiency, human evaluators provide essential depth of understanding, ethical judgment, and contextual awareness to ensure these systems' effectiveness.

2. Literature Survey

Bojic et al. (2023) proposed a scale-based evaluation scoring system and its application to a specific NLP system, thereby identifying gaps in existing methodologies. Mathewson & Pilarski (2022) designed and evaluated Human-Centred Interactive Machine Learning (IML) through decision-making processes initiated by humans to mitigate risks and ensure integrity, accountability, and reliability in system development. Human evaluation in ML systems could potentially be replaced by large language models (LLMs), as they are capable of providing consistent and reliable assessments of text quality, as suggested by Chiang & Lee (2023) in their research. LLMs are mainly used to evaluate texts, and their responses align with those of expert human evaluators.

For the consistent human evaluation of machine translation systems across various language pairs, focusing on semantic equivalence, the paper by Licht et al. (2022) introduces a new scoring metric

XSTS and a cross-lingual calibration method that enhance assessment across language pairs. It demonstrates the effectiveness of XSTS in large-scale evaluation studies. Research conducted by Freitag et al. (2021) highlights the importance of context and error analysis to prevent erroneous conclusions in evaluating system performance. With the use of explicit error analysis and a Multidimensional Quality Metrics (MQM) framework, desired results were achieved.

Rusinovich et al (2024) research focuses on integrating Artificial Intelligence (AI) with Machine Learning (ML) in healthcare and clinical settings to enhance patient outcomes and streamline healthcare processes. This approach considers user experience, patient engagement, and the ethical implications of deploying AI techniques in the healthcare sector, while also addressing data privacy concerns and emphasising the importance of interdisciplinary collaboration.

Human evaluation in ML systems plays a vital role in interpretability and feedback incorporation, as proposed by Kumar & Sharma (2022) in their research on human feedback adaptive learning using interpretable and interactive intelligent systems. The paper focuses on the principles, guidelines, metrics, and adaptive learning algorithms to design and provide human feedback interactive ML systems with conflict resolution. According to Van Der Lee et al. (2021), human evaluation in ML systems assesses automatically generated texts, highlighting the importance of human judgment in evaluating text quality. The paper by Kadam & Vaidya (2021) evaluated the cognitive aspects of machine learning agents, focusing on human-like evaluation methods for assessing their performance and decision-making processes. By analysing human-like cognitive processes in AI systems, one finds the positive impact of autonomy and the negative impact of rewards on intrinsic motivation. Issac et al. (2019) highlighted human evaluation in ML systems by focusing on understanding the factors that make models interpretable. Human-subject experiments were conducted to assess model interpretability, concluding that variable repetitions have relatively less effect on interpretability, while the number of cognitive chunks and model size significantly influence it. A framework was formulated by Hannah et al. (2022) for human evaluation by understanding cognitive differences and matching trials between human participants and algorithm evaluations. This paper demonstrates the utility of designing human performance evaluations for one-shot learning.

Nadia et al (2018) discussed the challenges of evaluating interactive machine learning systems, which learn from and adapt to humans while also receiving feedback. To study the system's behaviour and utility, combining algorithm-centred analysis and human-centred evaluation was suggested. To support both visible and hidden roles in interactive machine learning, the authors emphasised the importance of human-AI interaction methodologies for employing an interactive evolutionary approach for guided search.

3. Human Evaluation Methodology

3.1 Machine Translation Engines Used

In our study, we evaluated three Machine Translation (MT) engines—Google Translate, Bing Translator, and Anuvaad—which are distinguished by their target language coverage, underlying technological frameworks, and intended user applications. While automated evaluation metrics offer objective and scalable ways to compare, human evaluation provides a more nuanced

understanding of translation quality. It captures subtleties such as contextual appropriateness, cultural sensitivity, and idiomatic accuracy—dimensions often overlooked by algorithmic scoring systems. Therefore, integrating human judgment is essential for a comprehensive assessment of MT system performance.

3.2 Machine Translation Engines Used

The dataset of 1024 sentences was considered, and translations were carried out for each platform based on different MT engines. Once we obtained the translated sentences from Hindi to English, the human evaluation process was implemented. Using the 11 parameters (as seen in Table 1) proposed in this research, 1024 translated sentences were assessed. Additionally, ranks were assigned to each sentence based on the HEval (Human Evaluation Metric) Scale, with certain ranks given between 0 and 4 (as seen in Table 2). After ranking each translation from the various MT engines, a comparison was made between the results of the earlier experiments to reach a final conclusion.

Based on the parameters of Table 1, the evaluators or experts assign a ranking to each translated sentence according to the proposed HEval-Scale, as shown in Table 2. After each sentence is ranked based on the 11 parameters in table 1, an average is calculated for each sentence across each MT engine platform, enabling a comparison among the three proposed MT engines.

S.No	Parameter	Description
1.	Gender and Number of the Noun(s)	Change of gender in the translated sentence and the number of nouns.
2.	Proper Noun(s)	Proper noun of source sentence not identified as a proper noun in target sentence.
3.	Use of Adjectives and Adverbs	Incorrect use of adjectives and adverbs corresponding to the Nouns and Verbs.
4.	Proper words/Synonym (Lexical Choice)	Source text left untranslated or improper translation of words.
5.	Sequence of Phrases and Clauses	Incorrect translation of phrases and clauses in the target sentence or its misarrangement.
6.	Punctuation Marks	Incorrect punctuation.
7.	Tenses	Use of incorrect tense in the target sentence.
8.	Voice	Use of incorrect voices in the target sentence.
9.	Semantics	Problems with the semantics of the target sentence
10.	Fluency	Incorrect fluency in the target sentence
11.	Evaluating the translation of the source sentence (w.r.t. syntax and intended meaning).	Change in the syntax and the intended meaning of the target sentence from the source sentence.

Table 1. Human Evaluation Measure

Score	Description
4	Ideal
3	Perfect
2	Acceptable
1	Partially Acceptable
0	Not Acceptable

Table 2. Human Evaluation Scale

4. Results and Discussion

In this section, we examine the various translations generated by different MT engines when given a similar source sentence and the scores assigned by the human evaluator.

Example 1

Source Sentence: Some of the government initiatives like Swachh Bharat Abhiyan (to create mass awareness on cleanliness and environmental pollution, to eliminate open defecation, and improve solid waste management) and Namami Gange (significant reduction of pollution, conservation and resuscitation of the river Ganga) are worth mentioning here.

Google Translate Output: स्वच्छ भारत अभियान (स्वच्छता और पर्यावरण प्रदूषण पर व्यापक जागरूकता पैदा करने, खुले में शौच को खत्म करने और ठोस अपशिष्ट प्रबंधन में सुधार करने के लिए) और नमामि गंगे (प्रदूषण में उल्लेखनीय कमी, गंगा नदी का संरक्षण और पुनर्जीवन) जैसी कुछ सरकारी पहल सार्थक हैं। यहाँ उल्लेख कर रहे हैं।

Bing Translator Output: स्वच्छ भारत अभियान (स्वच्छता और पर्यावरण प्रदूषण पर जन जागरूकता पैदा करने, खुले में शौच को खत्म करने और ठोस अपशिष्ट प्रबंधन में सुधार करने के लिए) और नमामि गंगे (गंगा नदी के प्रदूषण में उल्लेखनीय कमी, संरक्षण और पुनर्जीवन) जैसी कुछ सरकारी पहलें यहां उल्लेखनीय हैं।

Anuvad MT Output: स्वच्छ भारत अभियान (स्वच्छता और पर्यावरण प्रदूषण पर जन जागरूकता पैदा करने, खुले में शौच को समाप्त करने और ठोस अपशिष्ट प्रबंधन में सुधार करने के लिए) और नमामि गंगे (प्रदूषण में महत्वपूर्ण कमी, गंगा नदी के संरक्षण और पुनर्जीवन) जैसी कुछ सरकारी पहल यहां उल्लेखनीय है।

Example 2

Source Sentence: The highest is the Xvaitu, who is self-sufficient and no longer requires religious instruction; the Airyamna, or friend, who befriends people to keep them from going astray; and the Verezanah, or worker, who is the active field worker on the ground and is to spread the faith.

Google Translate Output: सर्वोच्च Xvaitu है, जो आत्मनिर्भर है और उसे अब धार्मिक शिक्षा की आवश्यकता नहीं है; एयरयमना, या मित्र, जो लोगों को भटकने से बचाने के लिए उनसे मित्रता करता है; और वेरेज़ाना, या कार्यकर्ता, जो ज़मीन पर सक्रिय क्षेत्र कार्यकर्ता है और उसे विश्वास फैलाना है।

Bing Translator Output: उच्चतम Xvaitu है, जो आत्मनिर्भर है और अब धार्मिक निर्देश की आवश्यकता नहीं है; एयरयमना, या दोस्त, जो लोगों को भटकने से बचाने के लिए उनसे दोस्ती करता है; और वेरेज़ाना, या कार्यकर्ता, जो ज़मीन पर सक्रिय क्षेत्र कार्यकर्ता है और विश्वास फैलाने के लिए है।

Anuvad MT Output: सबसे ऊँचा है ज़्वैतू, जो आत्मनिर्भर है और जिसे अब धार्मिक शिक्षा की आवश्यकता नहीं है; ऐरियामना, या दोस्त, जो लोगों को भटकने से रोकने के लिए उनसे दोस्ती करता है; और वेरेज़ाना, या कार्यकर्ता, जो ज़मीन पर सक्रिय क्षेत्र कार्यकर्ता है और विश्वास फैलाने के लिए है।

Example 3

Source Sentence: The seat of the court was the Chamber of Princes in the Parliament Building in Delhi.

Google Translate Output: अदालत की सीट दिल्ली में संसद भवन में चैंबर ऑफ़ प्रिंसेस थी।

Bing Translator Output: अदालत की सीट दिल्ली में संसद भवन में चैंबर ऑफ़ प्रिंसेस थी।

Anuvad MT Output: अदालत की सीट दिल्ली में संसद भवन में चैंबर ऑफ़ प्रिंसेस थी।

Example 4

Source Sentence: On 28 December 2010, the non-governmental organisation, Coalition to Uproot Ragging from Education (CURE) co-founder Harsh Agarwal wrote a letter to the Raghavan Committee, a panel headed by former HRD Minister R. K. Raghavan, explaining distress over a particular scene, where the lead actors get ragged by seniors.

Google Translate Output: 28 दिसंबर 2010 को, गैर-सरकारी संगठन, कोलिशन टू अपरूट रैगिंग फ़्रॉम एजुकेशन (CURE) के सह-संस्थापक हर्ष अग्रवाल ने राघवन समिति को एक पत्र लिखा, जो पूर्व मानव संसाधन विकास मंत्री आर. जहां मुख्य कलाकारों की सीनियर्स द्वारा रैगिंग की जाती है।

Bing Translator Output: 28 दिसंबर 2010 को, गैर-सरकारी संगठन, शिक्षा से रैगिंग को उखाड़ने के लिए गठबंधन (क्योर) के सह-संस्थापक हर्ष अग्रवाल ने राघवन समिति, पूर्व मानव संसाधन विकास मंत्री आरके राघवन की अध्यक्षता वाली एक समिति को एक पत्र लिखा, जिसमें एक विशेष दृश्य पर संकट को समझाया गया, जहां मुख्य अभिनेताओं को वरिष्ठों द्वारा रैगिंग की जाती है।

Anuvad MT Output: 28 दिसंबर 2010 को, गैर-सरकारी संगठन, कोएलिशन टू अपरूट रैगिंग फ़्रॉम एजुकेशन (सी. यू. आर. ई.) के सह-संस्थापक हर्ष अग्रवाल ने राघवन समिति को एक पत्र लिखा, जो पूर्व

एच. आर. डी. मंत्री आर. के. राघवन की अध्यक्षता वाली एक समिति थी, जिसमें एक विशेष दृश्य पर संकट की व्याख्या की गई थी, जिसमें मुख्य अभिनेता वरिष्ठों द्वारा परेशान किए जाते हैं।

Example 5

Source Sentence: A 2003 Battelle Environmental Evaluation System (BEES) evaluation of Bangalore's physical, biological and socioeconomic parameters indicated that Bangalore's water quality and terrestrial and aquatic ecosystems were close to ideal, while the city's socioeconomic parameters (traffic, quality of life) air quality and noise pollution were poor.

Google Translate Output: बेंगलुरु के भौतिक, जैविक और सामाजिक-आर्थिक मापदंडों के 2003 बैटल पर्यावरण मूल्यांकन प्रणाली (बीईईएस) मूल्यांकन से संकेत मिलता है कि बेंगलुरु की जल गुणवत्ता और स्थलीय और जलीय पारिस्थितिकी तंत्र आदर्श के करीब थे, जबकि शहर के सामाजिक आर्थिक पैरामीटर (यातायात, जीवन की गुणवत्ता) वायु गुणवत्ता और ध्वनि प्रदूषण गरीब थे।

Bing Translator Output: बेंगलोर के भौतिक, जैविक और सामाजिक आर्थिक मापदंडों के 2003 के बैटल पर्यावरण मूल्यांकन प्रणाली (बीईईएस) ने संकेत दिया कि बेंगलोर की जल गुणवत्ता और स्थलीय और जलीय पारिस्थितिक तंत्र आदर्श के करीब थे, जबकि शहर के सामाजिक आर्थिक पैरामीटर (यातायात, जीवन की गुणवत्ता) वायु गुणवत्ता और ध्वनि प्रदूषण खराब थे।

Anuvad MT Output: बेंगलोर के भौतिक, जैविक और सामाजिक-आर्थिक मानकों के 2003 के बैटल पर्यावरण मूल्यांकन प्रणाली (बी. ई. ई. एस.) मूल्यांकन ने संकेत दिया कि बेंगलोर की जल गुणवत्ता और स्थलीय और जलीय पारिस्थितिकी तंत्र आदर्श के करीब थे, जबकि शहर के सामाजिक-आर्थिक मानकों (यातायात, जीवन की गुणवत्ता) वायु गुणवत्ता और ध्वनि प्रदूषण खराब थे।

The results of these five examples are shown in Tables 3, 3,4, and 5 respectively for Google Translate, Bing Translator, and Anuvad MT Engine. While performing translations using Google MT Engine, a common error was when the source sentence was longer than usual; as a result, the translation often broke into two sentences separated by a full stop, thus losing punctuation, semantics, and the original meaning of the target sentence. For example

Source Sentence: As per sub-section (3), on receipt of such an application, the Competent Authority, by order in writing, may either grant or refuse to grant such permission after giving an opportunity to the parties of being heard and after making such summary inquiries into the circumstances of the case as it thinks fit.

Google Translate Output: उपधारा (3) के अनुसार, ऐसे आवेदन की प्राप्ति पर, सक्षम प्राधिकारी, लिखित आदेश द्वारा, पक्षों को सुनवाई का अवसर देने और ऐसी संक्षिप्त पूछताछ करने के बाद या तो ऐसी अनुमति दे सकता है या देने से इनकार कर सकता है। मामले की परिस्थितियों में जैसा वह उचित समझे।

Another limitation of the Google Translation is when the source sentence contains formal salutations like Mr. or Mrs., the ' '. In the formal salutation is considered as the ending of the sentence by Google thus splitting the target sentence into two thereby affecting the overall meaning of the sentence. For example

Source Sentence: Later, an agreement was entered into with M/s. HLL with effect from March 22, 1997.

Google Translate Output: बाद में मैसर्स के साथ एक समझौता किया गया। एचएलएल 22 मार्च 1997 से प्रभावी।

Also in some sentences, it was noticed that some words of the source language are not translated into the target language and are available as it is in the target sentence. For example

Source Sentence: In the 1980s, the Spanish firm Esperanza y Cia developed a 120-mm mortar bomb that contained 21 anti-armour submunitions.

Google Translate Output: 1980 के दशक में, स्पैनिश फर्म Esperanza y Cia ने 120-मिमी मोर्टार बम विकसित किया था जिसमें 21 कवच-रोधी सबमिशन शामिल थे।

Human evaluation, when performed on Bing MT Engine, it was concluded that Bing Translations majorly lacked in providing the proper words repetitively, thus affecting the semantics and changing the meaning of the target sentence. For example

Source Sentence: Cinnamon is native to India and, as such, its uses range from that of a flavoring agent to a medicinal agent.

Bing Translator Output: दालचीनी भारत की मूल निवासी है और इस तरह, इसका उपयोग एक स्वादिष्ट बनाने वाले एजेंट से लेकर औषधीय एजेंट तक होता है।

Source Sentence: This template was modified in 1875 to the court design that exists today, with markings similar to Wingfield's version, but with the hourglass shape of his court changed to a rectangle.

Bing Translator Output: इस टेम्पलेट को 1875 में अदालत के डिजाइन में संशोधित किया गया था जो आज मौजूद है, विंगफील्ड के संस्करण के समान चिह्नों के साथ, लेकिन उनके दरबार के घंटे के चश्मे के आकार के साथ एक आयत में बदल गया।

While translations were performed on Anuvaad MT Engine, it was noticed that the least number of errors was noted here. In some sentences, the translation was not proper as it lacked the adjective's proper translation, thus affecting the semantics of the target sentence. For example

Source Sentence: An appearance is a bunch of attributes related to the service person, like their shoes, clothes, tie, jewelry, hair-style, make-up, watch, cosmetics, perfume, etc.

Anuvad Translator Output: उपस्थिति सेवा करने वाले व्यक्ति से संबंधित विशेषताओं का एक समूह है, जैसे कि उनके जूते , कपड़े , टाई , आभूषण , केश शैली , मेकअप , घड़ी , सौंदर्य प्रसाधन , इत्र आदि।

Therefore, in Anuvaad, there occurred the least number of errors among the translations of all the three above-discussed MT Engine.

5. Conclusion

Upon comparing the results of all the MT engines, it was found that Anuvaad MT Engine stood out as the best translation system with the fewest errors and the highest quality outputs, with Bing MT Engine ranking next with fewer errors compared to Google MT Engine. This indicates that Google Translator scores the lowest among the three. This research thus highlights the importance of Human Evaluation in the machine translation development process, thereby enhancing the relevance and effectiveness of machine translation systems.

Acknowledgement

This work is supported by the funding received from the Ministry of Electronics and Information Technology, Government of India, for the project "English to Indian Languages and vice versa Machine Translation System" under National Language Translation Mission (NLTM): Bhashini through administrative approval no. 11(1)/2022 HCC(TDIL) Part 5.

References

- Bojic, I., Chen, J., Chang, S. Y., Ong, Q. C., Joty, S., & Car, J. (2023). Hierarchical Evaluation Framework: Best practices for human evaluation. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2310.01917>
- Mathewson, K. W., & Pilarski, P. M. (2022). A Brief Guide to Designing and Evaluating Human-Centered Interactive Machine Learning. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2204.09622>
- Chiang, C., & Lee, H. (2023). Can large language models be an alternative to human evaluations? arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2305.01937>
- Licht, D., Gao, C., Lam, J., Guzman, F., Diab, M., & Koehn, P. (2022). Consistent Human Evaluation of Machine Translation across Language Pairs. arXiv (Cornell University). <https://doi.org/10.48550/arxiv.2205.08533>
- Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. <https://typeset.io/papers/experts-errors-and-context-a-large-scale-study-of-human-u6vnfom44l>
- Rusinovich, Y., Vareiko, A., & Shestak, N. (2024). Human-centered evaluation of AI and ML projects. Deleted Journal, 1(2). <https://doi.org/10.62487/ypqhkt57>
- Kumar, P., & Sharma, M. (2022). Human-Feedback adaptive learning using interpretable and interactive intelligent systems. In CRC Press eBooks (pp. 15–38). <https://doi.org/10.1201/9781003158165-2>

- Van Der Lee, C., Gatt, A., Van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67, 101151. <https://doi.org/10.1016/j.csl.2020.101151>
- Kadam, S., & Vaidya, V. (2021). Cognitive evaluation of machine learning agents. *Cognitive Systems Research*, 66, 100–121. <https://doi.org/10.1016/j.cogsys.2020.11.003>
- Isaac, Lage, Emily, Chen, Jeffrey, He, Menaka, Narayanan., Been, Kim, Samuel, J., Gershman, Finale, Doshi-Velez. (2019). Human Evaluation of Models Built for Interpretability.
- Hannah, P., Cowley, Mandy, L., Natter, Karla, Gray-Roncal, Rebecca, Rhodes, Erik, C., B., Johnson, Nathan, Drenkow, Tim, Shead, Frances, S., Chance, Brock, A., Wester, Will, Gray-Roncal. (2022). A framework for rigorous evaluation of human performance in human and machine learning comparison studies. *Dental science reports*, doi: 10.1038/s41598-022-08078-3
- Nadia, Boukhelifa., Anastasia, Bezerianos., Evelyne, Lutton. (2018). Evaluation of Interactive Machine Learning Systems. doi: 10.1007/978-3-319-90403-0_17

Riya Lonchenpa works as a Computational Linguist at the Speech and Language Processing Lab, Centre for Artificial Intelligence, Banasthali Vidyapith. She is also pursuing her PhD in English at Banasthali Vidyapith. Her areas of interest include Mythopoeia in Literature and Computational Psycholinguistics.

Anju Krishnia is working as a Computational Linguist at the Speech and Language Processing Lab, Centre for Artificial Intelligence, Banasthali Vidyapith. She has completed her PhD in Hindi from Banasthali Vidyapith. Her areas of interest include Comparative Literature and Computational Psycholinguistics.

Mhasilenuo Vizo works as a Computational Linguist at the Speech and Language Processing Lab, Centre for Artificial Intelligence, Banasthali Vidyapith. She is also pursuing her PhD in Hindi from Banasthali Vidyapith. Her areas of interest include Tribal Literature and Linguistics.

Nisheeth Joshi is working as a Professor of Artificial Intelligence at the Speech and Language Processing Lab, Centre for Artificial Intelligence, Banasthali Vidyapith. He has over 20 years of experience in designing complex AI systems. His areas of interest include Computational Linguistics, Machine Translation, and Language and Cognition.

Example	Noun/ Gender	Proper Noun	Adjectives and Adverbs	Proper Words Synonyms	Clause s Phrase s	Punctuati on	Tens es	Voic e	Semantic s	Fluency	Intended Meaning	Averag e Score
Example 1	3	3	3	3	3	3	0	0	0	0	0	0.4091
Example 2	0	3	3	0	3	0	0	3	0	0	0	0.2727
Example 3	3	3	3	3	3	3	3	3	3	3	3	0.7500
Example 4	0	3	3	0	3	0	0	0	0	0	0	0.2045
Example 5	3	3	3	3	0	0	1	3	0	0	0	0.3636

Table 3. Human Evaluation Scores of Google Translate

Example	Noun/ Gender	Proper Noun	Adjectives and Adverbs	Proper Words Synonyms	Clause s Phrase s	Punctuati on	Tens es	Voic e	Semantic s	Fluency	Intended Meaning	Averag e Score
Example 1	4	4	4	4	4	4	4	4	4	4	4	1.0000
Example 2	4	4	4	0	4	1	4	4	0	0	0	0.5682
Example 3	3	4	4	3	4	4	4	4	3	3	3	0.8182
Example 4	4	4	4	4	4	4	4	4	4	4	4	1.0000
Example 5	4	4	4	4	4	4	4	4	4	4	4	1.0000

Table 4. Human Evaluation Scores of Bing Translate

Example	Noun/ Gender	Proper Noun	Adjectives and Adverbs	Proper Words Synonyms	Clause s Phrase s	Punctuati on	Tens es	Voic e	Semantic s	Fluency	Intended Meaning	Averag e Score
Example 1	4	4	4	4	4	4	4	4	4	4	4	1.0000
Example 2	4	4	4	4	4	4	4	4	4	4	4	1.0000
Example 3	4	4	4	4	4	4	4	4	4	4	4	1.0000
Example 4	4	4	4	4	4	4	4	4	4	4	4	1.0000
Example 5	4	4	4	4	4	4	4	4	4	4	4	1.0000

Table 5. Human Evaluation Scores of Anuvad MT System