

Error Classification and Evaluation of Machine Translation Evaluation Metrics for Hindi as a Target Language

Samiksha Tripathi¹ and Dr. Vineet Kansal²

¹ AKTU, Lucknow, UP 226031, India

Samiksha.tripathi@gmail.com

² AKTU, Lucknow, UP 226031, India

Abstract. Evaluation of Machine Translation (MT) is an onerous but a critical task. Automatic evaluation metrics evaluates the adequacy and fluency of a translated sentence. Automatic evaluation of machine translation is able to compare between two different translation systems but it doesn't provide any insights into the kind of errors a translation system is making. Our error classification, inspired by Vilar et al, has extended categories more linguistically for Hindi language. In this paper, we will explore various evaluation metrics for machine translation and perform extensive linguistic and statistical analysis of the translation output to identify primary issues in existing framework of automated metrics for English-to-Hindi MT systems. This leads us to better insights for improvement of these metrics for English-to-Hindi automatic machine translation.

Keywords: Evaluation Metric, English-to-Hindi MT, Error Classification in MT.

1. Introduction

In Machine Translation, meaning of a text in source language is fully transformed to the equivalent meaning in target language. Translation can never be word for word substitution due to language specific characteristics, so it is a very challenging exercise to evaluate the Machine Translation. Most of the evaluation metrics focus on lexicon matching between the tokens of candidate translation (produced by the MT system) and reference translation [1].

There are many evaluation metrics being developed. These metrics work accurately for many languages but they fail in rating translation correctly for morphologically rich target languages such as Hindi. It is important to capture various types of divergence between English and Hindi and then try to evaluate against a specific evaluation metric system. We first did a linguistic error analysis to get qualitative insights into the variety of issues that crop up for English-to-Hindi machine translation. Subsequently a detailed statistical evaluation is performed to quantify the problems encountered with existing framework of MT evaluation metrics. Adequacy (Comprehensiveness) and Fluency (Naturalness) are most desirable features for correct translation [1]. A translation can be defined as adequate if it preserves the meaning of the source language and does not add additional information. Grammatical and natural text in the target language is known as fluency. Adequacy and fluency are the major aspects of the machine translation performance.

Evaluation of automated translation system can be done using Post Editing (PE) which makes correction in target language text that has been translated from source language using automated MT system. The problems with MT system can be understood with the help of PE; it automatically tests the MT system as well as evaluates it. Post Editing can be used in evaluating the quality of translation. It can be calculated based on correction or effort required in fixing the issues observed in automated machine translation.

As we can see more accurate translation needs less editing. Translated text from automated MT system is edited by human annotators. Post editing of translated text (output of MT System) is quicker compared to the one where translation is initiated from scratch by a human translator.

This paper discusses the challenges in designing the automated evaluation metrics for English-to-Hindi MT. In Section 2 we will talk about the brief history of human and automated evaluation metrics. Section 3 elaborates the automated evaluation. Issues in MT Evaluation with examples of Hindi sentences are discussed in section 4. Section 5 elaborates on statistical analysis for evaluation of such metrics and error classification.

Finally we conclude our discussion summarizing the issues in existing automatic English-to-Hindi MT evaluation metrics and provide suggestions for future work in this domain.

2. Related Work

Human evaluation of Machine Translation system is highly subjective and time consuming and it cannot be reused. Due to above mentioned reasons, nowadays automatic evaluation methods are more common. Widely and popularly known algorithm for evaluating the quality of machine translated text is BLEU [3]. It calculates n-gram precision and a brevity penalty between the candidate and reference translation. Some alterations have been done in BLEU metric to come up with NIST metric. NIST metric provides weights to n-gram based on how informative this n-gram is. Lower weight will be given for frequent n-grams. It is proposed by Doddington [4]. Turian et al. [5] introduced GTM (General Text Matcher) based on accuracy measures such as precision, recall and f-measure. The package for automatic evaluation of summaries is introduced in 2003 named as ROUGE [17]. In ROUGE, computer generated summary (candidate summary) is evaluated with the human created summary (reference summary). METEOR [16] came in 2005 that creates a word alignment between the two sentences i.e candidate translation string and reference translation string. The alignment is done through a word mapping such as i) Stem matching, ii) Exact matching, iii) Synonym matching. After getting the final alignment, score is calculated as the harmonic mean of unigram precision and recall. In recent years extension of METEOR translation evaluation metric is done to the phrase level in METEOR NEXT metric [8]. Additional paraphrase matcher is introduced where phrases and words are matched. Different human judgements are explored with Human mediated Translation Edit Rate (HTER) [9]. HTER is a semi-automatic measure where humans do not score the MT output but they generate a new reference translation which is closer to MT output and try to retain the fluency and adequacy of the original translated reference. The translated reference translation is used when the evaluation of MT is performed using Translation Edit Rate [9] or with other automated metrics. Annotator tries to minimize the number of edit to get the reference translation. So post editing is also one of the methods of estimating translation quality, more accurate translations require less editing. It is also helpful for finding the errors in the MT output. Correlation between automatic evaluation matrices with human judgements was attempted by Callison in 2007. They have tried to determine which automatic system produces the highest quality of translation from the list of nine different automatic evaluation metrics [2] and ranked them with the help of comprehensive human evaluation. Two categorical scales are currently being used to represent fluency and adequacy of MT system by the human evaluators.

Keeping different features of Hindi in mind, the METEOR Hindi was released by Ankush Gupta et. al [10] which is a modified version of METEOR based on Hindi specific features. METEOR Universal provides language specific evaluation by learning paraphrase table and function word list whereas earlier METEOR required human ranking judgments in the target language. METEOR Universal performed better for Russian and Hindi language [6].

There are various other metrics that were observed during the survey and many other evaluation metrics have released their new versions. During survey, it has been found that there are number of existing metrics but not all the metrics can correlate well with manual evaluation. They cannot work well with all the languages especially with free word order and morphologically rich languages like Hindi.

3. Classification of Hindi Translation Error

Evaluation of Machine Translation can be done more clearly when we are able to find the errors in machine translated outputs. One or more reference translations are required to find the errors in translation. It also helps in comparing the output of MT system with the correct text. Even though it is ambiguous due to several correct translations for the same source sentence, it is a worthwhile exercise to pinpoint the issues with MT and automatic MT evaluation.

A preliminary MT error typology for English language is defined by Llitjós et al [11] and Vilar has extended the error classification scheme [7].

We have classified the English-to-Hindi Translation errors in three segments. The errors have a hierarchal structure as shown in Figure 1 below.

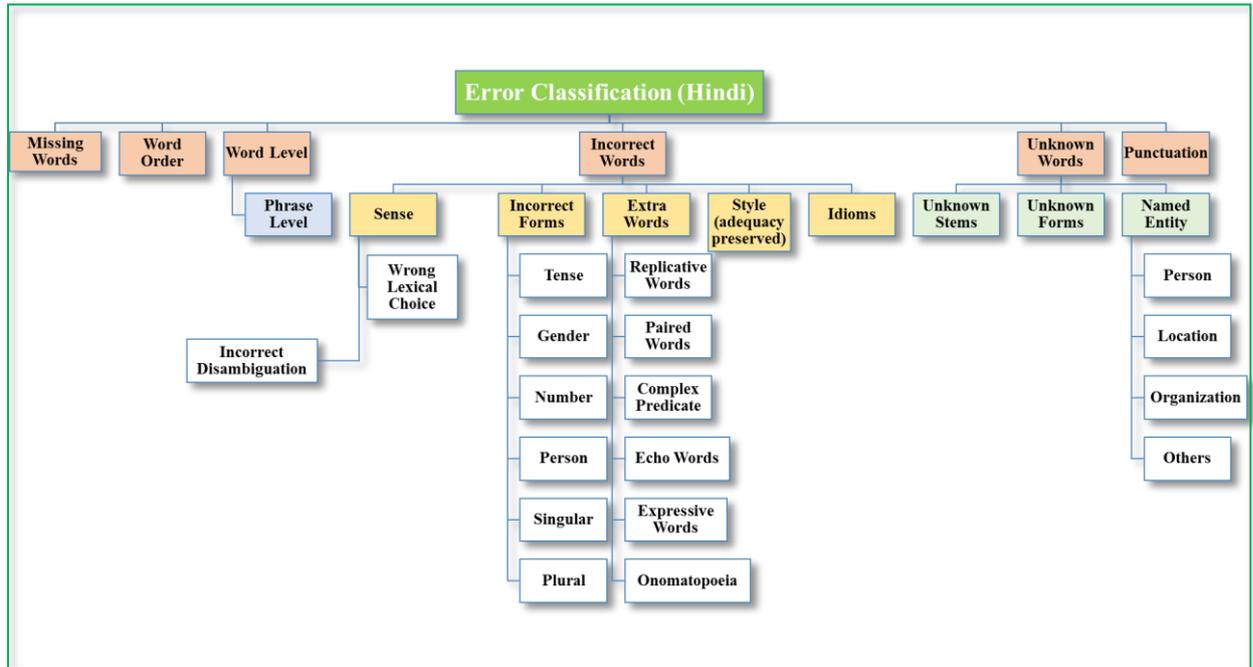


Fig. 1. Hierarchal Structure of Errors for English-to-Hindi Automatic MT

At the first level we have split the errors in eight big classes: “Missing Word”, “Word Order”, “Incorrect Words”, “Extra Words”, “Style”, “Idioms”, “Unknown Words” and “Punctuation”.

A “Missing Word” error is originated when few words are missing in translated sentence. Sometimes in Hindi translation missing of words is mandatory for expressing the translation. Eg: “It is raining outside” the translation in Hindi will be “बाहर बारिश हो रही है। baahar baarish ho rahii hai”. Meaning of “It” is missing in target translation to provide the *naturalness* in the translation.

There are several examples when missing word is essential for expressing the meaning of a source sentence. Missing words can be noun, verb or parsarg. Eg: Ram has two kids. “Raghav and Rajni have two kids” राघव और रजनी के दो बच्चे हैं। “raaghav aur rajanee ke do bachche hain .” When we tried to translate using Google it is giving output with missing “ke” “raaghav aur rajanee do bachche hain .” राघव और रजनी दो बच्चे हैं। This translation changes the meaning of source sentence due to missing parasarg from the original sentence to “Raghav and Rajni are two kids”

The second class of error is related to “word order”. Hindi is free word order language but sometimes word order makes a huge difference in the sense of the translated sentence. If we talk about the sentence like क्या आपने खाना खाया ? “kyaa aapne khaana khaaya?” and आपने क्या खाना खाया ? “aapne kyaa khaana khaaya?” Both the sentences have different meaning based on the position of word “kyaa” in the sentence. When “kyaa” is in first position then a person is asking that *Have u taken meal?* But when “kyaa” comes in second position then its sense is that *What items have you taken in meal?*

In the following sentence “not” has important place in target translation. If ‘not’ is associated with banana, then the translation will not be able to capture the correct sense and will change the meaning. “Raghav

likes banana not grapes.” Google: राघव नहीं केला अंगूर पसंद करता है। raaghava nahiin kelaa anguura pasand karta hai. Correct: राघव केला पसंद करता है न की अंगूर। raghav kela pasand karta hai na ki angura.

Errors due to wrong lexical choice and incorrect disambiguation fail to capture the correct sense of the word making the translation inaccurate. In a sentence “I want to meet American head.” “Head” is chief not a body part. We sometimes get Google translation as मैं अमेरिकी सिर से मिलना चाहता हूँ। The right lexicon choice for the given sentence is ”Mukhiya” मैं अमेरिकी मुखिया से मिलना चाहता हूँ। He had a great fall Google: वह एक महान गिर गया। Correct: वह धड़ाम से गिर गया। vaha dhadhaam se aa giraa. Example of wrong lexical choice and complex predicate: “The play is on.” Google: खेलने पर है। Correct: खेल चल रहा है। Khela chala rahaa hai.

Incorrect forms of words related to GNP or singular/plural etc. lies in “Incorrect Words” category. Sentences like “A Man scolded a boy.” एक आदमी ने लड़का को डांटा। Eka aadmi ne ladhka ko dantaa. एक आदमी ने लड़के को डांटा। Eka aadmi ne ladhke ko dantaa.

The additional word normally increased when we observe translation of expressive, replicative, paired or echo words in Hindi. “My hands are sticky”. Google : मेरे हाथ अवरुद्ध कर रहे हैं। Correct: मेरे हाथ चिप चिपा रहे हैं। This is a story of every house. Google: यह हर घर की कहानी है। Naturalness is better in the following sentence: यह घर घर की कहानी है। Please have some snacks. Google: कृपया कुछ नाश्ता लें। More fluent translation is: कृपया कुछ नाश्ता वाशता लें। Little fingers are so cute. Google: छोटी अँगलियों बहुत सुन्दर हैं। Replicative words give more fluent translation: छोटी छोटी अँगलियां बहुत सुन्दर हैं।

Interpretation of complex predicate is also an important factor. “He escaped drowning” Google: वह डूबने भाग निकले। vaha duubte duubte bacha. “वह डूबते डूबते बचा।” is a correct translation.

4. Automatic Evaluation

In this section we will talk about the methods to calculate the scores in automatic evaluation of MT system. There is also a brief discussion on few automated evaluation metrics.

Normally all automatic metrics of MT evaluation follow one of the following methods:

Precision Based: Total number of matched unigrams between candidate and reference translation are divided by the total length of the candidate translation of MT system.

Recall Based: Total number of matched unigrams between candidate and reference translation are divided by the total length of the reference translation of MT system.

F-measure Based: Collective scores of both precision and recall is being used.

Edit Distance Based: The number of insertion, deletion and substitution is counted for making candidate translation as reference translation.

4.1. BLEU

Papineni proposed BLEU (Bilingual Evaluation Understudy) based on n-gram metric. It evaluates candidate translations produced by an MT system and comparing them with reference translation done by human. For each n (ranges from 1 to maximum of 4) matching is done between candidate and corresponding reference translation. To compensate the difference in the length of candidate and reference translation the brevity penalty is used.

The final BLEU formula is

$$BLEU = BP \times \exp\left(\sum_{n=1}^4 \frac{1}{n} \log(pn)\right)$$

Where pn = modified n gram precision

Brevity penalty is calculated as

$$BP = \min(1, e^{1-lr/lc})$$

Where lc = length of the candidate translation

lr = effective reference corpus length

Discussion based on Hindi. In this section we will try to compile some examples from Hindi language where BLEU fails to score the MT evaluation. We have done both sentence level and corpus (multiple sentences) based analysis for the metrics. The detailed statistical analysis has been provided in Section 5. This section explores the issues from a qualitative linguistic perspective.

Table 1. Example

E: I am thirsty. H: मैं प्यासा हूँ। C1: मुझे प्यास लग रही है।
E-Source language, H- Human Reference Translation, C- Candidate Translation

BLEU is not able to correlate well with human judgements in all scenarios. In the above example, unigram precision is 0 out of 6; bigram precision is 0 out of 5 and so on. BLEU is not able to capture adequacy of the translation.

Table 2. Example

English Sentence	Human Reference Translation	Candidate Translation
The boy who is standing there is my brother.	वो लड़का वहाँ खड़ा है मेरा भाई है।	लड़का है जो वहाँ खड़ा है मेरा भाई है।
Even today he breaks down when he reminded of you.	तुम्हारी याद करके वह आज भी रो पड़ता है।	वह आप की याद दिला दी जब तक कि आज वह टूट जाती है।
Please check where the hotel is.	होटल तो देख लो की कहाँ है।	कृपया निरीक्षण करें जहाँ होटल है।
Anil used to say that but now his father also says same.	अनिल तो कहता ही था अब उनके पिता भी यही कहते हैं।	अनिल का कहना है कि करते थे. लेकिन अब उनके पिता भी एक ही कहते हैं।
Just look at the watch.	जरा घड़ी को तो देख लो।	आप सिर्फ घड़ी देखिये।

All the above four examples in Table 2 are revealing diverse features of Hindi language like the role of verb phrase, post positions and adverb. Candidate translations are failing in emphasizing the words whereas human translation could do that with the help of adverbs or postpositions.

After detailed analysis as presented in Section 5 and based on above observations, we can conclude that BLEU metric disappoints if the target language is Hindi.

4.2. METEOR

METEOR is an automatic metrics for MT system evaluation based on the concept of unigram matching between the MT systems produced translation and Human reference translation. METEOR is designed after observing the weaknesses of BLEU metrics. It is based on word to word alignment between machine translation and reference translation. Alignment between two sentences can be achieved by exact matching of words if their surface forms are identical. METEOR also matches words with simple morphological variants that can be aligned. If stems are identical and have similar synonym sets then matching will be done between system generated translation and reference translation.

The score is calculated as harmonic mean of unigram precision (matched n-grams out of the total number of n-grams in a MT system produced translation) and unigram recall (matched n-grams out of the total number of n-grams in reference translation). In METEOR-NEXT [12], paraphrase matcher is being introduced. It matches the phrases between the two strings.

Table 3. Example

E: Raghav had beaten Radha with stick.
H: राघव ने राधा को डंडे से मारा।
C1: राधा को राघव ने डंडे से मारा।
C2: राधा ने राघव को डंडे से मारा।
E-Source language, H- Human Reference Translation, C- Candidate Translation

METEOR-Hindi [10] used word based features and other linguistic parameters such as local word groups, part of speech tags and clause boundaries. Local Words Groups (LWG) [14] is a group of content and its associated function words. Function words tell the grammatical role of the content word in the sentence. Example sentence from Table 3 is showing the importance of Local Words Group. If we observe C1 and C2, we will see that all the words in both the sentences are showing exact match with the human reference sentence but C1 and C2 have opposite meaning. With the help of LWG, we can look into the difference and can assign the scores accordingly.

If all the words of a sentence are matched in METEOR-Hindi but POS is not same for the words then the sentence will get low scores. METEOR-Hindi also computes the exact matching of clauses and gives the scores according to the matched clauses.

4.3. TER (Translate Error Rate)

TER measure is proposed by Dorr and Snover in 2006 [15]. It calculates the amount of editing required in a MT system output to achieve the exact match of reference translation. TER counts the number of edits based on fluency and adequacy of a sentence. In TER we do not generate a new reference but try to match the system output with existing references.

$$TER = \frac{\text{number of edits}}{\text{average of reference words}}$$

In HTER (Human Mediated Targeted Error Rate), we find minimum edits required as per the new targeted reference generated by human. HTER is more complex as a human does not score directly on the MT output. Instead they generate a new reference translation which is closer to the MT output translation.

TER_p is an extension of TER where alignment is not based on exact match; rather it takes synonyms and stem of a word while matching. It does Stem match, Synonym match and phrase substitutions.

Table 4. Example

E: Father scolded one boy. H: पिता ने एक लड़के को डांटा । C: पिता ने एक लड़का को डांटा ।
E-Source language, H- Human Reference Translation, C- Candidate Translation

We can observe how post editing can improve the accuracy of translation. We can also consider replicative words in evaluation of machine translation. Replicative words can occur in almost all the South Asian Languages [16]. Replicative words enhance the naturalness of the translation. They improve the fluency of the translated output.

Table 5. Example

E: This is a story of every house. H: यह घर घर की कहानी है। C: यह प्रत्येक घर की कहानी है ।
E-Source language, H- Human Reference Translation, C- Candidate Translation

5. Evaluation & Error Analysis

Based on qualitative insights being gathered, as described in previous section, a detailed statistical analysis has been performed to quantify the correlations for BLEU and METEOR against human evaluation.

5.1. Sample Input Data

148 small paragraphs were randomly selected from following sources:

- 92 Paragraphs from Online Course Material from National Institute of Open Schooling (NIOS) [18]
- 32 Paragraphs from Online Stories
- 24 Paragraphs from Government Websites [19]

NIOS provides educational material for vocational, secondary and senior secondary courses in both English and Hindi languages. Course content is developed in English language. Hindi content for corresponding course material is then generated using the services of human translators. These are considered as reference translations whereas machine translations have been obtained using Google translator [20].

Second set of data has been collected using online Indian stories in English language. Examples include stories from Premchand, Akbar-Birbal anecdotes etc. 32 randomly selected paragraphs from these sources were given to people with Hindi as their native language and who can use English fluently as second language. Again these were translated using Google translate as well for further evaluation.

India has 22 official languages including Hindi and English. Government websites are generally multilingual, so we have taken 24 paragraphs from there in both English and Hindi languages.

5.2. Error Classification

From the data generated, 32 paragraphs were randomly selected for determining the top 5 Class of Errors. The distribution of errors for these randomly selected statements is shown in Table 6 below.

Table 6. Errors Distribution

Class of Errors	Count
Missing Word	40
Wrong Lexicon Choice	72
Extra Word	40
Word Order	66
Incorrect Forms	20

5.3. Metric Evaluation

BLEU and METEOR are selected for English-to-Hindi MT metric evaluation. All 148 statements have been analysed. Python NLP library (NLTK [21]) was used to tokenize the sample paragraphs. Subsequently, a Python program was written to compare the two translations (Human vs Automatic) for matched unigrams, bigrams, trigrams and four-grams. 32 of these 148 statements were also analysed manually. Figure 2 shows the sample of task carried out manually, whereas, Table 7 below shows the recorded observations.

Following formulae were used to evaluate BLUE and METEOR scores. Precision, Recall and F-Measure have been obtained using standard formulae as described in Section 4. Table 8 compiles all the results thus obtained.

$$BLEU\ Score = MIN\left(1, \frac{output\ length}{reference\ length}\right) \times \left(\prod_{i=1}^4 precision_i\right)^{\frac{1}{4}}$$

$$METEOR\ Score = \frac{10PR}{R + 9P}$$

English Sample (NIOS)

The term 'food' brings to our mind countless images. We think of items not only that we eat and drink but also how we eat them and the places and people with whom we eat and drink. Food plays an important role in our lives and is closely associated with our existence. It is probably one of the most important needs of our lives.

Human Translator (SSM)

भोजन शब्द से हमारे मस्तिष्क में असंख्य छवियां उभरती हैं। हमारे ध्यान में न केवल वे वस्तुएं आती हैं जिन्हें हम खाते और पीते हैं, बल्कि यह भी कि हम उन वस्तुओं को कैसे तथा कहां और किन व्यक्तियों के साथ हम खाते पीते हैं। हमारे जीवन में भोजन एक महत्वपूर्ण भूमिका निभाता है और हमारे अस्तित्व के साथ इसका गहरा सम्बन्ध है। यह संभवतः हमारे जीवन की सर्वाधिक महत्वपूर्ण आवश्यकताओं में से एक है
(Number of Words: 75)

Human Translator (SSM)

(BLEU): शब्द 'भोजन' हमारे दिमाग अनगिनत छवियों को लाता है हम वस्तुओं के बारे में नहीं सोचते केवल यही कि हम खाते हैं और पीते हैं लेकिन यह भी कि हम उन्हें और स्थानों को कैसे खाते हैं और जिनके साथ हम खाना खाते हैं और पीते हैं खाद्य हमारे में एक महत्वपूर्ण भूमिका निभाता है जीवन और निकटता से हमारे अस्तित्व के साथ जुड़ा हुआ है जीवन और निकटता से हमारे अस्तित्व के साथ जुड़ा हुआ है यह संभवतः इनमें से एक है हमारे जीवन की सबसे महत्वपूर्ण जरूरतों

(METEOR): शब्द 'भोजन' हमारे दिमाग अनगिनत छवियों को लाता है हम वस्तुओं के बारे में नहीं सोचते केवल यही कि हम खाते हैं और पीते हैं लेकिन यह भी कि हम उन्हें और स्थानों को कैसे खाते हैं और जिनके साथ हम खाना खाते हैं और पीते हैं खाद्य हमारे में एक महत्वपूर्ण भूमिका निभाता है जीवन और निकटता से हमारे अस्तित्व के साथ जुड़ा हुआ है यह संभवतः इनमें से एक है हमारे जीवन की सबसे महत्वपूर्ण जरूरतों

BLEU	
Matched Unigram	45
Matched Bigram	23
Matched Trigram	14
Matched Fourgram	6
Adequacy (Comprehensiveness)	3
Fluency (Naturalness)	2

METEOR		Adequacy (Comprehensiveness)	Fluency (Naturalness)
Matched Words	54	3	2

Fig. 2. Sample of Evaluations for English-to-Hindi MT

Table 7. Observations

S. No.	Paragraph #	Exact Matched Words-Unigrams BLEU	Total Words in Reference Translation	Total Words in Candidate Translation	Matched Words-Stem, Synset, Paraphrase-METEOR
1	Para 1	45	75	78	54
2	Para 2	26	66	56	44
3	Para 3	71	182	159	110
4	Para 4	48	109	107	71
.
.
147	Para 147	52	100	96	70
148	Para 148	76	194	169	118

Table 8. Metric Evaluation

S. No.	Paragraph #	Precision (Unigram)	Precision (Bigram)	Precision (Trigram)	Precision (Fourgram)	Recall	F-Measure	BLEU (Only Unigram)	BLEU (Till Bigram)	BLEU (Till Fourgram)	Meteor Hindi
1	Para 1	0.5769	0.2987	0.1842	0.0800	0.6000	0.5882	0.5769	0.4151	0.2245	0.5976
2	Para 2	0.4643	0.2545	0.1296	0.0755	0.3939	0.4262	0.3939	0.2917	0.1565	0.4000
3	Para 3	0.4465	0.2595	0.1592	0.1026	0.3901	0.4164	0.3901	0.2974	0.1822	0.3951
4	Para 4	0.4486	0.2830	0.1810	0.1154	0.4404	0.4444	0.4404	0.3498	0.2227	0.4412
.
.
147	Para 147	0.5412	0.3157	0.2021	0.1290	0.5202	0.5305	0.5202	0.3973	0.2483	0.5222
148	Para 148	0.4485	0.1790	0.1141	0.0725	0.3906	0.4175	0.3906	0.2467	0.1398	0.3957

5.4. Adequacy (Comprehensiveness) and Fluency (Naturalness) Evaluation

BLEU and METEOR scores for 32 randomly selected paragraphs have been compared against the adequacy and fluency scores for English-to-Hindi MT provided by human evaluator on the same set of 32 paragraphs. Adequacy and Fluency scores are categorical with the scales shown in Table 9 below.

Table 8. Adequacy and Fluency Scales

Adequacy (Comprehensiveness)	Fluency (Naturalness)
All Meaning (5)	Flawless Hindi (5)
Most Meaning (4)	Good Hindi (4)
Much Meaning (3)	Non-Native Hindi (3)
Little Meaning (2)	Disfluent Hindi (2)
None (1)	Incomprehensible (1)

Pearson Correlation Coefficient has been obtained to understand the behaviour of BLEU and METEOR with respect to adequacy and fluency. Figure 3 provides the graphical interpretation of this correlation for BLEU whereas Figure 4 provides the same for METEOR scores.

Pearson Correlation Coefficient is given by:

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y}$$

where, \bar{x} , \bar{y} are mean values and s_x , s_y are square roots of variance.

As can be clearly seen from the graphed data, METEOR performs much better in terms of both adequacy and fluency compared to BLEU. METEOR gives a correlation coefficient of more than 0.75 for both adequacy and fluency whereas BLEU has a correlation coefficient of 0.51 for adequacy and 0.61 for fluency.

The correlation for BLEU is especially bad whereas METEOR gives a correlation which is somewhat acceptable with lot of room left for further improvements.

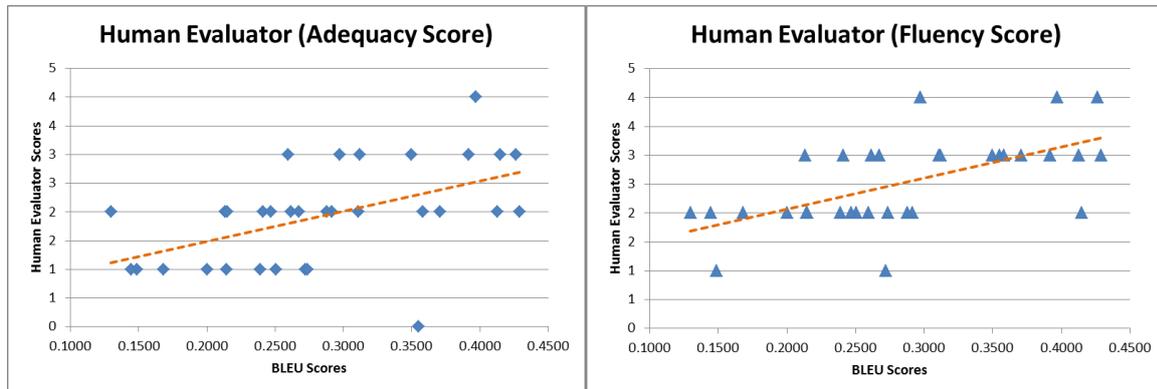


Fig. 3. BLEU vs. Human Evaluator

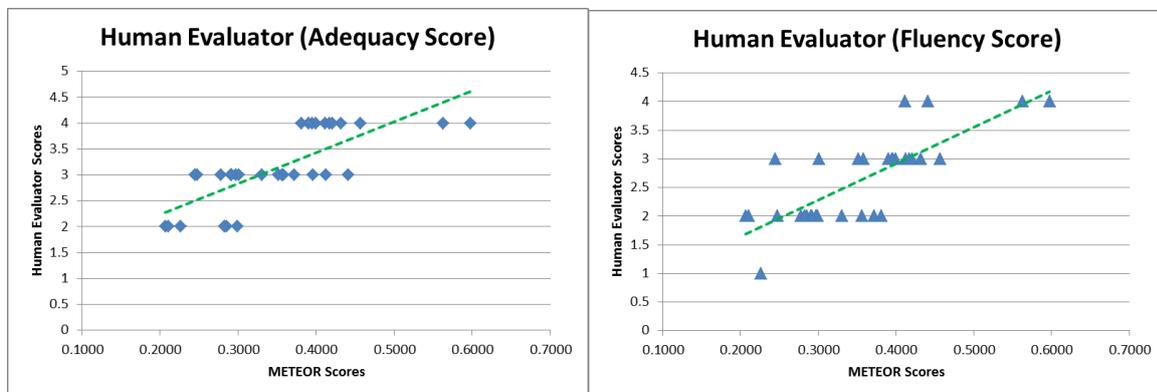


Fig. 4. METEOR vs. Human Evaluator

6. Conclusion

In the paper we discussed various important features of Hindi language and the anomalies that are encountered during evaluation of available translation systems. Various metrics have been studied and it has been established that these metrics are not able to evaluate English-to-Hindi translation appropriately.

We found that Post Position plays an important semantic role in Hindi. We should measure the Post Position Equivalence information to add one more level of matching word group. Even though Hindi is relatively free word order but Wrong Word/Phrase order impacts the naturalness of the MT translation. Incorrect form of verb influences the comprehensiveness of the text. We need to identify word groups such as NN + PSP, ADJ + NN and Verb groups, which includes verb and auxiliary verbs. We can score evaluation metrics based on strong and weak equivalence and Head Word matching of the reference and candidate translation.

As guidance for future work, all these parameters along with the tagged data obtained during the analysis contained in this paper can be utilized as input features for machine learning algorithms including neural nets to improve the performance of automatic English-to-Hindi MT evaluation metrics.

References

1. Specia, Lucia, et al.: "Predicting machine translation adequacy." Machine Translation Summit XIII, Xiamen, China (2011).
2. Callison-Burch, Chris, et al.: "(Meta-) evaluation of machine translation." Proceedings of the Second Workshop on Statistical Machine Translation. Association for Computational Linguistics (2007).
3. Papineni, Kishore, et al.: "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics (2002).
4. Doddington, George: "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics." Proceedings of the second international conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc. (2002).
5. Turian, J. P., L. Shen, and I. D. Melamed: Evaluation of Machine Translation and its Evaluation. In Proceedings of MT Summit IX, New Orleans, U.S.A (2003).
6. M Denkowski, et al.: "Meteor universal: Language specific translation evaluation for any target language" Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 376–380, Baltimore, Maryland USA, June 26–27, 2014, Association for Computational Linguistics (2014).
7. Vilar, David, et al.: "Error analysis of statistical machine translation output." Proceedings of LREC (2006).
8. Denkowski, Michael, and Alon Lavie: "Extending the METEOR machine translation evaluation metric to the phrase level." Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics (2010).
9. Snover, Matthew, et al.: "A study of translation edit rate with targeted human annotation." Proceedings of association for machine translation in the Americas (2006).
10. Gupta, Ankush, Sriram Venkatapathy, and Rajeev Sangal: "METEOR-Hindi: Automatic MT Evaluation Metric for Hindi as a Target Language." Proceedings of ICON-2010: 8th International Conference on Natural Language Processing (2010).
11. Ariadna Font Llitjós, Jaime G. Carbonell, and Alon Lavie.: A framework for interactive and automatic refinement of transfer-based machine translation. In Proc. of the 10th Annual Conf. of the European Association for Machine Translation (EAMT), Budapest, Hungary, May (2005).
12. Denkowski, Michael, and Alon Lavie: "Extending the METEOR machine translation evaluation metric to the phrase level." Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics (2010).
13. Kalyani, Aditi, et al.: "Assessing the Quality of MT Systems for Hindi to English Translation." arXiv preprint arXiv:1404.3992 (2014).
14. Bharati, Akshar, Vineet Chaitanya, and Rajeev Sangal : "Local word grouping and its relevance to Indian languages." Frontiers in Knowledge Based Computing (KBCS90), VP Bhatkar and KM Rege (eds.), Narosa Publishing House, New Delhi 277-296 (1991).
15. Ananthakrishnan, R., et al.: "Some issues in automatic evaluation of english-hindi mt: more blues for bleu." ICON (2007).
16. Banerjee, Satanjeev, and Alon Lavie.: "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (2005).
17. Hovy, Eduard, et al.: "Automated summarization evaluation with basic elements." Proceedings of the Fifth Conference on Language Resources and Evaluation (LREC 2006).
18. NIOS Online Course Material, <http://www.nios.ac.in/online-course-material.aspx>.
19. Center for Development of Advanced Computing, Govt. of India, <http://www.cdac.ac.in>.
20. Google Translate, <https://www.translate.google.com>.
21. NLTK NLP Library in Python, <http://www.nltk.org>.