MEE : An Automatic Metric for Evaluation Using Embeddings for Machine Translation

by

Ananya Mukherjee, Hema Ala, Manish Shrivastava, Dipti Misra Sharma

Report No: IIIT/TR/2020/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
January 2020

MEE : An Automatic *M*etric for *E*valuation Using *E*mbeddings for Machine Translation

Ananya Mukherjee[1], Hema Ala[1], Manish Shrivastava, Dipti Misra Sharma
Language Technology Research Centre, IIIT - Hyderabad, India
ananya.mukherjee@research.iiit.ac.in, hema.ala@research.iiit.ac.in, m.srivastava@iiit.ac.in, dipti@iiit.ac.in

Abstract—**We propose MEE, an approach for automatic Machine Translation (MT) evaluation which leverages the similarity between embeddings of words in candidate and reference sentences to assess translation quality. Unigrams are matched based on their surface forms, root forms and meanings which aids to capture lexical, morphological and semantic equivalence. We perform experiments for MT from English to four Indian Languages (Telugu, Marathi, Bengali and Hindi) on a robust dataset comprising simple and complex sentences with good and bad translations. Further, it is observed that the proposed metric correlates better with human judgements than the existing widely used metrics.**

Index Terms—**MT Evaluation, Automatic Metrics, Semantic Evaluation, Morphological Languages, Embeddings**

I. Introduction

Machine Translation (MT) is a task of converting the source text in one natural language to another natural language by preserving both faithfulness and fluency. Neural Machine Translation (NMT) has emerged as the most compelling approach to perform this task. With research in recent times, there are several adaptations of NMT currently being deployed. Therefore, it is essential to recognize a better performing system which can be achieved by evaluating the translated outputs. MT output can be evaluated by both humans and automatic evaluation metrics. Human judgements are more reliable but they are expensive, time-consuming and require a lot of effort for every evaluation. For such evaluation, evaluators must possess bilingual domain knowledge. Therefore, an automatic metric is chosen as it is fast, easy to run and has a high correlation with human judgements [1].

Widely used automatic evaluation metrics follow the idea of n-gram matching between a candidate and reference sentence like BLEU [2], NIST [3], etc. N-gram based metrics might fail to correlate well with human judgements for morphologically rich and free word order languages, as they fail to account semantic similarity. Moreover, lack of morphological processing often penalizes alternative word forms. We present a metric which provides means to evaluate candidate sentences on semantic similarity.

In this paper we use pre-trained fastText Embeddings [4] provided by Facebook. Our metric evaluates the MT output sentence (candidate sentence) using reference sentence (sentences from standard data-set mentioned in Section IV). We experiment our approach for MT from English to four Indian Languages (Telugu, Marathi, Bengali and Hindi) and compare evaluation scores with human judgements at segment level. This metric computes evaluation score using three modules namely exact match, root match and synonym match. In each module, fmean-score is calculated using harmonic mean of precision and recall by assigning more weightage to recall. We evaluate the final translation score by taking average of fmean-scores from individual modules. Using final translation scores and human judgements, we obtain correlation. When compared to other metrics, our metric correlates better with human judgments.

We are currently in the process of exploring several improvements to our proposed metric, which we believe to possess the potential to significantly enhance its sensitivity and correlation with human judgments. Our work on these directions is further discussed in Section VI.

II. Background and Motivation

BLEU is considered to be a default automatic evaluation metric till date. It counts the matching n-grams in candidate sentence to n-grams in reference sentence and uses modified n-gram precision. In spite of BLEU being the widely used metric for automatic MT evaluation, many researchers have explored and stated several limitations which depicted the inability of BLEU to capture the original quality dimensions of the translation output [5]. Also, BLEU is proven to be neither necessary nor sufficient for achieving improvements in translation output quality [6].

To overcome the limitations of BLEU, a new metric NIST [3] with some alterations to BLEU was proposed. While BLEU simply calculates n-gram precision adding equal weight to each n-gram, NIST calculates how informative a particular n-gram is which means it assigns more weight to rarer n-grams.

While BLEU and NIST prove to be simple and generic automatic evaluation metrics, they fail to capture the lexical and semantic diversities. Moreover, BLEU and NIST are purely based on modified n-gram precision along with Brevity Penalty which penalizes ”too short” translations. But this doesn’t make up for the absence of recall. There

[1] Equal contribution

are several other reference-based methods for automatic evaluation which provide a mechanically determined score of string similarity between candidate sentence (translated output) and reference sentence. If the candidate sentence contains same words in same order as the reference, it receives a high score else if other words are present or appear in a different order it receives a lower score. Hence, having only an *n-gram syntactic evaluation* is not sufficient for languages which are morphologically rich and follow relatively free word order. The tacit assumption is that translation quality can be measured based on similarity to human translation and thus, semantic evaluation is highly required [7].

METEOR [8] is one such metric that performs semantic evaluation using linguistic tools. METEOR modifies the precision and recall using weighted F-score based on unigram matching. This is done by exact matching, root matching, synonym matching, etc. Similar work was done for one Indian language in METEOR-Hindi [9] with few modifications to *alignment algorithm* and to the *scoring technique* of METEOR. As Indian Languages tend to follow relatively *Free Word Order*, the *Fragmentation Penalty* which is used in METEOR to compute the word-order similarity of candidate sentence and reference sentence is not used in METEOR-Hindi. METEOR has been extended in several ways. One such extension depicts to support evaluation of MT output in Spanish, French and German, in addition to English [10]. The latest being METEOR Universal [11] which claims to outperform baseline BLEU for two new languages, Russian and Hindi by (1) automatically extracting linguistic resources from the bitext used to train MT systems and (2) using a universal parameter set learned from pooling judgments from several languages.

Both METEOR and METEOR-Hindi consider morphological features of the target language using language specific resources like WordNet, Stemmer, etc. Despite this, METEOR is not chosen as an evaluation metric for low resource languages as it demands high-level stemmers and synonym extractors for these languages.

Hence, we propose an approach that can be easily used to syntactically and semantically evaluate the MT outputs using fastText Embeddings. FastText embedding is an extension of Mikolov's embedding [12] and is based on the skipgram model where each word is represented as a bag of character n-grams [13], [14]. Each character n-gram is associated with a vector representation and words are represented as sum of these vectors. Further, the word representation is learned by considering a large window of left and right context words. Unlike Mikolov's embeddings, fastText can provide an embedding for rare words, misspelled word, or words that are not present in the training corpus. This is because fastText uses character n-gram word tokenization.

FastText embeddings are available for 157 languages.[1] These pre-trained embeddings are used to perform synonym match and root match. We carry out our experiments for four Indian Languages (Telugu, Marathi, Bengali and Hindi) which are rich in morphology and follow relative free word order. Later, compare the scores of proposed metric with BLEU, NIST and METEOR.[2] We present the correlation of human judgements with experimental results of various metrics which demonstrates that our proposed approach outperforms the existing baseline metrics.

III. Approach

Our approach (MEE) evaluates the candidate sentence by computing a score based on explicit unigram matching using reference sentences. To achieve semantic evaluation, we use fasttext embeddings [4] provided by Facebook to calculate the word similarity score between the candidate and the reference words.

A. *MEE:* M*etric for* E*valuation using* E*mbeddings*

MEE contains three modules namely Exact Match, Root Match and Synonym Match. Segment level evaluation is done by computing scores at sentence level. For a given candidate sentence, fmean-scores are calculated in each module individually based on unigram match counts with reference sentence. Final score (fs) of the candidate sentence is obtained by taking a weighted average of all individual module scores (fmean-scores : $fm1$, $fm2$, $fm3$). In our experiment, all modules are assigned with equal weights. However, these weights can be tuned as required for each language pair.

Exact Match: Each unigram in candidate sentence is checked for an identical surface form match in reference sentence. Then total number of 'exact' unigram matches is used to calculate Exact Match module fmean-score $(fm1)$. Each unigram is matched only once.

Root Match: In this module, we find matches based on unigram similarity scores using fasttext embeddings. For each unigram pair, between candidate and reference sentence, a similarity score is obtained. Based on the threshold set for this module, we retrieve unigram root matches. This helps to identify morphological variants in a candidate sentence. Exact matches also contributes towards root match *count.* Further, this *count* is used to calculate the Root Match module fmean-score $(fm2)$.

Synonym Match: Here, the unigram pairs of candidate and reference sentences are matched to check for synonyms. Synonym match is based on a threshold different from that of root match. Semantically equivalent word choices are captured in this module. Exact matches and root matches contribute towards the synonym match count. Using this count, fmean-score $(fm3)$ is computed.

[1] https://fasttext.cc/docs/en/crawl-vectors.html

[2] We have re-implemented METEOR for Telugu, Marathi, Bengali and Hindi based on METEOR and METEOR-Hindi.

Table I: Examples with Word Similarity Scores.
S:Synonym (ranging from 0.4-0.5) R:Root (greater than 0.5)

Language		Word1,Word2	Score	English Equivalent
Telugu	S	తెలపండి, పేర్కొనండి telapaṇḍi, pērkonaṇḍi	0.42	to mention
		పళ్ళు, దంతాలు paḷḷu, dantālu	0.41	teeth
	R	శరీరం , శరీరంలో śarīraṁ, śarīranlō	0.59	body, in the body
		తీసుకోవడం, తీసుకోవాలి tīsukōvaḍaṁ, tīsukōvāli	0.56	act of taking, has to take
Marathi	S	साफ, स्वच्छ saaf, svaccha	0.4	clean
		खूप, भरपूर khupa, bharapūra	0.47	plenty
	R	नये, नका nayē ,nakā	0.63	not, do not
		आहे , आहेत āhē, āhēta	0.6	is, are
Bengali	S	কমজোর , দুর্বল kamajōra, durbala	0.4	weak
		সুতরাং ,তাহলে sutarāṁ, tāhalē	0.47	so, then
	R	নিয়ন্ত্রণের , নিয়ন্ত্রণে niẏantraṇēra, niẏantraṇē	0.55	of control, in control
		রাখতে, রাখার rākhatē, rākhāra	0.72	to keep
Hindi	S	गलने, पिघलने galane, pighalane	0.4	melting
		प्रारंभिक , प्राथमिक praarambhik, praathamik	0.44	initial, primary
	R	दवाइयों, दवाएं davaiyon, davaen	0.51	medicines, the medicines
		पहनने, पहने pahanane, pahane	0.63	wear

We have empirically observed that the similarity score of synonyms often ranges between 0.4 - 0.5. Likewise, the similarity score of any given word and its root, results to be greater than or equal to 0.5 (refer Table I for examples). For this reason, we adapted these thresholds in our experiments.

B. MEE Scoring

The translation score of a candidate sentence is computed as follows. Based on the number of matched unigrams (m), the total number of unigrams in candidate sentence (t) and the total number of unigrams in reference sentence (r), we calculate unigram precision $P = m/t$ and unigram recall $R = m/r$. For example, in exact match module, P and R are computed using the count of exact matches between the candidate and reference sentence. Individual fmean-scores are calculated for each module. This is achieved by parameterized harmonic mean of Precision and Recall [15] as mentioned in (1).

$$F_\beta = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \tag{1}$$

In (1), β is a parameter that controls a balance between P and R. When β= 1, F1 equals to the harmonic mean of P and R. If $\beta > 1$, F becomes more recall-oriented and if $\beta < 1$, it becomes more precision-oriented. We consider β=3 for fmean-score calculation (2). Recall is weighted nine times more heavily than precision as recall has a high significance in automatic MT evaluation [16].

$$F3 = \frac{10PR}{9P + R} \tag{2}$$

The calculation of final translation score of a candidate sentence is clearly described using an example mentioned in Table II. Also, matched unigrams (m), unigram precision (P), unigram recall (R) and fmean-scores (2) for each module is reported in Table III.

Table II: Example of a Telugu Sentence.

	Reference Sentence	Candidate Sentence
Telugu Sentence	పొగతాగడంతో కాన్సర్ తో పాటు చాలా వ్యాధులు వచ్చే ప్రమాదం ఉంటుంది .	ధూమపానంతో క్యాన్సర్‌తో సహా అనేక వ్యాధుల ప్రమాదం ఉంది.
Transliteration	Pogatāgaḍantō kānsar tō pāṭu cālā vyādhulu vaccē pramādaṁ uṇṭundi.	Dhūmapānantō kyānsartō sahā anēka vyādhula pramādaṁ undi.
English Equivalent	Smoking might have a risk of many diseases, along with cancer.	Smoking has a risk of many diseases, including cancer.

For example in Table II, *Exact Match* module identifies **one** exact match for the unigram pair [('ప్రమాదం', 'ప్రమాదం')]. *Root Match* module identifies **three** unigram pairs satisfying the root match criteria [('క్యాన్సర్‌తో', 'కాన్సర్'), ('వ్యాధుల', 'వ్యాధులు'), ('ఉంది', 'ఉంటుంది')] resulting in total module match count as **four** (Exact matches are considered in root match module. Thus, 1 + 3 = 4). The synonym matches of these **two** pairs [('సహా', 'పాటు'), ('అనేక', 'చాలా')] contributes towards *Synonym Match* module match count as **six** (Exact matches and root matches are considered in this module. Therefore, 1 + 3 + 2 = 6).

Table III: Module Wise Scores for the Example Sentence in Table II.

	Exact Match	Root Match	Synonym Match
Unigram Matches (m)	1	4	6
Unigram Precision (P)	0.13	0.5	0.75
Unigram Recall (R)	0.11	0.44	0.67
fmean-score	fm1= 0.11	fm2= 0.45	fm3= 0.68
Final Score (fs)	0.41		

$$fs = \frac{\sum_{j=1}^{N} w_{\text{j}} * fm_{\text{j}}}{\sum_{j=1}^{N} w_{\text{j}}} \tag{3}$$

Final translation score (fs) is calculated using (3) where w_{j} is weight of j^{th} module, fm_{j} is Fmean score of j^{th} module and N being the total number of modules ($N = 3$). All the three modules are treated with same weightage in our experiment ($w_{\text{j}} = 1$). However, these weights can be tuned as required for each language. This final translation score (fs) of each candidate sentence is used to compute the correlation with human judgements.

IV. Experimental Setup

We evaluate the translation quality using reference sentence for each candidate sentence. BLEU, NIST, METEOR and MEE metrics are used to automatically evaluate the translated output sentences and further, evaluation scores are compared with human judgements using correlation.

A. Dataset

To demonstrate the capability of our metric in terms of semantic and lexical evaluation, we experimented using ILCI (Indian Languages Corpora Initiative) corpus [17] as Indian Languages are rich in morphology and holds word order freedom. We considered three languages from the **Indo-Aryan family** (Hindi, Bengali and Marathi), one from **Dravidian family** (Telugu) and one from the **West Germanic family** (English).

Table IV: Dataset Statistics.

(a) Sentence Lengths.
R:Reference Sentences
G:Google Translated Outputs
B:Bing Translated Outputs

		Sentence Length		
Lang.	Sent.	Min.	Max.	Avg.
Tel	R	2	31	10.04
	G	3	32	9.83
	B	3	34	10.42
Mar	R	3	28	10.17
	G	3	28	9.7
Ben	R	2	34	10.9
	G	2	29	10.8
	B	2	29	10.7
Hin	R	3	48	14.08
	G	3	45	14.17
	B	3	51	14.31

(b) Diversity of Sentences.
Cl. : Clauses

	Tel.	Mar.	Ben.	Hin
Total Sentences	236	248	254	325
Simple Sentences	96	106	105	134
Two Clauses	58	59	68	84
Three+ Clauses	12	9	11	14
Two Verbs	36	40	43	49
Three Verbs	14	22	16	20
Four+ Verbs	98	96	103	134
Wh-Clauses	9	14	15	17
Subordinate Cl.	6	4	4	6
Complement Cl.	19	18	16	22
Relational Clauses	13	14	18	20
Conditional Cl.	8	12	14	15
Adverbial Clauses	85	76	85	113
Cleft Construction	58	67	58	80

Our experiment evaluates the MT output from two MT Systems: *Google Translate* and *Bing Translator*,[3] where English is source language and Telugu, Marathi, Bengali and Hindi are taken as target. Machine translated sentences are the **candidate sentences** and corresponding ILCI parallel corpus sentences are considered as **reference sentences**. Dataset statistics and distribution of sentence

[3] Note that Bing MT system does not support Marathi.

types (Simple and Complex Sentences) for each language is mentioned in Table IV(a) and Table IV(b) respectively.

B. METEOR Re-implementation

We have re-implemented METEOR for Telugu, Marathi, Bengali and Hindi using baseline METEOR [8] and METEOR-Hindi [9]. Drawing inspiration from these METEOR versions, five modules have been implemented namely Exact Match, Root Match, POS-tag match, Synonym match and Chunk match. Language specific resources are used such as shallow parser[4] by IIIT Hyderabad is used to extract roots, POS tags and chunks. Likewise, to extract synonyms, Indo Wordnet by IIT Bombay [18] is used.

C. Automatic Evaluation

Automatic evaluation of MT is easy and fast when compared to human evaluation in terms of manual effort, expense and time.

The translation quality of a candidate sentence is evaluated using a reference sentence (refer IV-A). Segment level evaluation is done using four automatic evaluation metrics namely BLEU, NIST, METEOR (refer IV-B) and MEE (proposed metric). The scores of these metrics are normalized to a range of 0-1. Table V shows normalized scores of BLEU, NIST, METEOR, MEE and human judgements for one example sentence per language.

D. Human Evaluation

Manual assessment is considered as a gold standard for the performance of automatic MT evaluation metrics. The perception of translation quality is subjective and depends on individual background and expectations of the human evaluators. Accordingly, three native speakers per language are chosen for evaluating the candidate sentences. The participants are highly proficient in English in addition to their native language. These judgements are made as per the rating scale [19] mentioned in Table VI. Further, average of three human judgements for each candidate sentence is taken and then normalized to a scale of 0-1 (see Table V).

E. Correlation

Ultimately, to evaluate the performance of our approach we compute the Pearson product-moment correlation (r) [20] between normalized automatic metric scores (A) and normalized human judgements (H), for all sentences (N) using (4). Pearson product-moment correlation is a measure of strength, association, as well as statistical relationship. Thus, if correlation value of an automatic metric with human judgements is high then it can be deduced that the metric correlates well with humans.

$$r = \frac{(N\sum_{i=1}^{N} H_{\mathrm{i}}A_{\mathrm{i}} - (\sum_{i=1}^{N} H_{\mathrm{i}})(\sum_{i=1}^{N} A_{\mathrm{i}}))}{\sqrt{N\sum_{i=1}^{N} {H_{\mathrm{i}}}^2 - (\sum_{i=1}^{N} H_{\mathrm{i}})^2}\sqrt{N\sum_{i=1}^{N} {A_{\mathrm{i}}}^2 - (\sum_{i=1}^{N} A_{\mathrm{i}})^2}} \qquad (4)$$

V. Results and Discussion

The consolidated correlation between normalized scores of averaged human judgements and normalized evaluation scores of various metrics is reported in Table VII. As correlation gives the degree of statistical relationship of metrics with human judgements, it can be inferred that the proposed MEE (Metric for Evaluation using Embeddings) outperforms baseline BLEU, NIST and METEOR in all cases. It is evident that BLEU and NIST has low correlation which proves that the n-gram matching syntactic evaluation lacks to capture semantic and morphological variations.

The graphical representation in Fig. 1 depicts the correlation values of normalized metric scores with normalized human judgements for Google and Bing MT Systems, per language. From the bar graphs, we can observe that the proposed approach holds high correlation with human judgements in case of both MT systems. Hence, having a higher correlation proves that MEE is more likely to evaluate the translation quality similar to that of humans than other metrics at assigning high(er) scores to better output, and low(er) scores to poorer output.

Also, in Fig. 2 we can see the *average of normalized scores of various metrics on bins of sentences* (Y-Axis) corresponding to a *normalized human score* (X-Axis)[5]. It is apparent from the graphs that the proposed metric exhibits higher correlation with human judgements when compared to other metrics for all kinds of output, whereas NIST and BLEU seem to under-perform for sentences which humans rated high.

The MEE performance is not optimum in a scenario where a reference sentence is a paraphrase of the corresponding MT output. Also, it was witnessed that in case of few improper translations, our proposed approach awarded decent scores in contrast with the low scores given by humans. The reason being, presence of a few common words irrespective of context. This suggests that in a few cases, MEE overlooks contextual information. However, in such instances, the performance of MEE drops only slightly, proving to be more robust than BLEU, NIST and METEOR.

VI. Future Work

We plan to extend our approach over all possible diverse language families. We look forward to considering paraphrases using pre-trained BERT models [21]. As they are

[4] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

[5] E.g.: In Fig.2(f), 72 sentences hold a normalized human score of 0.9 and, average of normalized BLEU, NIST and METEOR individually for these 72 sentences are about 0.3-0.4 whereas, average of normalized MEE is 0.5.

Table V: Normalized Scores of Example Sentence per Language.

			Normalized Scores				
Example Sentence	Transliteration	English Equivalent	Human	BLEU	NIST	METEOR	MEE
Telugu Candidate : ఒక్కసారిగా శరీర బరువు 10 శాతం తగ్గడం	okkasārigā śarīra baruvu 10 śātaṁ taggaḍaṁ	A 10 percent reduction in body weight at once.	0.8	0.4	0.6	0.7	0.8
Telugu Reference : శరీర బరువు ఒకేసారి 10 శాతం తగ్గిపోవడం	śarīra baruvu okēsāri 10 śātaṁ taggipōvaḍaṁ	Sudden decrease in body weight by 10 percent.					
Marathi Candidate : डावीकडील अन्नाचे तुकडे त्याद्वारे साफ केले जातात	Ḍāvīkaḍīla annācē tukaḍē tyādvārē sāpha kēlē jātāta.	The pieces of food on the left are cleaned by it.	0.2	0.3	0.1	0.1	0.2
Marathi Reference : ह्यामुळे जेवणातील राहिलेले कण निघून जातात	Hyāmuḷē jēvaṇātīla rāhilēlē kaṇa nighūna jātāta.	The left-over pieces of food is cleaned by it .					
Bengali Candidate : ছত্রাক আমাদের পায়ের আঙ্গুলের পেরেকের নীচে স্যাঁতসেঁতে জায়গায় তার আবাস তৈরি করে	Chatrāka āmādēra pāẏēra āṅgulēra pērēkēra nīcē syāṁtasēṁtē jāẏagāẏa tāra ābāsa tairi karē.	The fungus makes its home in damp places under our toe nails (metal nail).	0.2	0.4	0.1	0.1	0.2
Bengali Reference : ফাঙ্গাস আমাদের শরীরের এমন এমন জায়গা যেমন আঙুলের নখের নিচে নিজেদের বাসা বাঁধে	Phāṅgāsa āmādēra śarīrēra ēmana ēmana jāẏagā yēmana āṅulēra nakhēra nicē nijēdēra bāsā bāṁdhē	Fungus creates its abode in such places in our body, like in damped places under the nail of our toe .					
Hindi Candidate : यदि गुर्दे में एक ऑक्सालेट पत्थर है, तो रोकथाम क्या फायदेमंद हो सकती है	yadi gurde mein ek oksaalet patthar hai, to rokathaam kya phaayademand ho sakatee hai.	If kidney has an oxalate stone, what prevention can be beneficial.	0.5	0.2	0.2	0.3	0.4
Hindi Reference. : गुर्दे में ऑक्जेलेट पत्थरी रही हो तो आगे किन–किन चीजों से परहेज बरतना फायदेमंद हो सकता है	gurde mein okjelet pattharee rahee ho to aage kin-kin cheejon se parahej baratana phaayademand ho sakata hai.	If there are oxalate stones in kidney, then what further things are beneficial to avoid.					

Table VI: Rating Scale for Human Evaluation of Machine Translated Outputs.

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

Table VII: Language-Wise Correlation Values of Various Metrics with Human Judgements for Individual Machine Translation Systems.

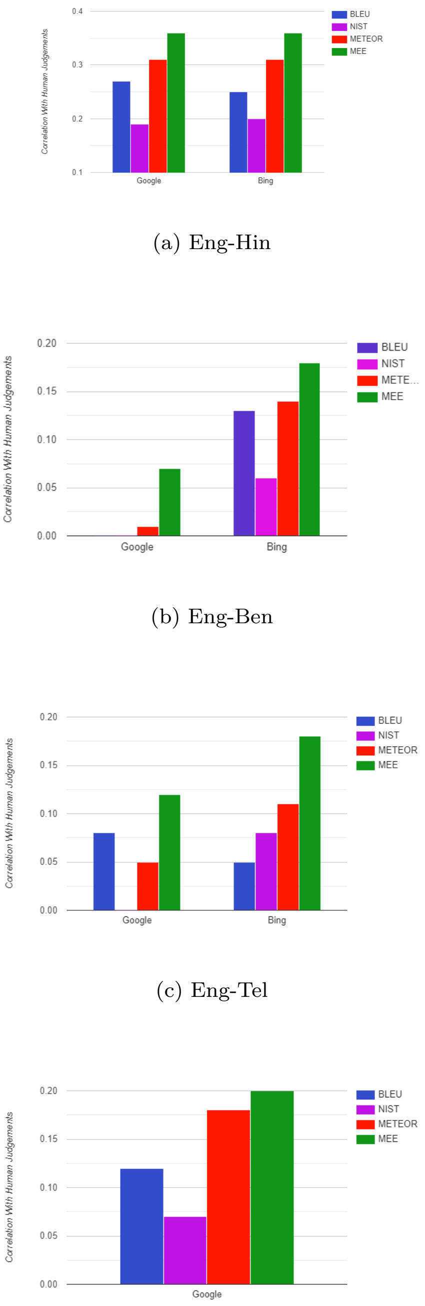
		Metrics			
Language	MT Sys.	BLEU	NIST	METEOR	MEE
Telugu	Google	0.08	0.00	0.05	**0.12**
	Bing	0.05	0.08	0.11	**0.18**
Marathi	Google	0.12	0.07	0.18	**0.20**
Bengali	Google	-0.01	-0.04	0.01	**0.07**
	Bing	0.13	0.06	0.14	**0.18**
Hindi	Google	0.27	0.19	0.31	**0.36**
	Bing	0.25	0.20	0.31	**0.36**

trained on contextual embeddings using attention on the entire sentence, it can help effectively to capture context, distant dependencies and ordering. We also intend to experiment with other existing datasets and correlate our metric with human judgements that are already available for such standard datasets. In exploration of MT evaluation, we believe, maximizing correlation with human judgements should be the primary focus.

VII. Conclusions

Automatic evaluation of machine translated output is a challenging task in *Natural Language Processing*. The proposed approach for automatic MT evaluation is based on the representation of text in a numerical form (embedding). We have tested and presented MEE (**M**etric for **E**valuation using **E**mbeddings) using two MT systems for four Indian Languages (Telugu, Marathi, Bengali and Hindi). The dataset considered in our experiments is robust and highly diverse in nature, consisting simple and complex sentences, with proper and improper translations.

Translation quality of these sentences were evaluated by humans and automatic metrics namely BLEU, NIST,



(a) Eng-Hin

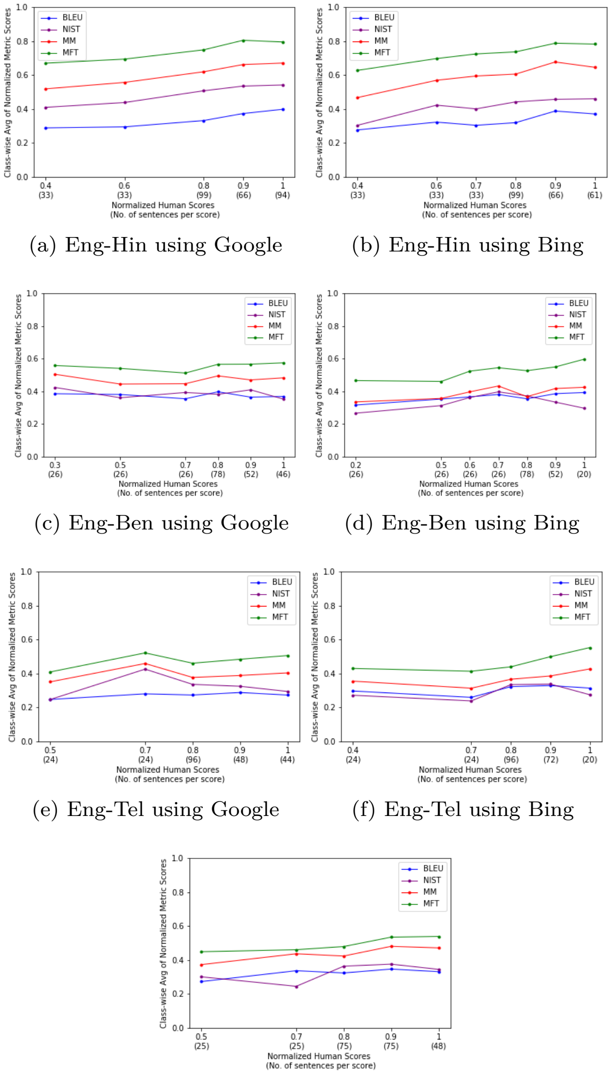
(b) Eng-Ben

(c) Eng-Tel

(d) Eng-Mar

Figure 1: Graphical Representation of Correlation of Various Automatic Metrics with Humans.

METEOR and MEE. Finally, Pearson product-moment correlation was used to compare the translation scores of various metrics with human judgements. Based on the reported correlation values, it is concluded that the proposed MEE significantly outperformed other widely used metrics. Besides, BLEU and NIST appeared to under-perform for sentences which humans rated high. Hence, an inference can be drawn that having only an *n-gram syntactic evaluation* fails to correlate well with humans for morphologically rich and free word order languages. Our approach captures **lexical, morphological and semantic similarity** without having a dependency on linguistic tools making it comparatively easier than existing metrics which demand several language-specific resources.



(a) Eng-Hin using Google

(b) Eng-Hin using Bing

(c) Eng-Ben using Google

(d) Eng-Ben using Bing

(e) Eng-Tel using Google

(f) Eng-Tel using Bing

(g) Eng-Mar using Google

Figure 2: Bin-wise (class-wise) Averaged Normalized Metric Scores Across Normalized Human Scores.
MM: METEOR, MFT: MEE (leveraging fastText embedding)

References

[1] Shterionov, Dimitar, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'dowd, and Andy Way. "Human versus automatic quality evaluation of NMT and PBSMT." Machine Translation 32, no. 3 (2018): 217-235.

[2] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. ”BLEU: a method for automatic evaluation of machine translation.” In Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311-318. Association for Computational Linguistics, 2002.

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

[4] Grave, Edouard, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. ”Learning word vectors for 157 languages.” arXiv preprint arXiv:1802.06893 (2018).

[5] Ananthakrishnan, R., Pushpak Bhattacharyya, M. Sasikumar, and Ritesh M. Shah. ”Some issues in automatic evaluation of english-hindi mt: more blues for bleu.” ICON (2007).

[6] Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. ”Re-evaluation the role of bleu in machine translation research.” In 11th Conference of the European Chapter of the Association for Computational Linguistics. 2006.

[7] Lommel, Arle. ”Blues for BLEU: Reconsidering the validity of reference-based MT evaluation.” In Proceedings of the LREC 2016 workshop “translation evaluation–from fragmented tools and data sets to an integrated ecosystem”, Portoroz, Slovenia, pp. 63-70. 2016.

[8] 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.

[9] Gupta, Ankush, Sriram Venkatapathy, and Rajeev Sangal. ”Meteor-Hindi: Automatic MT evaluation metric for Hindi as a target language.” In Proceedings of ICON-2010: 8th International Conference on Natural Language Processing. 2010.

[10] Lavie, Alon, and Abhaya Agarwal. ”METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments.” In Proceedings of the second workshop on statistical machine translation, pp. 228-231. 2007.

[11] Denkowski, Michael, and Alon Lavie. ”Meteor universal: Language specific translation evaluation for any target language.” In Proceedings of the ninth workshop on statistical machine translation, pp. 376-380. 2014.

[12] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. ”Efficient estimation of word representations in vector space.” arXiv preprint arXiv:1301.3781 (2013).

[13] Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. ”Fasttext. zip: Compressing text classification models.” arXiv preprint arXiv:1612.03651 (2016).

[14] Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. ”Enriching word vectors with subword information.” Transactions of the Association for Computational Linguistics 5 (2017): 135-146.

[15] Sasaki, Yutaka. ”The Truth of the F-Measure. 2007.” (2007).

[16] Lavie, A., Sagae, K. and Jayaraman, S., 2004, September. The significance of recall in automatic metrics for MT evaluation. In Conference of the Association for Machine Translation in the Americas (pp. 134-143). Springer, Berlin, Heidelberg.

[17] Jha, Girish Nath. ”The TDIL Program and the Indian Langauge Corpora Intitiative (ILCI).” In LREC. 2010.

[18] Panjwani, Ritesh, Diptesh Kanojia, and Pushpak Bhattacharyya. ”pyiwn: A Python-based API to access Indian Language WordNets.” In Proceedings of the 9th Global WordNet Conference (GWC 2018), p. 382. 2018.

[19] Ramanathan, Ananthakrishnan, Jayprasad Hegde, Ritesh Shah, Pushpak Bhattacharyya, and M. Sasikumar. ”Simple syntactic and morphological processing can help English-Hindi statistical machine translation.” In Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I. 2008.

[20] Benesty, Jacob, Jingdong Chen, Yiteng Huang, and Israel Cohen. ”Pearson correlation coefficient.” In Noise reduction in speech processing, pp. 1-4. Springer, Berlin, Heidelberg, 2009.

[21] Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. ”Bertscore: Evaluating text generation with bert.” arXiv preprint arXiv:1904.09675 (2019).