METEOR-Hindi : Automatic MT Evaluation Metric for Hindi as a Target Language

Ankush Gupta and **Sriram Venkatapathy** and **Rajeev Sangal**
Language Technologies Research Centre,
IIIT-Hyderabad, Hyderabad, India.
ankushgupta@students.iiit.ac.in,
{sriram,sangal}@iiit.ac.in

Abstract

BLEU (Papineni et al., 2002) is a widely used metric for machine translation evaluation. However, it fails to rate translations correctly for target languages that are morphologically rich and that have relatively free word order such as Hindi (Ramanathan et al., 2007).

In this paper, we present METEOR-Hindi, an automatic evaluation metric for a machine translation system where the target language is Hindi. METEOR-Hindi is a modified version of the metric METEOR, containing features specific to Hindi. We make appropriate changes to METEOR's alignment algorithm and the scoring technique.

In our experiments, we observed that METEOR-Hindi achieved high correlation of 0.703 with human judgments significantly outperforming BLEU that had a correlation of only 0.271.

1 Introduction

Machine translation is the process of transforming a sentence in one human language to another with the same meaning and similar construction. The quality of such systems can be measured, both based on human rating, and as well as using automatic scoring metrics. Human evaluation is the most reliable way for evaluating MT systems but it is subjective, expensive, time-consuming and involves human labor that cannot be reused. Automatic MT evaluation metrics play a prominent role in the evaluation of MT systems. Many automatic measures have been proposed to facilitate fast and cheap evaluation of MT systems, the most widely used of which is BLEU (Papineni et al., 2002), an evaluation metric that matches ngrams from multiple references. A variant of this metric, typically referred to as the "NIST" metric, was proposed by Doddington (Doddington., 2002).

While popular, weaknesses have been noted in BLEU in recent years, most notably the lack of reliable sentence-level scores (Liu et al., 2005; Liu and Gildea., 2006). Further, it is not suitable for evaluation of English-Hindi MT systems because of the properties of Hindi, such as, rich morphology and relative free word orderness. More details of the limitations of BLEU are present in section 2.

In order to overcome the weaknesses of BLEU, several metrics were proposed such as, METEOR (Banerjee and Lavie., 2005), GTM (Melamed et al., 2003), TER (Snover et al., 2006) and CDER (Leusch et al., 2006). Some metrics outperformed METEOR in the "Shared Task: Machine Translation for European Languages" in 2009 and 2010. Some of them are : i-letter-BLEU, SVM-rank, TESLA-M (Liu et al., 2010), Bkars. But among these measures, METEOR is the most suitable for evaluation of English-Hindi MT, as it offers immense flexibility in encoding parameters (other than just the word order sequences) that indicate quality of translation. It allows us to use linguistic cues for the task of evaluation. We present details of the advantages in using METEOR for English-Hindi evaluation in section 3.

METEOR, does not support Hindi by default, as it requires Hindi specific tools for computing synonyms, stem words, etc. Also, as METEOR is a language-pair specific evaluation tool, certain language specific characteristics need to be incorporated in the METEOR evaluation tool where the target language is Hindi. In this paper, we present

"Proceedings of ICON-2010: 8th International Conference on Natural Language Processing, Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2010"

METEOR-Hindi, which is a modified version of METEOR, containing features specific to Hindi. The characteristics of English-Hindi MT that help in the design of METEOR-Hindi are listed in section 4. The tools that we used to implement METEOR-Hindi are presented in section 5. The METEOR-Hindi algorithm (word-alignment and scoring) is described in section 6.

We show that METEOR-Hindi gave evaluation scores that correlated well with the human scores. The correlation value of METEOR-Hindi (0.703) was much higher than the correlation value of BLEU (0.271). We also conducted experiments on examining the importance of various features in METEOR-Hindi for evaluation. However, no conclusive points were observed yet on the importance of various features, because of the limited size of the evaluation dataset. The experiments and results are presented in greater detail in section 7, and the conclusion is illustrated in section 8.

2 Problems with BLEU metric

In this section, we list the reasons why BLEU is not an appropriate metric for English-Hindi evaluation (Ramanathan et al., 2007).

1. **Meaningless Sentence-level Score** (Liu et al., 2005; Liu and Gildea., 2006): As Hindi is a relatively free word order language (Rao., 2001), BLEU which is an n-gram precision based metric assigns meaningless scores to those sentences which have the same meaning as the reference but a different word order.

2. **Only Exact Matches** : Hindi is morphologically very rich and BLEU takes only exact matches into account, so it is not suitable for English-Hindi MT evaluation. Synonym matching is not there in BLEU, so "पुस्तक" and "किताब" are considered as different words.

3. **Lack of recall** : It is a significant weakness in BLEU, and the Brevity Penalty in the BLEU metric does not adequately compensate for the lack of recall. Recall has been confirmed by several metrics as being critical for high correlation with human judgments (Lavie et al., 2004).

4. **Admits too much variation by using Higher Order N-grams for Fluency and Grammaticality** : BLEU uses higher order n-grams to encapsulate and indirectly measure fluency and grammaticality in translation hypothesis. The criticism is that the n-gram matching technique is naive, allowing just too much variation. There are typically thousands of variations on a hypothesis translation, a vast majority of them both semantically and syntactically incorrect, that receive the same BLEU score. (Callison-Burch et al., 2006) note that phrases that are bracketed by bigram mismatch sites can be freely permuted, because reordering a hypothesis translation at these points will not reduce the number of matching n-grams and thus will not reduce the overall BLEU score. Example:
Reference : मुझे यकीन है कि जब खाद्य पदार्थों में विविधता की बात आती है तब कोई भोजनालय हमारे छात्रावास के भोजनालय की बराबरी नहीं कर सकता

Test1 : मुझे यकीन है जब खाद्य पदार्थों में विविधता की बात आती है कोइ भोजनालय हमारे छात्रावास के भोजनालय बराबरी नहीं कर सकता

Test2 : जब खाद्य पदार्थों में विविधता की बात आती है बराबरी नहीं कर सकता मुझे यकीन है कोई भोजनालय हमारे छात्रावास के भोजनालय

GLOSS: मुझे:I, यकीन है:believe, कि:that, जब:when, खाद्य पदार्थों:food, में:in, विविधता:variety, के,की:of, बात:issue, आती है:comes, तब:then, कोई:any, भोजनालय:mess, हमारे:our, छात्रावास:hostel, बराबरी:comparision, नहीं:not, कर सकता:can do.

The quality of Test1 translation is much better than that of Test2 but BLEU gives same score to Test1 and Test2.

5. **Geometric Averaging of n-grams** : Geometric averaging of n-gram scores produces a zero result whenever any of the individual n-gram scores are zero. As a result, sentence-level BLEU scores are highly unreliable (Liu et al., 2005; Liu and Gildea., 2006). For example,
Reference : उसका दोस्ताना व्यवहार अचानक क्रोध में बदल गया

Test : उसका दोस्ताना व्यवहार क्रोध में अचानक बदल गया

GLOSS: उसका:his, दोस्ताना:friendly, व्यवहार:behaviour, अचानक:suddenly, क्रोध:anger, में:in, बदल गया:transformed

BLEU precision metric :
1-gram precision : P1 = 8/8
2-gram precision : P2 = 4/7
3-gram precision : P3 = 1/6
4-gram precision : P4 = 0/5
BLEU precision score = 0 (weighted geometric average)

6. **Equal weightage** : BLEU equally weights all items in the reference sentences (Babych and Hartley., 2004). Therefore omitting content-bearing lexical items does not carry a greater penalty than omitting function words.

3 Advantages of METEOR for English-Hindi MT Evaluation

Following are the advantages of using METEOR metric for English-Hindi MT Evaluation :

1. METEOR addresses the problem of reference translation variability by utilizing flexible word matching, allowing for morphological variants and synonyms to be taken into account as legitimate correspondences. So we have the power of using linguistic tools like Hindi Morphological Analyzer, Hindi Wordnet etc.

2. METEOR uses unigram matching for lexical similarity hence does not rely totally on word order, thus suitable for Hindi which is a relatively free word order language.

3. METEOR uses and emphasizes recall in addition to precision, a property that has been confirmed by several metrics as being critical for high correlation with human judgments (Lavie et al., 2004).

4. The feature ingredients within METEOR are parameterized, allowing for the tuning of the metrics free parameters in search of values that result in optimal correlation with human judgments. Optimal parameters can be separately tuned for different types of human judgments and for different languages, thus suitable for Hindi (Lavie and Denkowski., 2009).

4 Important Aspects in English-Hindi MT

1. Word-order does not strictly convey the grammatical roles of words in a sentence. Hence, it is not the single most important criteria in relatively free word order languages like Hindi.
Example :
Reference : राम ने मोहन को पुस्तक दी

Test1 : राम ने मोहन को पुस्तक दी

Test2 : मोहन को राम ने पुस्तक दी

GLOSS : राम:Rama, ने:subj-marker, मोहन:Mohan, को:to, पुस्तक:book, दी:gave

Test1 and Test2 have different word orders but both are syntactically correct and have the same meaning, hence should be given similar scores.

2. What is important in English-Hindi MT evaluation is not just the existence of function words (such as "ने", "को", etc. in the above example) but whether they are correctly associated with their corresponding content words or not. The grammatical role of "राम" as the subject and "मोहन" as the object in both sentences [Test1 and Test2] comes from the case markers "ने" and "को". Therefore, even though Hindi is predominantly SOV in its word-order, correct case marking is a crucial part for translations to convey the same meaning.

3. Indian languages are morphologically rich. The following example illustrates the richer morphology of Hindi compared to English :

The word "boys" in English is translated as "लड़के" or "लड़कों" depending on the sentence. For example, in Sentence1 given below, the translation is "लड़के" while in Sentence2, the translation is "लड़कों".

Sentence1: The boys are playing cricket
लड़के क्रिकेट खेल रहे हैं

Sentence2: The boys brought the book from the market
लड़कों ने बाजार से किताब खरीदी

4. An English word can be translated to different Hindi words having the same meaning.

The following example illustrates the importance of synonym matching in English-Hindi MT:

English : I believe in God

Reference : मैं भगवान में विश्वास रखता हूँ

Test1 : मैं देवता में विश्वास रखता हूँ

Test2 : मैं अल्लाह में विश्वास रखता हूँ

Test3 : मैं खुदा में विश्वास रखता हूँ

GLOSS : मैं:I, भगवान , देवता , अल्लाह , खुदा:God, में:in, विश्वास:believe, रखता हूँ:keep

All these sentences should receive similar scores.

5 Tools Used

- Morph Analyzer (Hindi Morph 2.5.2)[1] : Given a hindi word, the morphological analyzer identifies the root and the grammatical features of the word like category, gender, number, person, case, vibhakti, Tam(Tense, Aspect, Modality).

- Hindi Wordnet (Hindi Wordnet 1.2) (Jha et al., 2001) : The Hindi WordNet is a system for bringing together different lexical and semantic relations between the Hindi words. In the Hindi WordNet the words are grouped together according to their similarity of meanings. For each word there is a synonym set, or synset, in the Hindi WordNet, representing one lexical concept. The Hindi WordNet deals with the content words, or open class category of words. Thus, the Hindi WordNet contains the following category of words- Noun, Verb, Adjective and Adverb.

- CRF Part of Speech (POS) Tagger (PVS and G, 2007) : POS tagger assigns part of speech tags to each word in the Hindi sentence. Identification of the parts of speech tags such as nouns, verbs, adjectives, adverbs for each word of the sentence helps in analyzing the role of each constituent in a sentence.

- Hindi Local Word Grouper (Bharati et al., 1998) : Local word grouper does the technical task of vibhakti computation.The main task here is to group the function words with the content words based on local information.

- Hindi Clause Boundary Identifier[2] : Traditionally, a clause is defined as a phrase containing at least a verb and a subject. It can be an independent clause or a dependent clause, based on whether it can stand alone when taken in isolation or not respectively. By the definition itself, the words inside a clause form a set of modifier-modified relations, thereby forming a meaningful unit, like a sentence. This makes most of the dependents of the words in a clause to be the words in the same clause, or we can say that the dependencies of the words in a clause are localized to the clause boundary. The taks of clause boundary identifier is to divide the given sentence into a set of clauses.

6 The METEOR-Hindi METRIC

6.1 METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee and Lavie., 2005) is a system that automatically evaluates the output of machine translation engines by comparing them to one or more reference translations. METEOR creates a word alignment between the two sentences, (1) the machine-produced candidate translation string, and (2) the human-produced reference translation string. The alignment is defined as a set of mappings between the words of the sentence pair, such that every word in each string maps to zero or one word in the other string, and no words in the same string. This alignment is created incrementally through a series of word mapping modules called, (1) Exact Matching, (2) Stem Matching, and (3) Synonym Matching. Each module only maps words that have not been mapped to any word in any of the preceding modules. After obtaining the final alignment, the score is computed as the harmonic mean of unigram precision and recall. Additional penalties are computed to directly capture how well-ordered the matched words in the machine translation are in relation to the reference. If more than one reference translation is available, the translation is scored against each reference independently, and the best scoring pair is used.

In their recent work, a matching module has been added which does Paraphrase match between words and phrases (Denkowski and Lavie.,

[1] http://ltrc.iiit.ac.in

[2] Developed at IIIT-Hyderabad as part of the Indian-to-Indian Languages Machine Translation Project

2010a). A new version of METEOR has been recently released "METEOR-Next" with an improved evaluation support (Denkowski and Lavie., 2010b).

6.2 METEOR-Hindi Aligner

We have extended the implementation of METEOR (METEOR 1.0) to support evaluation of translations into Hindi. As the properties of other Indian languages are very similar to those of Hindi, METEOR-Hindi can be easily extended to other Indian languages. For adapting METEOR to a new target language, two language-specific components need to be addressed. They are,

- Language-specific word matching module
- Language-specific parameters and the corresponding scoring function (Lavie and Agarwal., 2007).

The word-alignment algorithm for Hindi is the same as used in current METEOR. However, Hindi-specific tools need to be used. We created a new "stemming" module for Hindi. In this module, a hindi morph analyzer[3] is used for stemming. Synonym matches are detected using synonym sets from the Hindi WordNet version 1.2 (Jha et al., 2001). An example alignment is shown in Figure 1.

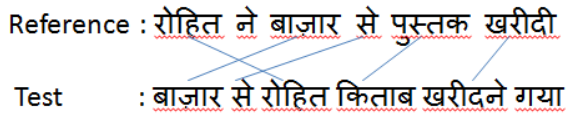

Figure 1: Example Alignment

6.3 METEOR-Hindi Scoring Function

Indian Languages are relatively free word order languages and are morphologically rich. For scoring the system outputs, apart from using only the word-based features, we also use other linguistic parameters such as local word groups, part-of-speech tags and clause boundaries. We motivate the use of these additional parameters in this section.

(1) **Local Word Group (LWG) match** : Local word groups (LWGs) (Bharati et al., 1998; Vaidya et al., 2009) consist of a content word and its associated function words. The function words assign the grammatical role information to their corresponding content words. Hence, the matching of the entire local word group conveys that the content words in both the reference and the system output are used for the same grammatical purpose. Hence, scoring based on the number of matching local word groups is crucial to convey the similarity of the reference sentence and the system output.

In METEOR, fragmentation penalty is used to compute the word-order similarity of the reference sentence and the system output. However, word order is not as important as the correct association of the case markers with their corresponding content words.

Hence, if the local word groups in the candidate does not match with the local word groups present in the reference, it should be given a lower score. METEOR-1.0 is limited to unigram matches, making it strictly a word-level metric. By taking into account LWG, we are going beyond word-level. The following example illustrates the importance of LWG : (Bharati et al., 1991)

Reference : बिल्लि ने कुत्ते को मारा
Test1 : कुत्ते को बिल्लि ने मारा
Test2 : कुत्ते ने बिल्लि को मारा
GLOSS : बिल्लि:cat, कुत्ते:dog, ने:subj-marker, को:object-marker, मारा:killed

Local Word Groups :
Reference : "बिल्लि ने" , "कुत्ते को" , "मारा"
Test1 : "कुत्ते को" , "बिल्लि ने", "मारा"
Test2 : "कुत्ते ने" , "बिल्लि को", "मारा"

Though all the words get matched in the "exact match" stage in both the test sentences but Test1 has the same meaning as Reference while Test2 means entirely the opposite. We can make out this difference through local word group which has a lesser match in Test2 compared to Test1. Hence, Test2 should be given lower score than Test1.

(2) **Part-of-Speech (POS) match** : POS refers to the lexical category of each word in a sentence on the basis of its context. In this parameter, we not just compute the number of matching words, but compute the number of matching words with same POS tag. Hence, those words are not considered

[3]Developed at IIIT-Hyderabad as part of the Indian-to-Indian Languages Machine Translation Project

as matched that have different syntactic category (conveyed through POS tags) in their respective sentences. The goal is, therefore, to capture the proportion of lexical items correctly translated, according to their shallow syntactic realization (Gimnez., 2008).

Sentences are automatically annotated using a CRF based POS Tagger[4]. The following example illustrates the importance of POS tags based scoring parameter for automatic MT evaluation having Hindi as the target language.

Reference : राम खेल रहा है
Test : खेल राम रहा है
GLOSS: राम:Ram, खेल रहा है:is playing

Reference	*POS*	*Test*	*POS*
राम	NN	खेल	NNP
खेल	VM	राम	NNP
रहा	VAUX	रहा	VM
है	VAUX	है	VAUX

Table 1: Assigned POS tags

In Table 1, all the words have matched but have different POS tags. For example, खेल in the reference had a POS tag 'VM' whereas, in the system output, it has a tag 'NNP'. Hence, it is not considered for computing the parameter score.

(3) **Clause match** : Traditionally, a clause is defined as a phrase containing atleast a verb and a subject. We used clause boundary identifier to find the clauses in the test and reference sentences. We then compute the exactly matching clauses between reference and the system output. The following example illustrates the idea.

Reference : रोहित स्कूल जाकर कूदने लगा
Test : रोहित स्कूल जाकर खेलने लगा
Gloss : रोहित:Rohit, स्कूल:school, जाकर:after going, कूदने:jumping, खेलने:playing, लगा:started

In Table 2, only one clause gets matched out of two, therefore candidate sentence should be given lower score.

Table 3 lists the complete set of parameters used in METEOR-Hindi.

$$s = \frac{(\sum W_i * f_i)}{(\sum W_i)} \quad (1)$$

[4]Developed at IIIT-Hyderabad as part of the Indian-to-Indian Languages Machine Translation Project

	Reference	*Test*
1.	रोहित स्कूल जाकर	रोहित स्कूल जाकर
2.	कूदने लगा	खेलने लगा

Table 2: Assigned Clauses

Stage	*Features*
Exact Match	Precision
	Recall
Stem Match	Precision
	Recall
Synonym Match	Precision
	Recall
LWG	Precision
	Recall
POS	Precision
	Recall
Clause	Precision
	Recall

Table 3: Parameters used for scoring in METEOR-Hindi

The scores obtained using the parameters are combined as weighted linear sum (see Equation 1). Here, s is the score and W_i is the weight of feature i.

In METEOR, only exact, stem and synonym modules are used with no additional language specific features, so the scoring function in METEOR is very specific. To take into account the extent to which the matched unigrams in the two strings are in the same word order, METEOR computes a penalty for a given alignment as follows. First, the sequence of matched unigrams between the two strings is divided into the fewest possible number of chunks such that the matched unigrams in each chunk are adjacent (in both strings) and in identical word order. The number of chunks (ch) and the number of matches (m) is then used to calculate a fragmentation fraction: frag = ch/m. The penalty is then computed as: Penalty = γ * frag$^{\beta}$.
We do not use chunks in METEOR-Hindi because for Hindi word order is not so important (See Section 4. Point 1). We use a very general equation for scoring in METEOR-Hindi which facilitates the use of standard Machine Learning techniques. As the scoring function is changed than what is currently in METEOR-1.0 for other languages, while calculating Precision and Recall in Stem Match module, words matched in the exact match mod-

ule are also considered as the exact words have the same stem. Similarly for synonym , matches in exact and stem matching modules are also considered while calculating Precision and Recall for Synonym Module.

For better accuracy, while scoring, for LWG, POS and Clause, we are also taking into account stem and synonyms, so "किताब को" and "पुस्तक को" will get matched as LWG, "किताब" and "पुस्तक" will get matched as POS [only if they have the same POS tag in the sentence] and "रमेश घर जाकर" and "रमेश घर गया" will get matched as Clauses.

The second main language-specific issue which required adaptation is the tuning of the 12 parameters within METEOR-Hindi. Due to the unavailability of high-quality training data, currently all the weights are taken to be 1.

7 Experiments and Results

We have evaluated METEOR-Hindi on a dataset of 100 sentences out of which 60 test translations are from the Anglahindi MT-system of IIT Kanpur and 40 are from IIIT-Hyderabad MT-system Shakti. The statistics of the Dataset we have used for our experiments are given in Table 4.

Number of Sentences	100
Avg. Test Sentence length	11.24
Avg. Ref Sentence length	11.23
Exact Matches	433
Stem Matches	574
Synonym Matches	622
LWG Matches	426
POS Matches	576
Clause Matches	9

Table 4: Dataset Statistics

Reference sentences and scores have been assigned by native hindi speakers with some knowledge of linguistics. The human judges rated each translation from 0 to 4. The following rating scheme was provided to them (Ramanathan et al., 2008):

The evaluation-metric for MT is evaluated by comparing the scores given by the metric with the scores provided by human raters. The comparision is performed using Pearson product-moment correlation. We normalized the human scores on a scale of 0 to 1. The METEOR-Hindi scores provided using different features on each sentence are calculated and the correlation is computed using the formula given in Equation 2.

$Rating$	$Translation - Quality$
4	Perfect
3	Good
2	Understandable
1	Roughly understandable
0	Nonsense

Table 5: Human Rating Criteria

$$r = \frac{(N\sum_{i=1}^{N} X_i Y_i - (\sum_{i=1}^{N} X_i)(\sum_{i=1}^{N} Y_i))}{(\sqrt{N\sum_{i=1}^{N} X_i{}^2 - (\sum_{i=1}^{N} X_i)^2}.\sqrt{N\sum_{i=1}^{N} Y_i{}^2 - (\sum_{i=1}^{N} Y_i)^2})} \quad (2)$$

where, N is the number of sentences, X_i is the metric's score on i^{th} sentence, Y_i is the human score on i^{th} sentence (see Table 6).

The reason of such a low correlation with BLEU (see Table 6) is that it gives score 0 in most of the sentences (see Figures 2 and 5). METEOR-Hindi assigns scores that are much closer to Human score as compared to BLEU (see Figures 3 and 4).

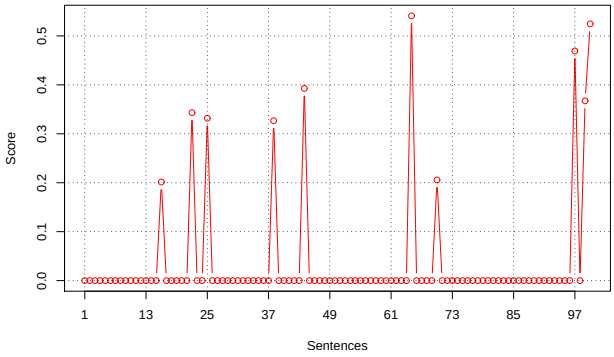

Figure 2: Output of BLEU

Highest Pearson product-moment **correlation of 0.703** is achieved using **Exact, Stem, Synonym and POS features**. Using linguistic features such as "stemming" and "synonym" have resulted in a better correlation.

However, surprisingly, the local word groups and clause based feature did not show an increase in correlation. To decipher the reason, we analyzed the test data. We found that there are a number of errors in the reference sentences like

Metric	*Features*	*Pearson − correlation*
BLEU	-	**0.27127353**
METEOR-Hindi	Exact	0.65626916
METEOR-Hindi	Exact + Stem	0.68747816
METEOR-Hindi	Exact + Stem + Synonym	0.70005324
METEOR-Hindi	Exact + Stem + Synonym + LWG	0.68148324
METEOR-Hindi	Exact + Stem + Synonym + POS	**0.70312150**
METEOR-Hindi	Exact + Stem + Synonym + Clause	0.65807648
METEOR-Hindi	Exact + Stem + Synonym + LWG + POS + Clause	0.66681955

Table 6: Correlation coefficients using different features

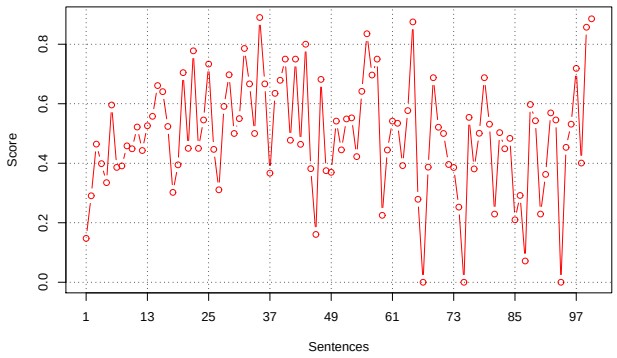

Figure 3: Output of METEOR-Hindi

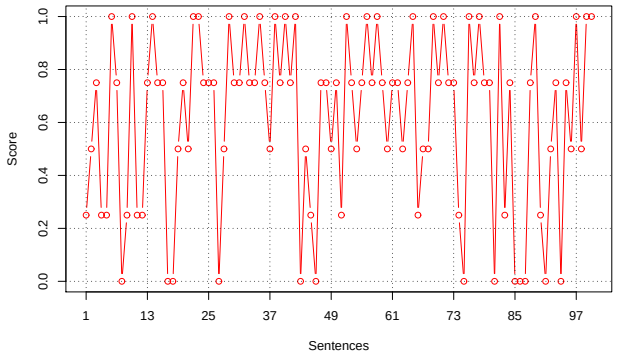

Figure 4: Human assigned scores

"हार कि" while it should be "हार की", "उनमे से" instead of "उनमें से" , "गाडियों मे" instead of "गाडियों में" due to which number of local word groups which matched decreased.

Adding clause feature also decreses the correlation as only 9 clauses get matched in 100 sentences as most of the sentences have only one verb, so by definition the whole sentence is a clause and it is very rare for the entire test sentence to match the reference. As clause is a much higher level concept than words, score should be less penalized if the clauses do not match. The correlation achieved when only clause feature is used is 0.297 comparable to BLEU because of zero match in most of the sentences. From Figure 6, it is clear that there is a linear relationship between METEOR-Hindi score and Human score except for some of the sentences where METEOR-Hindi assigns a score greater than zero when Human score is zero.

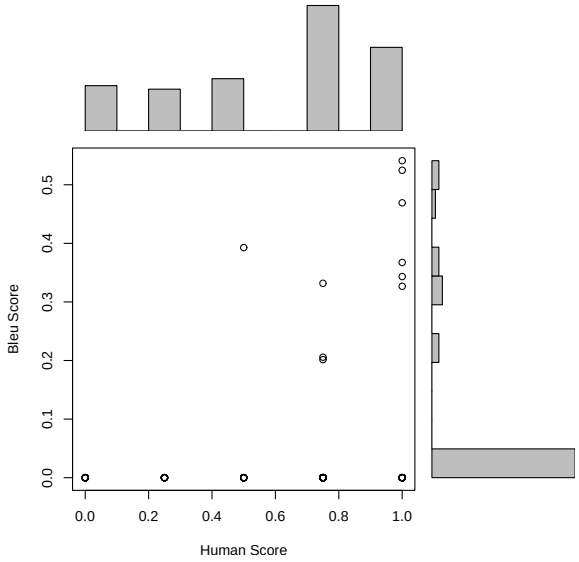

Figure 5: Scatterplot of BLEU and Human Score

The average scores provided by the metrics BLEU and METEOR-Hindi are compared with the average score provided by Human in Table 7. We observe that the difference between the average scores of METEOR-Hindi and Human raters is very less.

Metric	*Average − Score*
BLEU	0.0815
METEOR-Hindi	0.4919
Human	0.615

Table 7: Average scores

To see whether we can use METEOR-Hindi

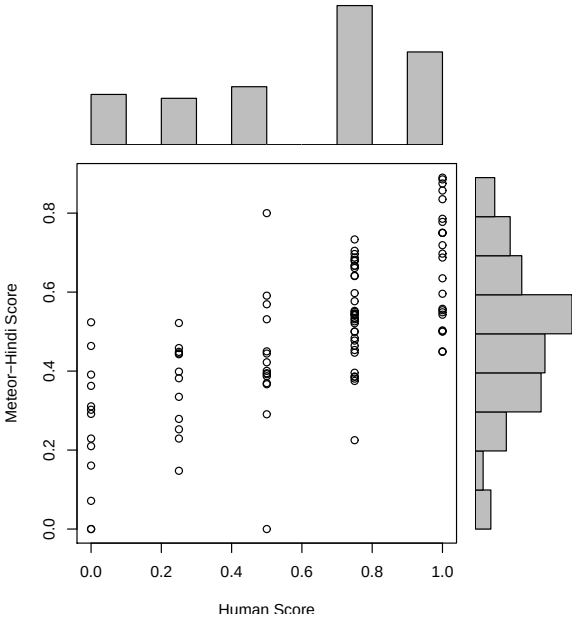

Figure 6: Scatterplot of METEOR-Hindi and Human Score

to compare two MT systems, we ran BLEU and METEOR-Hindi separately on the test sentences from IIT-Kanpur and IIIT-Hyderabad MT system. The results are given in Table 8.

We see that the METEOR-Hindi correlated better (as compared to BLEU) with the judgement of Human annotators. Human annotators as well as METEOR-Hindi had given greater score to the IIIT-H system, while BLEU gave greater score to IIT-K system.

We present some of the examples where METEOR-Hindi performed better than BLEU.

English Sentence : I grabbed the book
Reference : मैंने पुस्तक को पकडा
Test : मैंने किताब को पकडा
Gloss : मैंने:I, पुस्तक,किताब:book, को:object-marker, पकडा:grabbed

n-gram matches : unigrams : 3/4 ; bigrams : 1/3 ; trigrams : 0/2 ; 4-grams : 0/1
BLEU Score : 0.0
METEOR-Hindi Score[5] : 0.875
Human Score : 1.0

English Sentence : The ten best Aamir Khan performances (Ramanathan et al., 2007)
Reference : आमिर खान की दस सर्वोत्तम परफोर्मेनसस
Test : दस सर्वोत्तम आमिर खान परफोर्मेनसस
Gloss: आमिर खान:Aaamir Khan, दस:ten, सर्वोत्तम:best, परफोर्मेनसस:performances, की:of

n-gram matches : unigrams: 5/5; bi-grams: 2/4; trigrams: 0/3; 4-grams: 0/2
BLEU Score : 0.0
METEOR-Hindi Score[6] : 0.8708
Human Score : 0.9

8 Conclusions

In this paper, we presented METEOR-Hindi, an automatic evaluation metric for a machine translation system where the target language is Hindi. METEOR-Hindi is a modified version of the metric METEOR, containing features specific to Hindi. We made appropriate changes to METEOR's alignment algorithm and the scoring technique.

In our experiments, we observed that METEOR-Hindi achieved high correlation of 0.703 with human judgments significantly outperforming BLEU that had a correlation of only 0.271. We have highlighted the issues using BLEU for English-Hindi MT evaluation. Also, linguistic features have been identified as being useful in evaluation. We plan to train METEOR-Hindi on a large amount of high-quality data and using features like paraphrase match to achieve better correlation.

References

Bogdan Babych and Anthony Hartley. 2004. Extending the bleu mt evaluation method with frequency weightings. In *Proceedings of ACL.*

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaulation Measures for MT and/or Summarization.*

Akshar Bharati, Vineet Chaitanya, and Rajeev Sangal. 1991. Local word grouping and its relevance to indian languages. In *Frontiers in Knowledge Based Computing (KBCS90), V. P. Bhatkar and K. M. Rege (eds.), Narosa Publishing House, New Delhi, 1991, pp. 277–296.*

Akshar Bharati, Medhavi Bhatia, Vineet Chaitanya, and Rajeev Sangal. 1998. Paninian grammar framework applied to english. *South Asian Language Review*, (3).

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *Proceedings of the EACL.*

[5]Using 4 features : Exact, Stem, Synonym, POS

[6]Using 4 features : Exact, Stem, Synonym, POS

$Metric$	$System$	$Average - Score$	$Pearson - correlation$
BLEU	IIT-K	0.041	0.343
METEOR-Hindi	IIT-K	0.460	0.712
Human	IIT-K	0.567	-
BLEU	IIIT-H	0.032	0.163
METEOR-Hindi	IIIT-H	0.540	0.684
Human	IIIT-H	0.688	-

Table 8: MT System Comparison

Michael Denkowski and Alon Lavie. 2010a. Extending the meteor machine translation evaluation metric to the phrase level. In *Proceedings of NAACL/HLT, 2010.*

Michael Denkowski and Alon Lavie. 2010b. Meteor-next and the meteor paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR, 2010.*

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurence statistics. In *Human Language Technology: Notebook Proceedings, pages 128132.*

Jess Gimnez. 2008. Empirical machine translation and its evaluation. In *Ph.D. Thesis, Universitat Politcnica de Catalunya (defended July 2, 2008).*

S. Jha, D. Narayan, P. Pande, and P. Bhattacharyya. 2001. A wordnet for hindi. In *International Workshop on Lexical Resources in Natural Language Processing, Hyderabad, India, January, 2001.*

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second ACL Workshop on Statistical Machine Translation, pages 228231, Prague, Czech Republic, June.*

Alon Lavie and Micheal J. Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. In *Machine Translation*.

Alon Lavie, Kenji Sagae, and Shyamsundar Jayaraman. 2004. The significance of recall in automatic metrics for mt evaluation. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004), pages 134143, Washington, DC, September.*

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. Cder: Efficient mt evaluation using block movements. In *Proceedings of the Thirteenth Conference of the European Chapter of the Association for Computational Linguistics.*

Ding Liu and Daniel Gildea. 2006. Stochastic iterative alignment for machine translation evaluation. In *Proceedings of the COLING/ACL on Main conference poster sessions, pages 539546, Morristown, NJ, USA. Association for Computational Linguistics.*

Liu, D., and D. Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.*

Chang Liu, Daniel Dahlmeier, and Hwee Tou Ng. 2010. Tesla : Translation evaluation of sentences with linear-programming based analysis. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 329-334, Uppsala, Sweden, July. Association for Computational Linguistics, 2010.*

I. Dan Melamed, Ryan Green, and Joseph Turian. 2003. Precision and recall of machine translation. In *Proceedings of the HLT-NAACL 2003 Conference: Short Papers, pages 6163, Edmonton, Alberta.*

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.*

Avinesh PVS and Karthik G. 2007. Part-of-speech tagging and chunking using conditional random fields and transformation based learning. In *Proceedings of the SPSAL workshop during IJCAI, 2007.*

Ananthakrishnan Ramanathan, Pushpak Bhattacharyya, M. Sasikumar, and Ritesh M. Shah. 2007. Some issues in automatic evaluation of english-hindi mt: More blues for bleu. In *proceeding of 5th International Conference on Natural Language Processing(ICON-07), Hyderabad, India.*

Ananthakrishnan Ramanathan, Pushpak Bhattacharyya, Jayprasad Hegde, Ritesh M. Shah, and Sasikumar M. 2008. Simple syntactic and morphological processing can help english-hindi statistical machine translation. In *Proceedings of IJCNLP, 2008.*

Durgesh Rao. 2001. Machine translation in india: A brief survey. In *Proceedings of SCALLA 2001 Conference, Bangalore.*

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. 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 (AMTA-2006), pages 223231, Cambridge, MA, August.*

Ashwini Vaidya, Samar Husain, Prashanth Reddy, and Dipti M Sharma. 2009. A karaka based annotation scheme for english. In *Proceedings of CICLing , 2009.*