# **To Study Different Lexical Automatic Machine Translation Evaluation Metric for Indic Languages**

**Lalit Bisht Divyaprem Vanapamula Ankit Ghosal** *Apex Institute of Technology Apex Institute of Technology Apex Institute of Technology*

*(CSE) (CSE) (CSE)*

*Chandigarh University, Chandigarh University, Chandigarh University,*

*Mohali,140413, India Mohali,140413, India Mohali,140413, India* [*lalitbisht540@gmail.com*](mailto:lalitbisht540@gmail.com) [vdivyaprem@gmail.com](mailto:vdivyaprem@gmail.com) [ghosalankit3@gmail.com](mailto:ghosalankit3@gmail.com)

**Ayush Raj Shweta Chauhan**

*Apex**Institute of Technology Apex Institute of Technology*

*(CSE) (CSE)*

*Chandigarh University, University Centre for Research and Development,*

*Mohali,140413, India Chandigarh University,Mohali,140413, India*

[rajayush0221@gmail.com](mailto:rajayush0221@gmail.com) [shwetachauhansc.2000@gmail.com](mailto:shwetachauhansc.2000@gmail.com)

*Orcid id: 0000-0002-6598-1992*

**Abstract**

This research paper aims to study and compare different lexical automatic machine translation evaluation metrics for Indic languages. With the rise of machine translation systems, it has become essential to evaluate the quality of translations produced by these systems. However, the existing evaluation metrics designed for English and other European languages may not be suitable for Indic languages due to their complex morphology and syntax. Therefore, this study evaluates four different metrics, namely, BLEU, METEOR, TER, and NIST to identify the most suitable evaluation metric for Indic languages. The study uses datasets for three Indic languages, namely, Hindi, Bengali, and Telugu, and evaluates the metrics on various translation models. The study contributes to the field of machine translation by providing insights into suitable evaluation metrics for Indic languages.

The paper begins with an introduction to the importance of automatic machine translation evaluation and the challenges posed by Indic languages. It then provides a literature review of existing evaluation metrics and their limitations for Indic languages. Next, the paper presents the methodology used to conduct the study, which includes selecting the datasets and translation models and evaluating the metrics on the translations produced by these models. The paper then presents the results of the evaluation, which include the scores obtained by each metric for the three languages and the translation models. The paper concludes with a discussion of the implications of the study for the field of machine translation and suggestions for future research. Overall, the paper provides useful insights into the suitability of different automatic machine translation evaluation metrics for Indic languages and contributes to the development of better evaluation methods for these languages.

**Keywords:** indic, languages, metrics, translation, evaluation, machine, models.

**1 Introduction**

In our increasingly globalised world, machine translation has become an essential tool for communication. However, evaluating the quality of machine translations is a challenging task. Traditionally, human experts have been used to assess the quality of translations. However, this approach is time-consuming, expensive, and often subjective.

Machine translation (MT) is the task of automatically converting text from one natural language to another. MT is a challenging problem, especially for Indic languages, which are morphologically rich and have low availability of parallel corpora. Therefore, it is important to have reliable and robust methods to evaluate the quality of MT systems for Indic languages.

To address this issue, automatic evaluation metrics have been developed to assess the quality of machine translations. These metrics are based on various criteria such as fluency, adequacy, and accuracy. However, different metrics may provide different results, and it is important to understand their strengths and limitations to select the most appropriate metric for a particular application.

One of the most widely used methods for MT evaluation is to compare the output of a system with one or more human reference translations using automatic metrics. However, these metrics have limitations, such as relying on exact word matching, ignoring semantic similarity, and being sensitive to word order variations. Moreover, these metrics may not capture the linguistic diversity and complexity of Indic languages.

In this research paper, we aim to study and compare different lexical automatic machine translation evaluation metrics. We will explore the characteristics and performance of various metrics such as BLEU, METEOR, TER, and NIST. We will also discuss their advantages and limitations and provide insights into their suitability for different translation tasks.

Specifically, the paper will first provide an overview of the different types of automatic evaluation metrics and their main features. We will then conduct a comprehensive review of the literature to compare and contrast the most commonly used metrics. We will analyse the strengths and weaknesses of each metric, including their sensitivity to different types of errors, their ability to capture various aspects of translation quality, and their robustness across different languages and domains.

Next, we will conduct experiments to evaluate the performance of the selected metrics on a set of translation tasks. We will use different evaluation datasets and compare the results obtained using each metric. We will also investigate the correlations between the metrics and human judgments to assess their reliability and validity.

Finally, we will draw conclusions and provide recommendations for selecting the most appropriate metric for a given translation task based on our findings. We will also discuss future research directions and potential improvements for automatic machine translation evaluation metrics.

This research paper aims to contribute to the advancement of machine translation evaluation research by providing a comprehensive analysis of different lexical automatic evaluation metrics and their performance on various translation tasks.

**2 Indic Languages**

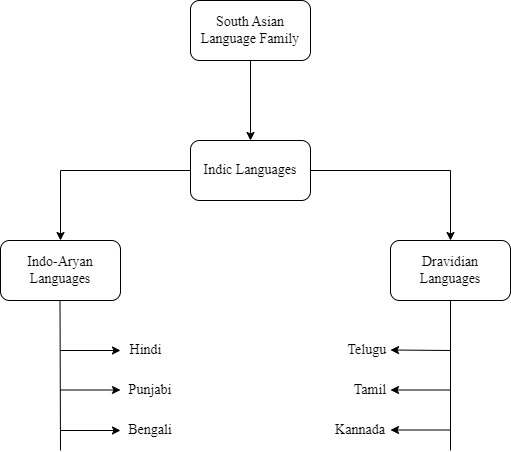
Indic languages are a group of languages that belong to the Indo-European language family and are mainly spoken in South Asia. The most widely spoken branch of Indic languages is the Indo-Aryan languages, which have more than 800 million speakers in India, Pakistan, Bangladesh, Nepal, Sri Lanka, and Maldives. Some of the major Indo-Aryan languages are Hindi, Bengali, Urdu, Punjabi, Marathi, Gujarati, Sindhi, Nepali, and Sinhala.

Another branch of Indic languages is the Dravidian languages, which are spoken by about 20% of Indians. The Dravidian languages are not related to the Indo-Aryan languages, but have influenced each other through contact and borrowing. Some of the major Dravidian languages are Tamil, Telugu, Kannada, Malayalam, and Odia.

The Indian Constitution recognizes 22 languages as official languages of India. These include 15 Indo-Aryan languages and 6 Dravidian languages. One of these languages is English, which is used as an associate official language along with Hindi. The Indian government also grants the status of classical language to six languages that have a long and rich literary tradition. These are Sanskrit, Tamil, Telugu, Kannada, Malayalam, and Odia.

Indic languages have a diverse and complex history and culture. They have developed various writing systems, such as Devanagari, Bengali-Assamese script, Gurmukhi script, Gujarati script, Oriya script, Sinhala script, Tamil script, Telugu script, Kannada script and Malayalam script. They have also produced many literary works of poetry, drama, epics, philosophy and religion. Some of the famous examples are the Vedas, the Ramayana, the Mahabharata, the Bhagavad Gita and the works of Kalidasa.

We have used Hindi, Bengali and Telugu for the purpose of this research paper. Hindi is the most used Indic language in the country with more than 500 million people calling it their native language. Bengali is the second most spoken language in India with a speaker base of more than 95 million and to add more variety we have also used Telugu which has a user base of more than 80 million. Using these languages, we aim to provide a comprehensive idea of how different evaluation metrics will perform when used to evaluate indic languages.



**Fig. 1** Various Indic Languages

**3 Role of Machine Translation Evaluation Methods in Machine Translation Evaluation**

Techniques for evaluating machine translation output are essential for determining its level of quality. The result of machine translation is evaluated in order to assess translation quality and pinpoint areas for improvement.

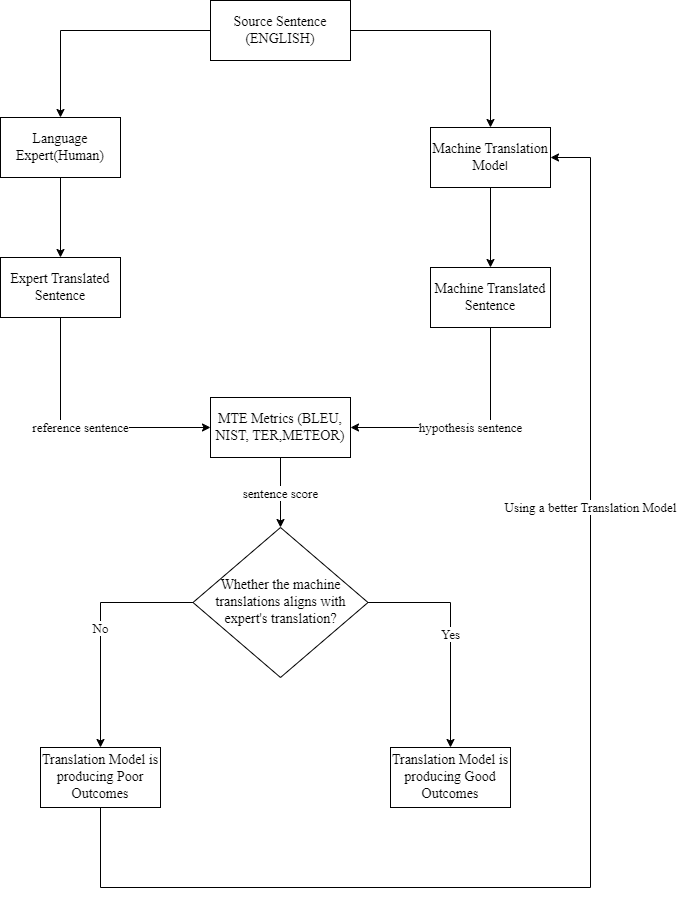
A number of machine translation evaluation techniques are available, including manual, automatic, and human evaluation. Human evaluation entails having translators rate the output of the machine translation using multiple criteria, including fluency, sufficiency, and correctness.

On the other hand, automatic evaluation uses metrics like BLEU, METEOR, and TER to gauge how well the output of machine translation is done. These metrics compare the output of the machine translation with the original text and give a score based on a number of factors, including grammatical correctness, sentence structure, and word overlap.

To evaluate the quality of machine translation output manually, a combination of human and automatic evaluation techniques is used. This approach combines the benefits of human and machine translation evaluation methods and provides more accurate and reliable results and this approach is also used in this paper to assess the quality of translations produced by machine translation systems as shown in fig. 1.

The role of MT evaluation methods in MT evaluation is to provide feedback and guidance for MT developers, users, and researchers. MT evaluation methods can help to identify the strengths and weaknesses of different MT systems, to compare and rank MT systems according to various criteria, to monitor and improve the quality of MT outputs over time, and to explore the impact of MT on various domains and applications. MT evaluation methods can also help to advance the scientific understanding of MT by providing empirical evidence and insights into the linguistic, cognitive, and social aspects of MT.

However, MT evaluation methods also face several challenges and limitations. For example, human evaluation is costly, time-consuming, subjective, and inconsistent. Automatic evaluation is fast, cheap, objective, and consistent, but it may not capture the nuances and complexities of natural language and human communication. Moreover, different MT evaluation methods may have different assumptions, objectives, and perspectives, which may lead to conflicting or incomparable results. Therefore, it is important to select appropriate MT evaluation methods for different purposes and contexts, and to combine multiple MT evaluation methods to obtain a comprehensive and reliable assessment of MT quality and performance.

****

**Fig. 2** Machine translation evaluation methodology

**4 Various Lexical Automatic Machine Translation Evaluation Metrics**

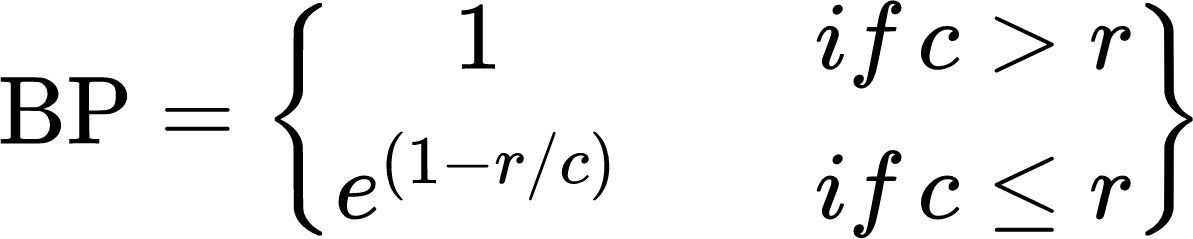
**4.1 Bilingual Evaluation Understudy (BLEU)**

BLEU (bilingual evaluation understudy) is a metric for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is considered to be the correspondence between a machine's output and that of a human: "the closer a machine translation is to a professional human translation, the better it is" [1]. BLEU was one of the first metrics to claim a high correlation with human judgements of quality [2][3], and remains one of the most popular automated and inexpensive metrics.

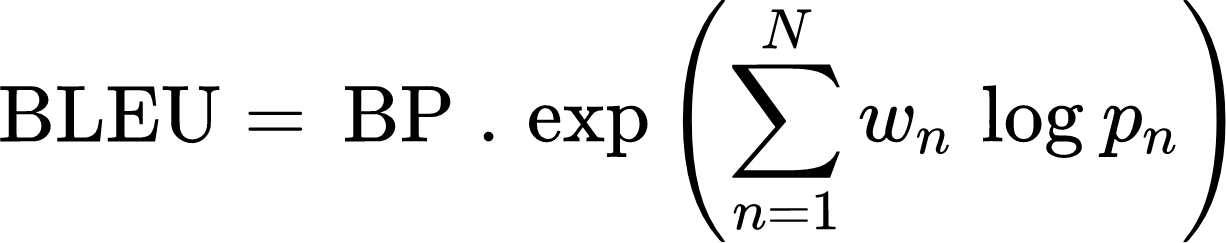
BLEU works by comparing the n-grams (sequences of n words) of the candidate translation with those of the reference translations. For each n-gram size (usually from 1 to 4), BLEU calculates a modified precision score, which is the ratio of matching n-grams to the total number of n-grams in the candidate translation. However, this precision score can be biased by repeating words or phrases in the candidate translation that are not in the reference translations. To avoid this, BLEU uses a clipping function that limits the number of times an n-gram can be counted based on its maximum frequency in any reference translation.

The modified precision scores for different n-gram sizes are then combined using a weighted geometric mean, which gives more weight to longer n-grams. The final BLEU score also incorporates a brevity penalty, which penalises candidate translations that are shorter than the reference translations. As shown in equation 1, The brevity penalty is calculated based on the ratio of the candidate translation length to the effective reference translation length, which is usually the closest length to the candidate translation among all reference translations. (However, in some versions of BLEU, such as NIST, the shortest reference translation length is used instead).

We compute the brevity penalty BP,

 (1)

Then,

 (2)

The BLEU score ranges from 0 to 1, with higher scores indicating more similar translations. However, it is not necessary to achieve a score of 1, as this would imply that the candidate translation is identical to one of the reference translations, which may not be possible or desirable. Moreover, adding more reference translations can increase the BLEU score, as there are more opportunities for matching n-grams.

BLEU has some limitations and challenges as a metric for evaluating machine translation quality. For instance, it does not account for grammatical correctness, semantic adequacy or stylistic variation. It also assumes that there is a single best translation for each source sentence, which may not be true in practice. Furthermore, it relies on exact word matching, which can miss synonyms, paraphrases or other linguistic variations that convey the same meaning. Additionally, it may not correlate well with human judgements at the sentence level, as humans may consider other factors besides lexical similarity.

Despite these drawbacks, BLEU is widely used as a simple and fast way to compare different machine translation systems or approaches. It can also provide feedback for improving machine translation models or identifying errors. However, it should not be used as the sole criterion for assessing translation quality, and it should be complemented by other metrics and human evaluations.

**Table 1** BLEU score computation

Source Sentence: He didn’t do his work on time.

MT Sentence: नहीं किया अपना कार्य समय से

Reference Sentence: उसने अपना कार्य समय से नहीं किया

Unigram Precision नहीं किया अपना कार्य समय से

1 1 1 1 1 1

Unigram precision = 6/6

Bigram Precision नहीं किया किया अपना अपना कार्य कार्य समय समय से 1 0 1 1 1

Bigram Precision = 4/5

Trigram Precision नहीं किया अपना अपना कार्य समय समय से नहीं

0 1 0

Trigram Precision = 1/3

BLEU score 0.7 (Using Eq. 2)

**4.2 National Institute of Standards and Technology (NIST)**

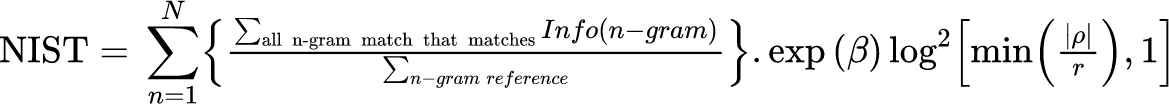
NIST (National Institute of Standards and Technology) is an organisation that evaluates the performance of natural language processing (NLP) systems through the development of benchmark datasets and standardised evaluation metrics.

The organisation has developed several widely-used benchmark datasets, such as the Multilingual Information Retrieval (MLIR) dataset and the TREC (Text Retrieval Conference) dataset, to evaluate the performance of various NLP tasks such as text classification, information retrieval, and question answering.

NIST differs from BLEU in two main aspects: how it calculates n-gram precision and how it applies the brevity penalty. N-gram precision is a measure of how well the n-grams (sequences of words) in the machine-translated text match those in the reference text. BLEU simply adds equal weight to each n-gram, regardless of how common or rare it is. NIST, on the other hand, also calculates how informative a particular n-gram is. That is to say, when a correct n-gram is found, the rarer that n-gram is, the more weight it will be given. For example, if the bigram "on the" is correctly matched, it will receive lower weight than the correct matching of bigram "interesting calculations", as this is less likely to occur.

The brevity penalty is a factor that penalises machine-translated texts that are too short compared to the reference texts. BLEU applies a harsh brevity penalty that can significantly lower the score if the translation length deviates from the reference length. NIST applies a more lenient brevity penalty that does not impact the overall score as much for small variations in translation length.

NIST score is computed using a formula that combines n-gram precision and brevity penalty. The formula is as follows:

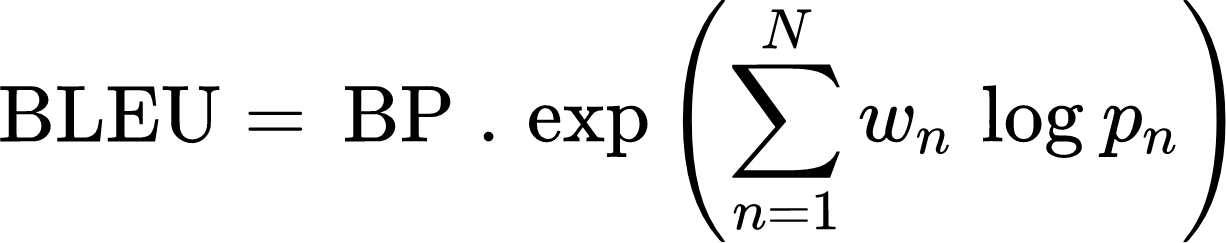


(3)

Where,

{"backgroundColorModified":false,"type":"$$","font":{"size":14,"color":"#000000","family":"Times New Roman"},"code":"$$\\text{Info(ngram)}\\;\\text{=}\\;\\text{Info(wi}_{1},\\,...\\,wi_{n}\\text{)}\\;\\text{=}\\;\\text{log}_{2}\\,\\tfrac{\\text{number}\\;\\text{of}\\;\\text{occourances}\\;\\text{of}\\;wi_{1}\\,...,\\,wi_{n-1}}{\\text{number}\\;\\text{of}\\;\\text{occourances}\\;\\text{of}\\,wi_{1}\\,...,wi_{n}}$$","id":"5","aid":null,"backgroundColor":"#ffffff","ts":1682710392511,"cs":"S8THQCrGVEBXh5YgiCL08Q==","size":{"width":561,"height":28}} (4)

Given that the brevity penalty factor in Equations 3 and 4 is set to 0.5, the average word count in the MT sentences is 2/3 the average word count in the reference sentences.

 (5)

where BP is the brevity penalty, w\_n is the weight for each n-gram order (based on information theory), p\_n is the modified n-gram precision for each n-gram order, and exp is the exponential function.

In addition to developing benchmark datasets, NIST has also developed standardised evaluation metrics such as precision, recall, and F1 score to assess the performance of NLP systems on these datasets. These metrics allow for objective comparisons between different NLP systems and help researchers to identify areas of improvement in their models.

Several research papers have used NIST datasets and evaluation metrics to evaluate the performance of their NLP systems. For example, the paper "Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds" by Wang et al. (2020) [5] used the TREC dataset to evaluate the performance of their question-answering system. Similarly, the paper "BERT-based Lexicalized Topic Models for Political Text Analysis" by Le et al. (2021) [6] used the MLIR dataset to evaluate the performance of their topic modelling system.

NIST plays a crucial role in the development and evaluation of NLP systems, providing standardised benchmark datasets and evaluation metrics to facilitate objective comparisons between different models and to drive advancements in the field.

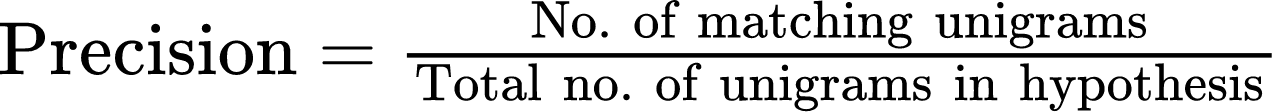
**4.3 Metric for Evaluation of Translation with Explicit Ordering (METEOR)**

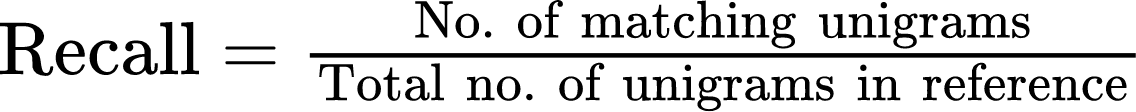
The Metric for Evaluation of Translation with Explicit Ordering (METEOR) is a machine translation evaluation metric that aims to measure the quality of machine translation results in a way that aligns with human judgments of translation quality. This is accomplished by comparing the output of the machine translation to one or more reference translations and assessing the output's quality using a mix of precision, recall, and alignment error.

METEOR takes into account synonyms, paraphrases, and word order in addition to exact word matches, unlike other machine translation assessment metrics that emphasise word matching. To find semantic distinctions between words and to take into account variations in word order, this is accomplished by using multiple linguistic resources such as WordNet and synonym sets.

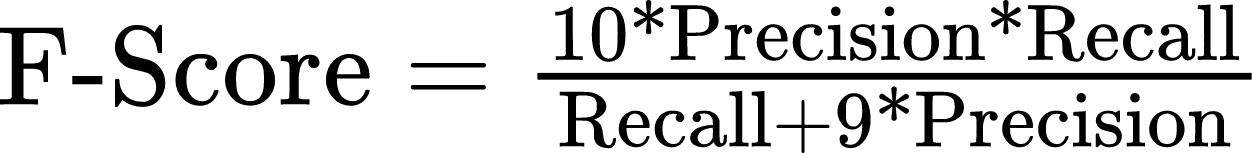
To compute the METEOR score, the machine translation output and reference translations are tokenized and stemmed to eliminate inflections and variations. The accuracy and recall of the alignments are then determined based on the alignment between the machine translation output and the reference translations. The final score is determined by taking the harmonic mean of accuracy and recall and applying an F-mean penalty to account for length disparities between the machine translation output and the reference translations.

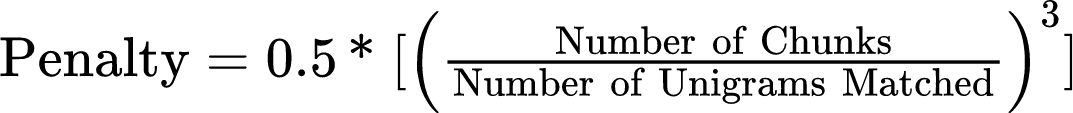
METEOR Score is calculated by:

 (6)

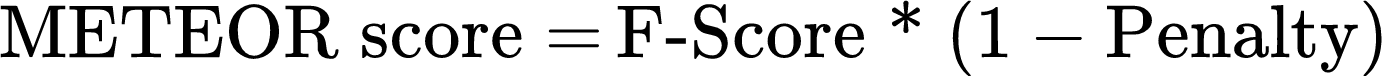
 (7)

Using eqn. 6 & 7, we can derive the F-score i.e., eqn. 8.

 (8)

 (9)

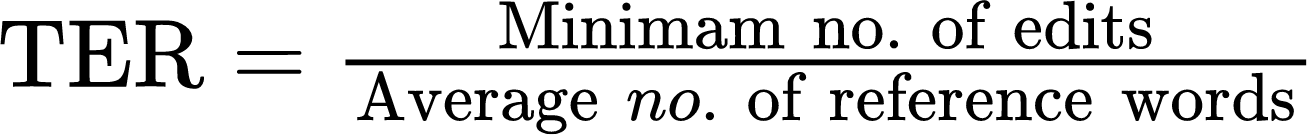
Now, using the eqn. 8 & 9, we can get the METEOR score.

 (10)

METEOR is commonly used in machine translation assessments, such as those conducted by the Conference on Machine Translation (WMT), and is recognised to correspond strongly with human translation quality ratings. Its capacity to account for word semantic distinctions in addition to word.

**4.4 Translation Error Rate (TER)**

TER (Translation Error Rate) is a metric commonly used in Natural Language Processing (NLP) to evaluate the quality of machine translation output. TER is a distance-based metric, which means it calculates the edit distance between the machine-generated translation and the reference translation. The edit distance is calculated by counting the number of operations required to transform the machine-generated translation into the reference translation. These operations can be insertion, deletion, substitution, or reordering of words. TER is defined as the minimum edit distance normalised by the total number of words in the reference translation. The lower the TER score, the better the machine translation. The TER score is calculated in Eq. 11.

 (11)

TER was first introduced by Snover et al. in 2006 in their paper "A Study of Translation Error Rate with Targeted Human Annotation" [7]. They proposed TER as an alternative to the widely used BLEU (Bilingual Evaluation Understudy) metric. Snover et al. demonstrated that TER had a stronger correlation with human judgments of translation quality than BLEU.

Since then, TER has been used in many research papers to evaluate the quality of machine translation output. For example, in their paper "Improving Lexical Choice in Neural Machine Translation" [8], Bawden et al. used TER to compare the performance of their model to other state-of-the-art machine translation models. Similarly, in their paper "Multi-Task Learning for Multimodal Machine Translation" [8], Libovický et al. used TER to evaluate the performance of their model.

**5 Experimental Setup**

**5.1 Dataset**

The Indic Corp corpus of data was developed by AI4Bharat, a nonprofit organisation devoted to the promotion of artificial intelligence (AI) technologies for Indian languages, and is the source of the dataset that we used. The Indic Corp expanded over the course of many months by locating and scraping hundreds of web sources, mostly news, magazines, and books, crawling news items, and blogging.

IndicCorp is one of the largest publicly available corpora for Indian languages. Additionally, it was used to train our publicly accessible models, which now perform cutting-edge on a range of tasks. The Corpus consists of a significant monolingual sentence-level corpus of 11 languages from two language families (Indo-Aryan and Dravidian), including Indian English.

**5.2 Translators**

In this experimental setup we have used two widely used and famous translators which are available on the internet and support Indic Language Translations.

These translators are:

1. Google Translate
2. Yandex Translate

And we have tried to apply machine translation evaluation metrics on the translations produced by these translators and compare both the translator’s based on their results.

**5.3 Test Dataset**

To compare the effectiveness of the Automatic Machine Translation Evaluation metrics we need a reliable and robust reference point. This need is fulfilled by assigning a human language expert to translate the given English language sentences to one of the three languages used for the purpose of this research paper, i.e., Hindi, Bengali and Telugu.

The translations made by the Human language expert are then stored in a Reference dataset that will be further used to evaluate the Automatic Machine Translation Evaluation Metrics.

**5.4 Evaluation Metrics**

Once we have our Datasets ready, we will use the Automatic Machine Translation Evaluation Metrics to evaluate the translations. The Evaluation Metrics we have used are Bilingual Evaluation Understudy (BLEU), National Institute of Standards and Technology (NIST), Translation Error Rate (TER) and Metric for Evaluation of Translation with Explicit Ordering (METEOR).

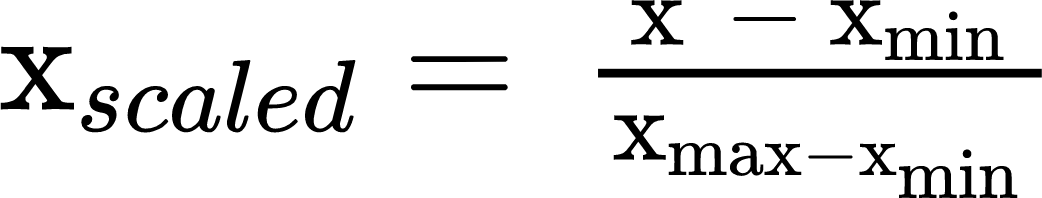
These metrics will take the Machine translated sentences and the sentences translated by the Human language expert to give output. These outputs will be in the range of 0 - 1, with 0 being the worst translation and 1 being the best possible translation.

**5.5 Analysis by Human Expert**

Based on the provided reference and the hypothesis data, a human expert will give scores on whether the hypothesis sentences align with the reference sentences in the range of 0 to 1. Here, the main criteria of the scores will be how close the hypothesis sentences match with the reference sentences. A score higher than 0.5 will indicate a high level of similarity with 1 indicating an absolute perfect translation and vice versa.

**5.6 Normalisation**

The Normalisation technique used here isMin-Max Normalisationwhich makes use of Minimum and Maximum values from a given set of values in order to scale down the value to a specified range, usually between 0 and 1. With the help of scaling we were able to improve the evaluation metrics which are somewhat sensitive to certain input features present in the dataset. The min-max normalisation formula is provided in eqn. 12.

 (12)

**5.7 Pearson Correlation**

Pearson Correlation Coefficient (r) is a statistical measure of the linear relationship between two quantitative variables. It ranges between -1 to +1 with -1 indicating that there is a negative correlation, 0 indicates that there is no correlation and +1 indicates that there is a positive correlation between the two variables.

We have used the Pearson Correlation Coefficient to compare the results of the different Automatic Machine Translation Evaluation Metric Scores and the scores provided by the human language expert.

**6 Result**

1. **Hindi Language**

Table 2 Pearson Correlation of Hindi Language using MT and human annotation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Automatic Machine** | **BLEU** | **METEOR** | **NIST** | **TER** |
| **Google** | **0.136** | **0.340** | **0.233** | **0.009** |
| **Yandex** | **0.026** | **0.116** | **0.028** | **0.258** |

1. **Bengali Language**

Table 3 Pearson Correlation of Bengali Language using MT and human annotation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Automatic**  **Machine** | **BLEU** | **METEOR** | **NIST** | **TER** |
| **Google** | **0.205** | **0.195** | **0.282** | **0.008** |
| **Yandex** | **0.203** | **0.142** | **0.244** | **0.387** |

1. **Telugu Language**

Table 4 Pearson Correlation of Telugu Language using MT and human annotation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Automatic**  **Machine** | **BLEU** | **METEOR** | **NIST** | **TER** |
| **Google** | **0.194** | **0.289** | **0.246** | **0.152** |
| **Yandex** | **0.191** | **0.226** | **0.061** | **0.140** |

As shown in the tables 2, 3 & 4 above, among the Machine Translation Evaluation Metrics METEOR and NIST have outperformed BLEU and TER in most cases. Here, the closer the value is to 1, the better it is. Also, we can see from the tables that the translations from Google translate are consistently of a better quality than the translations produced by Yandex translate, it is especially true for languages other than Hindi which has a larger corpus given it is the most widely spoken language in the subcontinent and hence have better translations. Bengali and Telugu in comparison have a smaller corpus which makes them comparatively harder to translate which is reflected in the tables above.

**7 Discussion and Conclusion:**

In conclusion, the study aimed to explore various lexical automatic machine translation evaluation metrics for Indic languages. Several famous metrics such as BLEU, METEOR, TER, and NIST were compared and evaluated on their effectiveness in assessing the quality of the machine-translated text. The evaluation was carried out on multiple datasets, and the results were analysed to determine which metric performed better.

The findings revealed that BLEU performed relatively well on most datasets and was the most widely used metric for evaluating machine translation systems. However, the study also highlighted the limitations of BLEU and the need to use multiple metrics for a more comprehensive evaluation of machine translation quality.

The study recommends using a combination of BLEU, METEOR, and TER metrics to evaluate machine translation systems for Indic languages. This approach provides a more comprehensive evaluation and a better understanding of the quality of the machine-translated text. Additionally, the study suggests that future research should focus on developing new evaluation metrics specifically for Indic languages to improve the accuracy and effectiveness of machine translation evaluation.

**7.1 Future Work**

In this paper, we have presented a comparative analysis of different lexical automatic machine translation evaluation metrics for indic languages. We have evaluated the performance of these metrics on three different datasets of English - Hindi, English - Bengali and English - Telugu translation pairs.

As a future work, we plan to extend our study to other Indic languages and domains. We also aim to incorporate syntactic and pragmatic features to capture the structural and contextual aspects of translation quality. Furthermore, we intend to explore the correlation of the metrics with human judgments and conduct a user study to validate its usefulness and reliability.

Another possible direction for future work is to explore the use of neural machine translation (NMT) models for Indic languages and evaluate them using Lexical Automatic Machine Translation LAMT metrics. NMT models are based on deep learning techniques that learn to translate from large parallel corpora without relying on explicit rules or features. NMT models have shown remarkable results for high-resource languages, but their performance may degrade for low-resource languages, such as many Indic languages. Moreover, NMT models may face challenges in handling the long and complex sentences, the domain mismatch, and the data sparsity of Indic languages. Therefore, it is interesting to examine how NMT models perform for Indic languages and how they compare with statistical machine translation (SMT) models using LAMT metrics. We hope that our work will contribute to the advancement of machine translation research and evaluation for indic languages.

**References**

[1] Papineni et al., "BLEU: a Method for Automatic Evaluation of Machine Translation", ACL 2002.

[2] Doddington et al., "Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics", HLT 2002.

[3] Callison-Burch et al., "Re-evaluating the Role of BLEU in Machine Translation Research", EACL 2006.

[4] Wang, S., Liu, C., & Wei, F. (2020). Learning to Answer by Learning to Ask: Getting the Best of GPT-2 and BERT Worlds. arXiv preprint arXiv:2010.00459.

[5] Le, H., Nguyen, N. T., & Nguyen, T. H. (2021). BERT-based Lexicalized Topic Models for Political Text Analysis. arXiv preprint arXiv:2101.06379.

[6] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation error rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (pp. 223-231).

[7] Bawden, R., Gašić, M., & Mrkšić, N. (2018). Improving lexical choice in neural machine translation. arXiv preprint arXiv:1808.09381.

[8] Libovický, J., Kopeček, I., Bojar, O., & Maršík, L. (2018). Multi-task learning for multimodal machine translation. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 794-805).

[9] Chauhan, Shweta, and Philemon Daniel. "A comprehensive survey on various fully automatic machine translation evaluation metrics." Neural Processing Letters (2022): 1-55.

[10] Wong, B.T. and Kit, C., 2012, July. Extending machine translation evaluation metrics with lexical cohesion to document level. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 1060-1068).

[11] Chauhan, S., Saxena, S., & Daniel, P. (2022, July 29). Enhanced unsupervised neural machine translation by cross lingual sense embedding and filtered back-translation for morphological and endangered Indic languages. Taylor & Francis Online. Retrieved May 1, 2023, from <https://www.tandfonline.com/doi/full/10.1080/0952813X.2022.2135612>

[12] S. Lee et al., “A Survey on Evaluation Metrics for Machine Translation,” Mathematics, vol. 11, no. 4, p. 1006, Feb. 2023, doi: 10.3390/math11041006.

[13] Frisoni, G., Carbonaro, A., Moro, G., Zammarchi, A., & Avagnano, M. (2022, October). NLG-Metricverse: An End-to-End Library for Evaluating Natural Language Generation. In Proceedings of the 29th International Conference on Computational Linguistics (pp. 3465-3479).

[14] Rocha, Á., Adeli, H., Reis, L. P., & Costanzo, S. (Eds.). (2018). Trends and Advances in Information Systems and Technologies: Volume 1.

[15] G. Singh Sodhi and J. Singh Sodhi, "A Robust Invariant Image-Based Paper-Currency Recognition Based on F-kNN," 2021 International Conference on Intelligent Technology, System and Service for Internet of Everything (ITSS-IoE), Sana'a, Yemen, 2021, pp. 1-6, doi: 10.1109/ITSS-IoE53029.2021.9615287.

[16] Forcada, M. L., Depraetere, H., & Vandeghinste, V. (2011). Proceedings of the 15th Annual conference of the European Association for Machine Translation. In Proceedings of the 15th Annual conference of the European Association for Machine Translation.

[17] Kos, K. (2008). Adaptation of new machine translation metrics for Czech (Doctoral dissertation, Bachelor’s thesis, Charles University in Prague).

[18] Sun, M., Liu, Y., Liu, Z., & Zhang, M. (2013). Chinese computational linguistics and natural language processing based on naturally annotated big data. Springer.

[19] A. Anand, N. K. Trivedi, V. Gautam, R. G. Tiwari, D. Witarsyah and A. Misra, "Applications of Internet of Things(IoT) in Agriculture: The Need and Implementation," 2022 International Conference Advancement in Data Science, E-learning and Information Systems (ICADEIS), Bandung, Indonesia, 2022, pp. 01-05, doi: 10.1109/ICADEIS56544.2022.10037505.

[20] Rocha, Á., Adeli, H., Reis, L. P., & Costanzo, S. (Eds.). (2018). Trends and Advances in Information Systems and Technologies: Volume 1.

[21] Olive, J., Christianson, C., & McCary, J. (Eds.). (2011). Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation. Springer Science & Business Media.

[22] Zhang, Y., & Vogel, S. (2010). Significance tests of automatic machine translation evaluation metrics. Machine Translation, 24, 51-65.

[23] Rodriguez-Torrealba, R., Garcia-Lopez, E., & Garcia-Cabot, A. (2022). End-to-End generation of Multiple-Choice questions using Text-to-Text transfer Transformer models. Expert Systems with Applications, 208, 118258.

[24] Zhao, D., Li, J., Feng, Y., & Ji, H. (2015). Natural Language Processing and Chinese Computing. Springer International Publishing.

[25] Sathiyamurthy, K., Panimalar, D., & Pandian, S. L. (2014, December). Multilingual acquiring of e-content definition based on universal networking language. In 2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE) (pp. 240-244). IEEE.

[26] House, J. (1997). Translation quality assessment: A model revisited. Gunter Narr Verlag.

[27] Ke, X., & Ma, Q. (2014). Study on an impersonal evaluation system for English-Chinese translation based on semantic understanding. Perspectives, 22(2), 242-254.

[28] Dey, S., Vinayakarao, V., Gupta, M., & Dechu, S. (2022, May). Evaluating commit message generation: to BLEU or not to BLEU?. In Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results (pp. 31-35).

[29] Condon, S., Arehart, M., Parvaz, D., Sanders, G., Doran, C., & Aberdeen, J. (2012). Evaluation of 2-way Iraqi Arabic–English speech translation systems using automated metrics. Machine translation, 26, 159-176.

[30] Joty, S., Guzmán, F., Màrquez, L., & Nakov, P. (2017). Discourse structure in machine translation evaluation. Computational Linguistics, 43(4), 683-722.

[31] Livingstone, R. W., Elder, M. K., Singh, A., Westlake, C. M., Tate, W. P., Abraham, W. C., & Williams, J. M. (2021). Secreted amyloid precursor protein-alpha enhances LTP through the synthesis and trafficking of Ca2+-permeable AMPA receptors. Frontiers in Molecular Neuroscience, 14, 660208.

[32] Dey, S., Vinayakarao, V., Gupta, M., & Dechu, S. (2022, May). Evaluating commit message generation: to BLEU or not to BLEU?. In Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: New Ideas and Emerging Results (pp. 31-35).

[33] Gelbukh, A. (Ed.). (2006). Computational Linguistics and Intelligent Text Processing: 7th International Conference, CICLing 2006, Mexico City, Mexico, February 19-25, 2006, Proceedings (Vol. 3878). Springer.

[34] Gelbukh, A. (2009). Computational linguistics and intelligent text processing. Springer Berlin/Heidelberg..

[35] Clark, A., Fox, C., & Lappin, S. (Eds.). (2012). The handbook of computational linguistics and natural language processing (Vol. 118). John Wiley & Sons.

[36] Singh, J., & Gupta, V. (2016). Text stemming: Approaches, applications, and challenges. ACM Computing Surveys (CSUR), 49(3), 1-46.

[37] Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., ... & Lim, H. (2023). A Survey on Evaluation Metrics for Machine Translation. Mathematics, 11(4), 1006.

[38] Pérez-Ortiz, J. A., Sánchez-Martínez, F., Esplà-Gomis, M., Popović, M., Rico, C., Martins, A., ... & Forcada, M. L. (2018). Proceedings of the 21st Annual Conference of the European Association for Machine Translation: 28-30 May 2018, Universitat d'Alacant, Alacant, Spain.

[39] Lopez, A. (2008). Statistical machine translation. ACM Computing Surveys (CSUR), 40(3), 1-49.

[40] Bilbao, V. D., Lopes, J. P., & Ildefonso, T. (2005, December). Measuring the impact of cognates in parallel text alignment. In 2005 portuguese conference on artificial intelligence (pp. 338-343). IEEE.

[41] Sharma, D., Dhiman, C., & Kumar, D. (2023). Evolution of visual data captioning Methods, Datasets, and evaluation Metrics: A comprehensive survey. Expert Systems with Applications, 119773.

[42] Nakanishi, A. (1990). Writing Systems of the World. Tuttle Publishing.

[43] Sanders, G. A., Weiss, B. A., Schlenoff, C., Steves, M. P., & Condon, S. (2013). Evaluation methodology and metrics employed to assess the TRANSTAC two-way, speech-to-speech translation systems. Computer Speech & Language, 27(2), 528-553.

[44] Takakura, S., Han, D., & Furugori, T. (2005). Recognition and utilization of clausal relations in complex sentences for improving the performance of machine translation systems. Journal of Quantitative Linguistics, 12(2-3), 239-261.

[45] Dakwale, P. (2020). Strategies for effective utilization of training data for machine translation. Universiteit van Amsterdam.

[46] Kos, K. (2008). Adaptation of new machine translation metrics for Czech (Doctoral dissertation, Bachelor’s thesis, Charles University in Prague).

[47] Bizzoni, Y., Teich, E., i Bonet, C. E., & van Genabith, J. (2021). Proceedings of the First Workshop on Modelling Translation-Translatology in the Digital Age.

[48] Ngoc Tien Le. Advanced Quality Measures for Speech Translation. Computation and Language [cs.CL]. Université Grenoble Alpes, 2018. English. ffNNT : 2018GREAM002ff. Fftel-01891892.

[49] Vyas P, Vyas G, Dhiman G. RUemo—The Classification Framework for Russia-Ukraine War-Related Societal Emotions on Twitter through Machine Learning. Algorithms. 2023; 16(2):69. <https://doi.org/10.3390/a16020069>.

[50] Home. (n.d.). YouTube. Retrieved May 3, 2023, from <https://www.localisation.ie/sites/default/files/best-thesis/OEDCMTT-BTA14.pdf>

[51] Chunseong Park, C., Kim, B., & Kim, G. (2017). Attend to you: Personalized image captioning with context sequence memory networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 895-903).

[52] Bawden, R. (2018). Going beyond the sentence: Contextual machine translation of dialogue (Doctoral dissertation, Université Paris-Saclay (ComUE)).

[53] Pino, J., Waite, A., & Byrne, W. (2012). Simple and efficient model filtering in statistical machine translation. The Prague Bulletin of Mathematical Linguistics, 98, 5.

[54] Schmidt, C. (2016). Handling multimodality and scarce resources in sign language machine translation (Doctoral dissertation, Dissertation, RWTH Aachen University, 2016).

[55] Association for Computational Linguistics. (2014). Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics.

[56] Home. (n.d.). YouTube. Retrieved May 3, 2023, from https://aclanthology.org/W/W14/W14-33.pdf

[57] Mondal, S. K., Zhang, H., Kabir, H. D., Ni, K., & Dai, H. N. (2023). Machine translation and its evaluation: a study. Artificial Intelligence Review, 1-90.

[58] Home. (n.d.). YouTube. Retrieved May 3, 2023, from https://www.gabormelli.com/RKB/index.php?mobileaction=toggle\_view\_mobile&title=National\_Institute\_of\_Standards\_And\_Technology\_%28NIST%29\_Metric

[59] Castilho, S. (2016). Measuring acceptability of machine translated enterprise content (Doctoral dissertation, Dublin City University).

[60] Metrics for Trustworthy AI - OECD.AI. (n.d.). OECD AI Policy Observatory. Retrieved May 3, 2023, from <https://oecd.ai/en/catalogue/metrics>

[61] Livingstone, R. W., Elder, M. K., Singh, A., Westlake, C. M., Tate, W. P., Abraham, W. C., & Williams, J. M. (2021). Secreted amyloid precursor protein-alpha enhances LTP through the synthesis and trafficking of Ca2+-permeable AMPA receptors. Frontiers in Molecular Neuroscience, 14, 660208.

[62] Pérez-Ortiz, J. A., Sánchez-Martínez, F., Esplà-Gomis, M., Popović, M., Rico, C., Martins, A., ... & Forcada, M. L. (2018). Proceedings of the 21st Annual Conference of the European Association for Machine Translation: 28-30 May 2018, Universitat d'Alacant, Alacant, Spain.

[63] Nakanishi, A. (1980). Writing Systems of the World: Alphabets, Syllabaries, Pictograms (Sekai no Moji). Rutland, VT: Charles E. Tuttle Co.

[64] How Many Official Language In India (Updated 2023). (2023, April 4). Japanese Language Interpreter. Retrieved May 3, 2023, from <https://japaneselanguageinterpreter.com/blog/how-many-official-language-in-india/>