International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Performance analysis of MT tools through Hindi to English web query translation

Amit Asthana*[1], Sanjay K. Dwivedi[2]

Submitted: 01/10/2024 Revised : 01/11/2024 Accepted:10/11/2024

Abstract: Query translation plays crucial and important role in cross-lingual information retrieval (CLIR) systems, where retrieval efficiency is closely tied with translation accuracy. Indian languages require an effective and efficient machine translation (MT) tool to effectively transform query intent into other language. As machine translation becomes increasingly prevalent in the translation industry, understanding its quality is gaining greater importance. However, the focus on the acceptance of MT output based on performance, and more importantly, how acceptable it is to human translators, has been relatively limited. The complexity of MT arises from factors such as words with multiple meanings, sentences with various interpretations, and differing grammatical structures across languages. This complexity is further intensified by the lack of structural constraints and the presence of ambiguity, particularly in the case of web queries. The goal of this work is to evaluate the accuracy of free online MT tools in translating Hindi web queries. The accuracy has been measured using several metrics, including METEOR, BLEU, NIST, hLEPOR, CHRF, and GLEU. Our results show that translation accuracy is higher for longer queries compared to shorter ones. Among the translators tested, Google Translate performed the best, while Systran performed the worst, with a performance gap of more than 42% between the two. The present research work assesses the performance and effectiveness of the popular MT tools for Hindi to English query translation

Keywords: *machine translation, web query translation, information retrieval, online translators, translation quality*

1. Introduction

Machine Translation (MT) tools automate the translation process, and significant improvements in their accuracy have been made over the years. However, evaluating their quality across diverse parameters remains a challenging task. Most quality assessments of MT outputs have focused on the sentence level, with little attention paid to how a larger text is structured. This is evident in metrics like BLEU [1], METEOR [2], and TER [3], which primarily evaluate translations at the sentence level. As a result, translated texts may sometimes lack readability and coherence.

Sentence-based metrics, which assess translations segment by segment, often fail to capture document-level coherence because they ignore the broader context of the entire text. For instance, while a machine translation system may produce grammatically correct individual sentences, these sentences may not logically flow when combined. Additionally, these metrics overlook lexical consistency between sentences, leading to disjointed translations that may not convey the intended meaning of the original text. Consequently, even if each sentence receives a high score for its accuracy, the overall document may still fail to effectively communicate the original content. For MT users, the general meaning and coherence of the text are more important than the grammatical accuracy of each sentence, making document-level accuracy crucial [4].

While document-level accuracy is essential, sentence-level accuracy also plays a key role, especially in cross-lingual information retrieval (CLIR) research. In CLIR, the effectiveness of retrieval depends heavily on the performance of the MT system. Unlike conventional sentence or paragraph translation, translating web queries poses unique challenges due to the lack of context. Web queries are typically short and may not provide sufficient information, making accurate translation more difficult. Therefore, the focus in CLIR is on translating web queries effectively, despite these challenges.

1.1. MT Evaluation Metrics

The MT evaluation metric BLEU is the first to show a substantial correlation with human judgement. It, along with its close variant, the NIST metric, became the standard way for optimizing statistical machine translation systems. While BLEU is remarkably simple and effective tool, recent research has demonstrated that a number of next-generation metrics can outperform BLEU in terms of human evaluation correlation (Callison-Burch, 2009),[6]. [3] proposed TER that measures the number of edits (insertions, deletions, and substitutions) required to transform the generated text into the reference text and considers the structure and grammar of the sentences. Word insertion, substitution, deletion, and changes to the sequence of words or phrases are all considered as edits [7]. PORT is a metric that evaluates MT tuning automatically by considering a wide range of features including precision, recall, strict brevity penalty, ordering metric, and redundancy penalty [8]. The quality of MT in older methods was determined by how well the output sentences of MT matched the human-translated texts. It uses a neural network to predict human judgments of translation quality and considers a wide range of features such as lexical semantics, fluency, and faithfulness to the source text [9]. Fonseca et al. [10] showed that it is possible to attain a high level of

[1] *Babasaheb Bhimrao Ambedkar University, Lucknow – 226025, India*
ORCID ID : 0000-0002-4244-0309
Email: aamitonline@gmail.com
[2] *Babasaheb Bhimrao Ambedkar University, Lucknow – 226025, India*
ORCID ID : 0000-0003-3839-3017
Email: skd200@yahoo.com
** Corresponding Author Email: aamitonline@gmail.com*

correlation with human judgement without using any reference translation. AdaBLEU, a variant of BLEU, adapts the n-gram order based on the length of the reference text and considers the lexical and syntactical aspects of the words and is helpful for morphologically rich languages to make up for the lack of tight string matching in BLEU [11], [12]. Metrics considered for the evaluation of the translation quality are explained below:

1.1.1. BLEU

The foundation of BLEU (Bi-Lingual Evaluation Understudy) is n-gram match precisions and was developed in 2002 [1]. It is the most extensively used metric for MT evaluation because of its claim of strong correlation with human rankings of MT output. It works on per-word algorithm that judges translations at the segment level.

The n-gram precision is thus defined as

$$Pn=\frac{|BNGnR \cap BNGnT|}{|BNGnT|}$$

Given a reference text R and a translation candidate T, all n-grams included in R and T for n=1, 2, 3, 4, and labelled as BNGnR and BNGnT respectively.

BLEU incorporates a brevity penalty to address the lack of a recall measure, and hence the tendency to produce brief translations, as

$$BP = \begin{cases} 1 \; if \; c > r \\ e^{1-r/c} \; if \; c \leq r \end{cases}$$

Where c and r mean the sentence length of output candidate translation and reference translation respectively.

The metric is finally defined as:

$$BLEU(R,T) = BP \times \sqrt[4]{P1P2P3P4}$$

1.1.2. Metric for Evaluation of Translation with Explicit Ordering (METEOR)

It is an automatic metric which evaluates the MT output by comparing it with one or more reference translations with improved correlation with Human Judgments [2]. In contrast to BLEU, which concentrates entirely on precision-based features, METEOR emphasizes recall in addition to precision, as it has been proven by various metrics to be crucial for strong correlation with human assessments. It calculates a score based on explicit word-to-word matches between the translation and a reference translation when evaluating a translation. If more than one reference translation is available, then the given translation is scored independently against each reference, with the highest score being recorded.

1.1.3. National Institute of Standards and Technology (NIST)

NIST is an evaluation metric that measures the similarity between the generated translation and the reference translation at the n-gram level. It is similar to BLEU, but it weights the n-grams based on their frequency in the reference text [13]. NIST is used in the annual Machine Translation Evaluation (MT) organized by the National Institute of Standards and Technology. NIST does not adequately consider the word order in sentences or the language's grammatical structure. NIST is used in the annual Machine Translation Evaluation (MT) organized by the National Institute of Standards and Technology.

1.1.4. The LEPOR_v3.1 System (hLepor)

hLepor is an enhanced version of LEPOR i.e. Length Penalty, Precision, n-gram Position difference Penalty and Recall [14] that combines the sub factors using weighted mathematical harmonic mean instead of the simple product value. It considers the linguistic features, such as the POS of the words.

1.1.5. Character n-gram F-score (CHRF)

CHRF is an evaluation metric used to assess the similarity between two texts at the character level. It is commonly employed in natural language processing tasks where word boundaries may not be well-defined or in cases where the exact arrangement of characters matters more than individual words.

$$F - score = \frac{2*(\text{Precision} * \text{Recall})}{(\text{Pr}\,ecision + \text{Recall})}$$

1.1.6. GLEU

GLEU stands for "Google-BLEU", which is a variant of the BLEU (Bilingual Evaluation Understudy) metric commonly used to evaluate the quality of machine translation. GLEU was proposed as a modification to the original BLEU metric, with the aim of addressing some of its limitations [15].

GLEU calculates the precision of n-grams in the candidate translation that match the reference translation, and also considers the precision of the reference n-grams in the candidate translation. It then combines these two scores to give a final score. The main difference between GLEU and BLEU is that GLEU uses a different length normalization method, which is intended to better handle short translations. In general, GLEU has been found to correlate well with human judgments of translation quality, and has been shown to outperform BLEU in some settings. A comparison of these metrics in terms of their advantages and disadvantages has been shown in Table 1.

Table 1. Pros & cons of the evaluation metrics

Metric	Type	Advantages	Disadvantages
BLEU	n-gram	Easy to compute and widely used	Limited to n-gram matching, doesn't consider word order
METEOR	Hybrid	Incorporates semantic similarity, handles synonyms	Requires language-specific resources, computationally expensive
NIST	n-gram	Incorporates frequency-based weighting, easy to compute	Can produce unnatural translations, doesn't consider word order
TER	Edit distance	Measures fluency, handles reordering and paraphrasing	Not suited for agglutinative languages, can be expensive

PORT	Edit distance	Sensitive to fluency and meaning, considers word order	Requires training data, not widely used
hLEPOR	Hybrid	Incorporates syntactic and semantic analysis, handles ellipsis and co-reference	Limited to specific languages, requires additional resources
CHRF	Character level	Incorporates character-level information, useful for languages with complex morphology and Can handle spelling variations and minor word order differences.	Sensitive to tokenization errors, as it relies heavily on character-level information. May not fully capture semantic similarity between sentences.
GLEU	Adaptive	captures some level of lexical and syntactic similarity and measures fluency, adequacy	Doesn't consider word order, which may lead to inaccurate evaluations, especially for languages with free word order.

Evaluation metrics has its strengths and weaknesses as the metrics like BLEU, NIST, and TER provide valuable insights but fail to capture word order and syntactical structure, hence, a combination of metrics tailored to specific contexts is crucial for comprehensive MT evaluation and system optimization. In this work we have selected some popular free online MT tools for evaluation of their performance on Hindi web queries. The purpose of this work is to address the limitations of current evaluation measures in machine translation (MT) and emphasize the importance of assessing translation quality. The target audience for the work includes researchers, practitioners, and professionals in machine translation, and information retrieval, particularly those interested in CLIR research involving Hindi language.

2. Related Work

The assessment of translation quality has long been a crucial area of study in the field of translation, and interest in it has grown as a result of the industry's rising need for translation. Numerous methods of MT evaluation have also been developed in response to the lack of general agreement over what constitutes a "good" translation.

In early machine translation (MT) evaluation studies, [16] utilized Babel Fish to assess translation quality across multiple languages. Their study focused on translating English proverbs into various languages using Babel Fish, aiming to identify any errors or inconsistencies in the translations produced by MT systems. Following this, [17] discussed the limitations of an English-Thai MT, emphasizing the importance of addressing lexical disambiguation issues to enhance translation performance. Based upon these findings the National Institute of Standards and Technology conducted a comprehensive comparison of 22 MT systems, concluding that Google Translator consistently performed well across translations from Arabic and Chinese to English. [18] and [19] evaluated web translation systems, with a particular focus on translations from English to Spanish using various platforms such as LogoMedia, Systran, and PROMT translators. Their studies highlighted variations in translation quality among different systems, with Systran often producing unsatisfactory translations. further explored the ranking of free machine translation systems, specifically from French to English. Despite Google Translate emerging as the top performer, the study also pointed out significant issues with lexical ambiguity and idiomatic translations. Subsequent studies by [20] and [21] highlighted the dominance of Google Translate across different evaluations and languages. [22] conducted a user survey, revealing that Google Translate was preferred for longer texts and Bing Translator for the shorter texts. [23] and [24] confirmed these findings by further validated Google Translator's superiority in accuracy, particularly in translations between English and other languages. Finally, [25] conducted a comprehensive evaluation of various MT tools using German and English, with Google Translate outperforming other systems in terms of translation accuracy. Neural network-based methods, such as neural machine translation (NMT) have also shown promising results in improving translation quality by capturing complex linguistic patterns and context dependencies, leverage advanced architectures to improve accuracy and fluency [26]. Neural networks are increasingly being employed to develop more sophisticated evaluation metrics which leverage neural network architectures to analyze translations at various levels, from sentence-level fluency to document-level coherence. Human-in-the-loop [27] approaches incorporate human feedback into the evaluation process, allowing for more nuanced assessments of translation quality by capturing aspects that automated metrics may overlook. [28] proposed the Word Predicted and Substituted (WPS) method for query reformulation using word2vec focusing on the Indonesian QA System (IQAS). They conducted two experiments one measuring the impact of reformulated queries on search engine results (E-1), and assessing the effectiveness of reformulated queries in IQAS (E-2). The experimental evaluation result showed high effectiveness in providing additional information (E-1: 81%) and enhancing IQAS search capabilities (E-2: 274.74%). [29] introduces a self-supervised query reformulation (SSQR) method that does not require parallel query data. It processes reformulation as a masked language modeling task on unannotated query data, extending T5(a sequence-to-sequence Transformer based model) with a new pre-training objective called corrupted query completion. SSQR identifies expansion locations and utilizes T5 to generate content

based on information gain. Evaluation shows SSQR outperforms unsupervised methods significantly and competes well with supervised ones and generating useful and natural-sounding reformulated queries. [30] proposed RLQR (Reinforcement Learning for Query Reformulations) to generate reformulations that maximize product coverage. They compared it with supervised generative models and strong RL-based methods showing 28.6% increase in coverage compared to standard models. RLQR outperformed state-of-the-art benchmarks significantly and also demonstrated increased coverage on an external Amazon shopping dataset. User surveys may lack fairness, as user preferences can be subjective and influenced by factors unrelated to translation quality. System comparisons are informative but may overlook small differences in translation quality and user preferences [35], [36]. Using a combination of evaluation methods, provides a more comprehensive understanding of machine translation performance. This multidimensional approach ensures a balanced assessment by identifying both quantitative and qualitative details, leading to improvements in machine translation systems.

Studies indicate that the Google Translate consistently performs well across various evaluations, emerging as a leading machine translation system. While some studies indicate variations in translation quality among different systems, Google Translate generally demonstrates superior performance in terms of accuracy. User surveys indicate that preferences for selecting MT system may vary depending on factors like translation task and text length, Google Translate often preferred for longer texts. Overall, while Google Translate emerges as a leading MT system in many studies, the variability in results highlights the importance of considering different evaluation criteria in assessing MT system performance. This paper compares and analyses the efficiency of the translators for web queries over a set of metrics for the purpose of effective evaluation of different aspects of translation quality which provides a more comprehensive and balanced assessment of translation quality as each evaluation metric has its strengths and weaknesses, and by combining them, we can compensate for individual limitations and gain a more accurate understanding of the performance of machine translation systems. In this work we have combined multiple metrics for the purpose of translation quality evaluation mitigating the limitations of the individual metrics and adding more dimensions to the evaluation. This work aims to improve the accuracy and reliability of machine translation systems, ultimately enhancing their usability and effectiveness for CLIR by addressing the limitations of online free translators particularly for the translation of web queries.

3. Experimental Analysis

We collected 50 Hindi web queries (from various sources like FIRE 2008, 2010, 2011)[1] and some of these queries consist out of vocabulary (OOV) words collected from a popular Hindi newspaper "Dainik Jagran"[2] out of which 20 are short and 30 are the longer ones. The queries are collected in a way that there are 20 short and 30 long queries. The queries have been categorization according to their size. Queries having three or less words are termed short query and rest are longer ones. We used popular free online translators i.e. google translator[3], Microsoft bing[4], translate dict[5], translate.com[6], Collins translator[7], imTranslator[8], Yandex translate[9] and systran translator[10] for the purpose of translation of web queries from Hindi to English. Queries have been translated as collected from the respective sources regardless of their content by simply ignoring any OOV word present in the query. Table 3 shows the initial queries and their translations performed by the translators.

We took help from linguistics, who are the experts in linguistic studies and belong to the respective university department, for the purpose of obtaining reference translation of the queries. The quality of the translation is then assessed with evaluation metrics i.e. BLEU, NIST, METOR, hLepor, CHRF and GLEU. Each selected evaluation metric offers unique advantages for evaluating translation quality. BLEU is generally used for its simplicity and effectiveness in evaluating lexical similarity. NIST provides a unified view by considering both precision and recall with respect to the reference. METEOR utilizes linguistic features such as stemming, synonymy, and word order allowing METEOR to capture instances where the translated output fails to convey the intended meaning accurately thus providing assessment in terms of semantic similarity and adequacy. hLEPOR extends BLEU by considering additional factures like word order. CHRF offers robustness against variations in morphological changes, whereas GLEU focuses on fluency and syntactic accuracy. The metrics and their specific focus have been shown in Table 2.

Table 2. The Evaluation metrics with their specific focus

Metrics	Specific Focus
BLEU	Lexical Similarity
NIST	Informativeness
METEOR	Semantic Equivalence, Adequacy
hLepor	Word Order
CHRF	Robustness
GLEU	Fluency and accuracy

Translators' performance has been evaluated and analysed based on the individual metrics score as well as the average metric score. A comparative analysis of translation quality for short and long queries and also for the queries irrespective of their categories has been performed. The significance of performance differences among the translators based upon individual and average metric scores for translation performed by the respective translators has been analysed and discussed in the next section.

Table 3. Sample web queries with translations

[1] http://fire.irsi.res.in
[2] https://epaper.jagran.com/epaper
[3] https://translate.google.co.in
[4] https://www.bing.com/translator
[5] https://www.translatedict.com
[6] https://www.translate.com/machine-translation
[7] https://www.collinsdictionary.com/us/translator
[8] https://imtranslator.net/translation/hindi/to-english/translation
[9] https://translate.yandex.com
[10] https://www.systran.net/en/translate

International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Original query	टाटा की नैनो गाड़ी	आत्मा का स्वरुप	कर्मभूमि का महत्व	देशहित के प्रकल्प	ज्ञानीजन की पहचान	2जी स्पेक्ट्रम घोटाले में ए. राजा	मुंबई ताज हमला	जॉर्ज बुश का आतंकवाद दमन अभियान	चक्रवाती तूफ़ानों का तांडव	नजरबन्द नेता सू-की
Reference translation	tata's nano car	form of soul	importance of work field	projects of national interest	identity of the wise	a. raja in 2g spectrum scam	Mumbai taj attack	george bush's terrorism suppression campaign	fury of cyclonic storms	detained leader suukyi
google	tata nano car	form of soul	importance of land	national interest projects	identity of the wise	in the 2g spectrum scam, a. king	Mumbai taj attack	george bush's terrorist suppression campaign	cyclonic storms	intern leader su-ki
bing	tata's nano car	the nature of the soul	importance of karmabhoomi	project of national interest	identification of knowledgeable people	a. raja in 2g spectrum scam	Mumbai taj attack	george bush's terrorism suppression campaign	cyclones storms	the undersigned leader suukyi
Translate dict	tata nano car	form of soul	importance of land	national interest projects	identity of the wise	in the 2g spectrum scam, a. king	Mumbai taj attack	george bush's terrorist suppression campaign	cyclonic storms	the undersigned leader suukyi
Translate .com	tata's nano car	the nature of the soul	importance of karmabhoomi	project of national interest	identification of knowledgeable people	a. raja in 2g spectrum scam	Mumbai taj attack	george bush's terrorism suppression campaign	cyclones storms	the undersigned leader suukyi
collins translator	tata's nano car	the nature of the soul	importance of karmabhoomi	project of national interest	identification of knowledgeable people	a. raja in 2g spectrum scam	Mumbai taj attack	george bush's terrorism suppression campaign	cyclones storms	the undersigned leader suukyi
imtranslator	tata's nano car	soul form	importance of	projects of country hit	identification of knowledge	a. in 2g spectrum scam raja	Mumbai taj attack	george bush's terrorism	cyclone of	seen leader su-ki

			karmabhoomi					daman abhiyan	cyclonic storms	
yandex translate	tata's nano car	the nature of the soul	importance of karmabhoomi	the state of the nation	identity of the knowledgeable person	a. in the 2g spectrum scandal king	Mumbai taj attack	george bush's terrorism suppression campaign	orgy of cyclonic storms	hsu – ki
systran translate	tata nano	nature of soul	importance of karma bhoomi	projects of national interest	identifying the knower	2g scam: a raja	mumbai attack	george bush's counter-terrorism campaign	cyclonic storm	detained leader suukyi

As already discussed in the previous section, the present work aims to include popular free online translators for the evaluation of their efficiency over web query translation from Hindi to English.

4. Results & Discussion

Information retrieval in cross-lingual system is a challenging task because queries written in the native scripts need to be matched with the documents written in the Roman scripts. The initial step in CLIR i.e. translation of web query, plays a significant role as the accuracy of the query translation is directly proportional to the retrieval effectiveness. Thus, the selection of suitable translator in order to achieve the quality translations for specific language pair plays a crucial role. Quality estimation of translators in terms of evaluation metrics are essential in order to identifying the suitability of translators for different purposes. Evaluation metrices may have certain scores or range above which translations may be considered understandable. We have found some benchmark scores based on the analysis of few research papers to show the value range of evaluation metrics for various translators. For example, the BLEU score below 0.3 shows that the translations may contain significant grammatical errors, 0.3 to 0.4 understandably good and above 0.6 resembles better quality translations than human translators respectively on the scale of 0 to 1 [31]. METEOR scores below 0.5 generally contains grammatical errors, 0.5 and above reflect understandable translations and over 0.7 reflect good and fluent translations on the scale of 0 to 1 [32].

Web queries and sentences has basic differences in their structure. A sentence has a standard syntax which contains suitably placed part of speech (POS) and helping verb in it, but a web query usually misses a proper syntax. Another difference is the absence of context in the web queries usually resulting in ambiguity which makes it very difficult for the translators to produce correct translation. Translation ambiguity may occur due to the varied meaning of the same word which is termed as homonymy and polysemy [33]. Homonymy which is a word that has completely different meanings, for example the word “bank” can either mean a river bank or a financial institution, and polysemy which has two distinct meanings but are related for example “head” may refer to family head or the human head.

Table 4. Example of ambiguous web queries with translation

Translator/Queries	**गुलाबकीकलम**	**रेखाकाजन्मस्थान**
Google	rose pen	Rekha's birthplace
Bing	rose pen	Birthplace of the line
Translate Dict	rose pen	Rekha's birthplace
Translate.com	rose pen	Birthplace of the line
Collins Translator	rose pen	Birthplace of the line
Im Translator	rose pen	Line's birthplace
Yandex	rose pen	The birthplace of the line
Systran	rose pen	Line Birthplace

There are various parameters to view MT quality like adequacy, measurement of how much information from the original text is preserved in MT output, fluency, measurement of how naturally the MT output sounds in the target language, informativeness and no method can meet all of the requirements for quality estimation [34]. Until an evaluation metric doesn’t considers all the parameters for the evaluation of MT, it may not correctly evaluate or evaluate closer to human evaluator.

Let us take some of few examples to understand the translation ambiguity in Hindi, like {गुलाबकीकलम} ‘Gulab ki kalam’ and {रेखाकाजन्मस्थान} ‘Rekha ka janmsthaan’ as shown in Table 4. First one means ‘a branch of rose’ but all the translators translated the query wrongly as they are getting confused by the word {कलम} ‘Kalam’ which means a ‘small branch’ as well as ‘pen’. The logical translation of the second example is ‘birth-place of Rekha’ but translators other than Google and TranslateDict produces wrong output as the word {रेखा} ‘Rekha’ has two meaning i.e. a line and a name ‘Rekha’. As per our understanding the error in the translations may exist due to the fact that the datasets used to train the models may not have or comparatively less data availability of the ambiguous word having another meaning in the query. If one

meaning of a homonymy and polysemy word present in the dataset is used majorly for translations then the neural connections with the respective meaning will get stronger and thus, will be given preference in further translations. Another possible reason might be the structural difference of a query than a complete sentence as a sentence contains the required POS in proper place unlike a web query which has no definite structure. Web queries are usually shorter and hence faster to translate than a document. However, because of the limited context query translation suffers due to translation ambiguity.

Table 5. MT evaluation metrics scores for translator

	BLEU	NIST	MET EOR	hL epor	CHRF	GLEU
Google	0.449	0.189	0.715	0.759	0.725	0.594
Bing	0.323	0.167	0.602	0.663	0.629	0.426
TranslateDict	0.455	0.182	0.721	0.756	0.718	0.586
Translate.com	0.328	0.167	0.599	0.664	0.632	0.429
Collins Translator	0.323	0.168	0.602	0.662	0.629	0.426
ImTranslator	0.248	0.199	0.501	0.572	0.533	0.359
Yandex	0.26	0.183	0.529	0.628	0.536	0.364
Systran	0.24	0.206	0.501	0.601	0.52	0.347

As mentioned in the previous section, the web queries have been taken mainly from Fire and a popular Hindi newspaper for the translation. Translators and their scores for web queries (irrespective of their categories) for respective evaluation metric has been shown in table 5. Google Translator scores the highest whereas, Systran, the lowest in terms of average score of all metric with performance difference of 42.06% between the two. Google translator scores the highest in case of hLepor, CHRF and GLEU which indicate that the translations achieved from Google translator has the highest quality in terms of word order, robustness, fluency and accuracy. Translate Dict shows the highest METEOR and BLEU score reflecting good translation quality including semantic similarity, adequacy and lexical similarity respectively. Systran's being the lowest scorer in majority of metrics shows inadequacy and the relative weakness in terms of fluency, accuracy and lexical similarity whereas Im translator's weakness include inadequacy and word order.

Table 6. MT evaluation metrics scores for short queries

	BLEU	NIST	MET EOR	hL epor	CHRF	GLEU
Google	0.301	0.148	0.618	0.678	0.649	0.543
Bing	0.219	0.143	0.515	0.58	0.565	0.37
Translate Dict	0.312	0.147	0.647	0.691	0.652	0.549
Translate .com	0.219	0.143	0.515	0.58	0.565	0.37
Collins Translator	0.219	0.143	0.515	0.58	0.565	0.37
Im Translator	0.149	0.151	0.325	0.39	0.385	0.254
Yandex	0.135	0.14	0.388	0.497	0.388	0.27
Systran	0.171	0.171	0.373	0.506	0.409	0.286

Translation accuracy for the long queries has been observed to be higher than the shorter ones, as shown in table 6 and table 7. It may be due to the fact that the longer queries generally contain some more contextual information as a result these are less ambiguous in comparison to the shorter queries, which might be the reason that translators are able to achieve higher accuracy in case of such queries.

Table 7. MT evaluation metrics scores for longer queries

	BLEU	NIST	MET EOR	hLepor	CHRF	GLEU
Google	0.548	0.25	0.78	0.814	0.775	0.628
Bing	0.393	0.203	0.66	0.719	0.672	0.464
Translate Dict	0.551	0.234	0.77	0.799	0.762	0.611
Translate .com	0.4	0.203	0.656	0.721	0.676	0.469
Collins Translator	0.393	0.205	0.66	0.718	0.672	0.463
Im Translator	0.314	0.271	0.618	0.694	0.632	0.428
Yandex	0.344	0.246	0.623	0.715	0.635	0.426
Systran	0.285	0.259	0.587	0.665	0.593	0.388

It has occasionally been observed that in certain queries although only a few words get correctly translated with respect to the reference translation, scores better in terms of BLEU and NIST as compared to the queries having most of the words correctly translated but placed at inappropriate places. This is possibly due to the word order after translation which is considered in these metrics. This is possibly due to the fact that the proper word order gets more score than the correct word translation. In relation to the issues of how much credit is provided for appropriate lexical choice and how much additional credit is given for appropriate word order, the BLEU and the NIST metrics reflects opposing behavior.

TranslateDict scores the highest in majority of metrics for shorter queries while Im Translator the lowest and the performance difference between these two is 81.25%. The performance difference between Google, which excels in the majority of metrics and Systran which obtains the lowest score for long queries, is 36.67%. It has also been observed that the translation accuracy for long queries is higher in comparison to the shorter ones for all the translators. This difference in accuracy is mainly due to the presence of more contextual information in the longer queries. The

difference in translation accuracy between short and long queries for Google and Translate Dict, (i.e. the highest scorer for long and short queries respectively) is 29.21% and 24.31% respectively.
The present work may be useful for researchers in identifying strengths and limitations of the translation system guiding improvements in accuracy and fluency, it may also assist users in selecting the appropriate MT system based on the needs and also insights the underlying issues to be further explored in various MT systems. It also motivates development of research and innovations, leading to advancements in CLIR and MT technologies and also ensures continuous improvement with the evolving demands of cross-lingual communication and information retrieval. This work may have limitations such as a relatively small dataset size, focusing on a specific language pair and considering only free online translators. Future research could explore larger and more diverse set of queries involving some other MT tools to improve the robustness of results, provide greater understandings of effectiveness and areas for improvement.

5. Conclusion

This study compared the efficiency of online translators for Hindi web queries to English with various evaluation metrics. Our findings show that the translation of web queries is more challenging as compared to the complete sentence, and also translators often produce incorrect translations at many occasions due to the lack of context and proper syntax. The research findings include that the Google Translator outperformed other translators, while Systran scoring the lowest average metric score, resulting a performance difference of 42.06% between them. We also found that shorter queries with limited context posed significant challenges for accurate translation, while longer queries generally achieved higher accuracy. Among the tested translators, Google Translate consistently ranked highest, suggesting its potential for effective handling of Hindi web queries in CLIR. The findings of this work offer an understanding into how free online translation system performs for queries with varying size. In this work while the focus is only on Hindi-English and free online translators, hence we are planning to expand the further research on diverse language pairs also including other MT systems as well that could extend on these findings and refine optimal translator choices for specific context and length-related factors.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[2] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[3] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, “A study of translation edit rate with targeted human annotation,” in Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, 2006, pp. 223–231.

[4] E. M. Visser and M. Fuji, “Using sentence connectors for evaluating MT output,” arXiv preprint cmp-lg/9608019, 1996.

[5] C. Callison-Burch, “Fast, cheap, and creative: Evaluating translation quality using Amazon’s Mechanical Turk,” in Proceedings of the 2009 conference on empirical methods in natural language processing, 2009, pp. 286–295.

[6] C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan, “Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation,” in Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, 2010, pp. 17–53.

[7] P. Koehn et al., “Moses: Open source toolkit for statistical machine translation,” in Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, 2007, pp. 177–180.

[8] D. Chiang, S. DeNeefe, Y. S. Chan, and H. T. Ng, “Decomposability of translation metrics for improved evaluation and efficient algorithms,” in Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pp. 610–619.

[9] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, “COMET: A neural framework for MT evaluation,” arXiv preprint arXiv:2009.09025, 2020.

[10] E. Fonseca, L. Yankovskaya, A. F. T. Martins, M. Fishel, and C. Federmann, “Findings of the WMT 2019 shared tasks on quality estimation,” in Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), 2019, pp. 1–10.

[11] S. Chauhan, S. Saxena, and P. Daniel, “Monolingual and parallel corpora for Kangri low resource language,” arXiv preprint arXiv:2103.11596, 2021.

[12] S. Chauhan, P. Daniel, A. Mishra, and A. Kumar, “Adableu: A modified bleu score for morphologically rich languages,” IETE J Res, vol. 69, no. 8, pp. 5112–5123, 2023.

[13] G. Doddington, “Automatic evaluation of machine translation quality using n-gram co-occurrence statistics,” in Proceedings of the second international conference on Human Language Technology Research, 2002, pp. 138–145.

[14] A. L. F. Han, D. F. Wong, and L. S. Chao, “LEPOR: A robust evaluation metric for machine translation with augmented factors,” in Proceedings of COLING 2012: Posters, 2012, pp. 441–450.

[15] Y. Wu, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” arXiv preprint arXiv:1609.08144, 2016.

[16] P. A. Watters and M. Patel, “Semantic processing performance of Internet machine translation systems,” Internet Research, vol. 9, no. 2, pp. 153–160, 1999.

[17] P. Boonkwan and A. Kawtrakul, “Plaesarn: machine-aided translation tool for English-to-Thai,” in COLING-02: Machine Translation in Asia, 2002.

[18] O. Bezhanova, M. Byezhanova, and O. Landry, “Comparative analysis of the translation quality produced by three MT systems,” McGill University, Montreal, Canada, 2005.

[19] M. Aiken and Z. Wong, “Spanish-to-English translation using the Web,” Southwest Decision Sciences Institute, 2006.

[20] M. Aiken, K. Ghosh, J. Wee, and M. Vanjani, “An evaluation of online Spanish and German translation accuracy,” Communications of the IIMA, vol. 9, no. 4, pp. 67–84, 2009.

[21] S. Hampshire and C. P. Salvia, “Translation and the Internet: evaluating the quality of free online machine translators,” Quaderns: revista de traducció, pp. 197–209, 2010.

[22] A. Guerberof Arenas, “Exploring Machine Translation on the Web,” revistatradumática, no. 8, pp. 1–6, 2010.

[23] R. de Oliveira and D. Anastasiou, “Comparison of SYSTRAN and Google Translate for English→ Portuguese,” Tradumática, no. 9, pp. 118–136, 2011.

[24] H. Li, A. C. Graesser, and Z. Cai, “Comparison of Google translation

with human translation," in the twenty-seventh international flairs conference, 2014.
[25] S. Chand, "Empirical survey of machine translation tools," in 2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2016, pp. 181–185.
[26] I. Sutskever, "Sequence to Sequence Learning with Neural Networks," arXiv preprint arXiv:1409.3215, 2014.
[27] C. Chandler, P. W. Foltz, and B. Elvevåg, "Improving the applicability of AI for psychiatric applications through human-in-the-loop methodologies," Schizophr Bull, vol. 48, no. 5, pp. 949–957, 2022.
[28] A. S. Utami, N. Yusliani, M. D. Marieska, and A. Abdiansah, "Query Reformulation for Indonesian Question Answering System Using Word Embedding of Word2Vec," Computer Engineering and Applications Journal, vol. 11, no. 1, pp. 1–14, 2022.
[29] Y. Mao, C. Wan, Y. Jiang, and X. Gu, "Self-supervised query reformulation for code search," in Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2023, pp. 363–374.
[30] S. Agrawal, S. Merugu, and V. Sembium, "Enhancing e-commerce product search through reinforcement learning-powered query reformulation," in Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 4488–4494.
[31] C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan, "Findings of the 2011 workshop on statistical machine translation," in Proceedings of the sixth workshop on statistical machine translation, 2011, pp. 22–64.
[32] A. Agarwal and A. Lavie, "Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output," in Proceedings of the Third Workshop on Statistical Machine Translation, 2008, pp. 115–118.
[33] D. Zhou, M. Truran, T. Brailsford, V. Wade, and H. Ashman, "Translation techniques in cross-language information retrieval," ACM Computing Surveys (CSUR), vol. 45, no. 1, pp. 1–44, 2012.
[34] E. Hovy, M. King, and A. Popescu-Belis, "Principles of context-based machine translation evaluation," Machine Translation, vol. 17, pp. 43–75, 2002.
[35] Amit Asthana, Ganesh Chandra, Sanjay K. Dwivedi, "Approach to Handle Compound Out of Vocabulary Words in Hindi Web Queries", Journal of Information Systems Engineering and Management, Vol. 10No.13s (2025), https://doi.org/10.52783/jisem.v10i13s.2158
[36] Asthana Amit, and Sanjay K. Dwivedi. "Exploring snippets as a dataset to overcome challenges in CLIR." *ITM web of conferences*. Vol. 54. EDP Sciences, 2023.