ISSN No. 0976-5697

Volume 3, No. 7, Nov-Dec 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Evaluating BLEU Metric for English to Hindi Language Using ManTra Machine Translation Engine

Neeraj Tomer*
AIM & ACT Banasthali University Banasthali
Jaipur, India
tneeraj12@rediffmail.com

Deepa Sinha
Department of Mathematics
South Asian University New Delhi, India
deepasinha2001@gmail.com

Piyush Kant Rai
AIM & ACT Banasthali University Banasthali
Jaipur, India
raipiyush5@gmail.com

Abstract: Evaluation of MT is required for Indian languages because the same MT is not works in Indian language as in European languages due to the language structure. So, there is a great need to develop appropriate evaluation metric for the Indian language MT. The present research work aims at studying the Evaluation of Machine Translation Evaluation's BLEU metric for English to Hindi for tourism domain using the output of ManTra, a translation system. Machine Translation Evaluation has been widely recognized by the Machine Translation community. The main objective of MT is to break the language barrier in a multilingual nation like India.

Keywords: MTE- Machine Translation Evaluation, MT – Machine Translation, EILMT –Evaluation of Indian Language Machine Translation, ManTra – MAchiNe Assisted TRAnslation Technology, Tr – Tourism.

I. INTRODUCTION

Indian languages are highly inflectional, with a rich morphology, relatively free word order, and default sentence structure as Subject-Object-Verb. In addition, there are many stylistic differences. So the evaluation of MT is required for Indian languages because the same MT does not work in Indian language as in European languages. The same tools are not used directly because of the language structure. So, there is a great need to develop appropriate evaluation metric for the Indian language MT.

English is understood by less than 3% of Indian population. Hindi, which is official language of the country, is used by more than 400 million people **[1]**. MT assumes a much greater significance in breaking the language barrier within the country's sociological structure. The main objective of MT is to break the language barrier in a multilingual nation like India. English is a highly positional language with rudimentary morphology, and default sentence structure as Subject-Verb-Object. The present research work aims at studying the "Evaluation of Machine Translation Evaluation's BLEU Metric for English to Hindi" for tourism domain. The present research work is the study of statistical evaluation of machine translation evaluation for English to Hindi. The research aims to study the correlation between automatic and human assessment of MT quality for English to Hindi. The main goal of our experiment is to determine how well a variety of automatic evaluation metric correlated with human judgment.

In the present work we propose to work with corpora in the tourism domain and limit the study to English – Hindi language pair. It may be assumed that the inferences drawn from the results will be largely applicable to translation for English to other Indian Languages. Our test data consisted of a set of English sentences that have been translated from expert and non-expert translators. The English source sentences were randomly selected from the corpus of tourism domain. These sentences are taken randomly from the different resources like websites, pamphlets etc. Each output sentence was scored by Hindi speaking human evaluators who were also familiar with English. It may be assumed that the inferences drawn from the results will be largely applicable to translation for English to other Indian Languages, as assumption which will have to be tested for validity. We intend to consider the following MT engine in our study-

a. ***ManTra:*** C-DAC Pune has developed a translation system called ManTra. The work in ManTra has to be viewed in its potentiality of translating the bulk of texts produced in daily official activities. The system is facilitated with pre-processing and post-processing tools, which enables the user to overcome the problems/errors with minimum effort. The strategy used for translation is: NOT Word to Word; NOR Rule to Rule; BUT Lexical Tree to Lexical Tree **[2]**.

II. OBJECTIVE

The main goal of this work is to determine how well a variety of automatic evaluation metrics correlated with human scores. The other specific objectives of the present work are as follows.

a) To design and develop the parallel corpora for deployment in automatic evaluation of English to Hindi machine translation systems.

b) Assessing how good the existing automatic evaluation metrics BLEU, will be as MT evaluating strategy for evaluation of Indian language machine translation systems by comparing the results obtained by this with human evaluator's scores by correlation study.
c) To study the statistical significance of the evaluation results as above, in particular the effect of-
   (a). size of corpus
   (b). sample size variations
   (c). increase in number of reference translations

a. Creation of parallel corpora: Corpus quality plays a significant role in automatic evaluation. Automatic metrics can be expected to correlate very highly with human judgments only if the reference texts used are of high quality, or rather, can be expected to be judged high quality by the human evaluators. The procedure for creation of parallel corpora is as under:

a) Collect English corpus from the domain from various resources.
b) Generate multiple references (we limit it to three) for each sentence by getting the source sentence translated by different expert translators.
c) XMLise the source and translated references for use in Automatic evaluation

Description of Corpus

Domain	Source Language	Target Language	No. of Sentences	No. of Human Translation	Name of MT Engine
Tourism	English	Hindi	1000	3	Mantra

For the corpus collection our first motive was to collect as possible to get better translation quality and a wide range vocabulary. For this purpose the first corpus we selected to use in our study is collected from different sources. We have manually aligned the sentence pairs.

In our study for tourism domain we take 1000 sentences. When the text has been collected, we distributed this collected text in the form of Word File, each word files having the 100 sentences of the particular domain. In this work our calculation will be based on four files- source file and three reference files. Reference files are translated by the language experts. We give the file a different identification. For e.g. our first file name is Tr_0001_En where Tr_ for tourism 0001 means this is the first file and En means this is the Candidate file. We treat this as the candidate file. In the same way our identification for the Hindi File is Tr_0001_Hi, in this Hi is for the Hindi file and we have called this a reference file. As we already mention that we are taking the three references we named them reference 1(R1), reference 2(R2), reference 3(R3). In the study we take the candidate sentence and the reference sentences, as shown below in e.g.:

b. Source Sentence:

In Mexico City is the Plaza de las Tres Culturas, which celebrates the three major cultures that have shaped Mexico: there are Aztec ruins, the 17th-century colonial church of San Diego and several late 20th-century buildings **[3]**.

c. Candidate Sentence:

मेक्सिको नगर नगर चौक ड लास ट्रेस कुल्टुरस में, जो तीन प्रमुख संस्कृतियों मनाते है कि मेक्सिको: वहाँ को आकार धारण किया गए सन डीगो और कई देर 20थ शताब्दी भवनों का अज़टेक खंडहर, 17थ शताब्दी औपनिवेशिक गिरजाघर हैं

d. Reference Sentences:

R1: मैक् सिको शहर में tres culturas होता है, जो कि खुद उपप्रधानमंत्री hotel plaza de जश्न मनाया मैक् सिको के तीन प्रमुख संस्कृतियों आकार है कि अवशेषों aztec हैं, जो 17वीं शताब् दी के सैन डिएगो और औपनिवेशिक चर्च भवन कई देर से 20वीं शताब्दी

R2: मैक् सिको शहर में tres culturas होता है, जो कि खुद उपप्रधानमंत्री hotel plaza de जश्न मनाया मैक् सिको के तीन प्रमुख संस्कृतियों आकार है कि अवशेषों aztec हैं, जो 17वीं शताब् दी के सैन डिएगो और औपनिवेशिक चर्च भवन कई देर से 20वीं शताब्दी

R3: मैक्सिको शहर बहुत सी संस्कृतियों का संग्रह हैं, जिनमें से मुख्यतया तीन ने मैक्सिकों को आकार प्रदान किया हैः यहां है एजटेक सर्वनाश, 17वीं सदी की सैन डियेगो की कोलोनियल चर्च व पुरानी 20वीं सदी के कुछ मकान।

III. HUMAN EVALUATION

Human evaluation is always best choice for the evaluation of MT but it is impractical in many cases, since it might take weeks or even months (though the results are required within days). It is also costly, due to the necessity of having a well trained personnel who is fluent in both the languages, source and targeted. While using human evaluation one should take care for maintaining objectivity. Due to these problems, interest in automatic evaluation has grown in recent years **[4]**. Every sentence was assigned a grade in accordance with the four point scale for adequacy.

IV. AUTOMATIC EVALUATION BY BLEU METRIC

We used BLEU evaluation metric for this study. This metric is specially designed for English to Hindi. BLEU metric, designed for evaluating MT quality, scores candidate sentences by counting the number of n-gram matches between candidate and reference sentences. BLEU metric is probably known as the best known automatic evaluation for MT. To check how close a candidate translation is to a reference translation, an n-gram comparison is done between both. Metric is designed from matching of candidate translation and reference translations. We have chosen correlation analysis to evaluate the similarity between automatic MT evaluations and human evaluation. Next, we obtain scores of evaluation of every translated sentence from both MT engines. The outputs

from both MT systems were scored by human judges. We used this human scoring as the benchmark to judge the automatic evaluations. The same MT output was then evaluated using both the automatic scoring systems. The automatically scored segments were analyzed for Spearman's Rank Correlation with the ranking defined by the categorical scores assigned by the human judges. Increase in correlation indicates that the automatic systems are more similar to a human in ranking the MT output.

Statistical significance is an estimate of the degree, to which the true translation quality lays within a confidence interval around the measurement on the test sets. A commonly used level of reliability of the result is 95% **[5]**. To reach at decision, we have to set up a hypothesis and compute p-value to get final conclusion.

The present research is the study of statistical evaluation of machine translation evaluation's BLEU metric. The research aims to study the correlation between automatic and human assessment of MT quality for English to Hindi. While most studies report the correlation between human evaluation and automatic evaluation at corpus level, our study examines their correlation at sentence level. The focus in this work is to examine the correlation between human evaluation and automatic evaluation and its significance value, not to discuss the translation quality. In short we can say that this research is the study of statistical significance of the evaluated results, in particular the effect of sample size variations.

So, firstly we take source sentences and then get these sentences translated by our MT engine, here we consider the Anuvadaksh. We have the different references of these sentences. After doing this we do the evaluations of these sentences human as well as the automatic evaluations and we collect the individual scores of the given sentences considering all the three references one by one. The following table shows the individual scores of the five sentences (particular sentences can be seen at the end of the paper) using different no. of references **[6]**.

Table 1: Human Evaluation and BLEU Evaluation scores

S. No.	BLEU Score			
	Human Eval.	one no. of reference	two no. of references	three no. of references
1.	0.75	0.3101	0.3169	0.3917
2.	0.75	0.3398	0.35	0.3837
3.	0.75	0.6965	0.7364	0.8011
4.	0.75	0.3289	0.3289	0.4498
5.	0.5	0.7071	0.7071	0.8133

In this way we also collect the individual scores of all the sample sizes like 20, 60,100,200,300,500 and 1000 sentences. After this we do the correlation analysis of these values. In order to calculate the correlation with human judgements during evaluation, we use all English–Hindi human rankings distributed during this shared evaluation task for estimating the correlation of automatic metrics to human judgements of translation quality, were used for our experiments. In our study the rank is provided at the sentence level **[7]**.

For correlation analysis we calculate the correlation between human evaluation and automatic evaluations one by one by the Spearman's Rank Correlation method. The Spearman's rank correlation coefficient is given as (when ranks are not repeated)-

$$\rho = 1 - \left(\frac{6\sum_{i=1}^{n} d^2}{n(n^2-1)} \right)$$

Where d is the difference between corresponding values in rankings and n is the length of the rankings. An automatic evaluation metric with a higher correlation value is considered to make predictions that are more similar to the human judgements than a metric with a lower value. Firstly, we calculate the correlation value in between the human evaluation and automatic evaluation BLEU metric means human evaluation with BLEU for sample size 20, 60, 100, 200, 300, 500 and 1000.

Table 2: Correlation (ρ) values

Sample Size	ρ values		
	one no. of reference	two no. of references	three no. of references
20	.194	.230	.200
60	.260	.239	.279
100	.157	.159	.166
200	-.066	-.091	-.066
300	-.066	-.121	-.135
500	-.091	-.123	-.116
1000	-.108	-.119	-.106

After calculating the correlation, we need to find out which type of correlation is there between the variables and of which degree and whether the values of the correlation are significant.

V. ANALYSIS OF STATISTICAL SIGNIFICANCE TEST FOR HUMAN EVALUATION AND AUTOMATIC EVALUATION

Statistical significance is an estimate of the degree, to which the true translation quality lays within a confidence interval around the measurement on the test sets. A commonly used level of reliability of the result is 95%, for e.g. if, say, 100 sentence translations are evaluated, and 30 are found correct, what can we say about the true translation quality of the system? To reach at decision, we have to set up a hypothesis and compute p-value to get final conclusion that whether there is any correlation between the human evaluations and automatic evaluations. If yes, then what is the type and degree of correlation? Also what is the significance of the correlation value? In this work we set the hypothesis that there is no correlation between the values of human and automatic evaluation. The p-value will provide the answer about the significance of the correlation value.

A Z-test is a statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. For each significance level, the Z-test has a single critical value (for example, 1.96 for 5% two tailed) which makes it more convenient than the Student's t-

test which has separate critical values for each sample size **[8]**. The test statistic is calculated as:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, s_1^2 and s_2^2 are the sample variances, n_1 and n_2 are the sample sizes and z is a quartile from the standard normal distribution.

Table 3: p-values of output of Anuvadaksh using different no. of references

Sample Size	p-values		
	one no. of reference	two no. of references	three no. of references
20	0.0001	0.0001	0.0001
60	0.0001	0.0001	0.0001
100	0.0001	0.0001	0.0001
200	0.015	0.0548	0.015
300	0.017	0.0364	0.0301
500	0.0174	0.0202	0.0158
1000	0.0158	0.0179	0.015

Now on the basis of these values we conclude our results like which type and degree of correlation is there between the given variables and whether the correlation results are significant. In the above example we have done all the calculations by considering the single reference sentence and in tourism domain using 5 numbers of sentences **[9]**.

But in our research work we consider the different references like 1, 2, 3 and we use the different sample sizes like 20, 60, 100, 200, 300, 500, and 1000. We see whether the results remains uniform for different sample sizes and different number of references in particular domains **[10, 11]**. For above calculation we used following sentences:

a. This section includes basic facts on a number of the US External Territories: Baker & Howland Islands, Jarvis Island, Johnston Atoll, Kingman Reef, Midway Islands, Navassa Island, Palmyra and Wake Island.
b. But traces of earlier inhabitants remain in the remarkable temples and pyramids at Palenque and Teotihuacán, and in the traditions of dozens of indigenous cultures.
c. Though expensive and exclusive, heliskiing ultimate adventure sport is fast gaining popularity.
d. Camel safaris generally cover the area around Jaisalmer, Bikaner or Jodhpur, popularly known as the desert circuit.
e. History comes to life in Mexico: the scars of recent history is still apparent.

Candidate Sentences (translated by ManTra):

a. यह अनुभाग हम बाह्य क्षेत्रों: बेकर और होवलांड द्वीपों, जर्विस द्वीप, जोहंस्टोन प्रवाली, किंग्मान समुद्री चट्टान, मध्यम मार्ग द्वीपों, नवस्स द्वीप, पल्म्यरा और जागना द्वीप के संख्या पर बुनियादी तथ्य को शामिल करता है
b. किंतु का पहले निवासियों उल्लेखनीय मंदिरों और पिरैमिडो में रहते हैं में, और देशी संस्कृतियों के दर्जनों के परंपराओं में खोल पर चलते हैं
c. महंगी और अनन्य, हेलिस्कींग चरम साहस प्रदर्शित यद्यपि लोकप्रियता को तेज प्राप्त कर रहे हैं
d. ऊँट यात्राएँ अधिकांश सम्मिलित क्षेत्र जैसलमेर, बीकानेर अथवा जोधपुर के बारे में, योग्यता दौरा सामान्यतया जानते हैं
e. इतिहास तक लिफ में मेक्सिको: हाल ही का इतिहास का निशाने प्रकट अभी भी हैं

VI. RESULTS

In the domain tourism there is significance difference between the average evaluation score of human with BLEU at 5% level of significance and this is for sample sizes 20, 60 and 100.

In Table 2 (Correlation (ρ) values) correlation value for BLEU is .230 and .200 these values are for sample size 20 and for two and three number of references which is significant at 5% level of significance. A similar result is seen in the case of sample size 60 and 100 for all three references. But for the sample sizes 200, 300, 500 and 1000 value of correlation is insignificant on the given level of significance.

VII. CONCLUSION

This work will help to give the feedback of the MT engines. In this way we may make the changes in the MT engines and further we may revise the study. Corpus quality plays a significant role in automatic evaluation. Automatic metrics can be expected to correlate highly with human judgments only if the reference texts used are of high quality.

VIII. ACKNOWLEDGMENT

The present research work was carried under the research project “English to Indian Languages Machine Translation System (EILMT)”, sponsored by TDIL, Ministry of Communications and Information Technology, Government of India. With stupendous ecstasy and profundity of complacency, we pronounce utmost of gratitude to Late Prof. Rekha Govil, Vice Chancellor, Jyoti Vidyapith, Jaipur Rajasthan.

IX. REFERENCES

[1]. en.wikipedia.org/wiki/English_language

[2]. http://pune.cdac.in/html/aai/mantra_rajbhasha_en.aspx

[3]. http://www.destinationservices.com/destinations/americas/mexico

[4]. Bilbao V. (2005): “Evaluation of Machine Translation Systems and of Parallel Text Alignment”, Centro de Informatica e Technologies da Informacao.

[5]. Koehn P. (2004): "Statistical Significance Tests for Machine Translation Evaluation" Computer Science and Artificial Intelligence Laboratory Massachusetts Institute of Technology, The Stata Center, 32 Vassar Street, Cambridge, MA 02139.

[6]. Tomer N. and Sinha D. (2012): "Evaluating Machine Translation Evaluation's BLEU Metric for English to Hindi Language Machine Translation", The International Journal of Computer Science & Application-TIJCSA, 1(6), 48-58.

[7]. Tomer N., Sinha D. and Rai P.K. (2012): "Evaluating Machine Translation Evaluation's F-Measure Metric for English to Hindi Language Machine Translation", International Journal of Academy Research Computer Engineering and Technology-IJARCET, 1(7), 151-156.

[8]. http://en.wikipedia.org/wiki/Z-test

[9]. Tomer N., Sinha D. and Rai P.K. (2012): "Evaluation of Modified-BLEU Metric for English to Hindi Language Using ManTra Machine Translation Engine", International Journal of Advanced Research in Electronics & Communication Engineering -*IJARECE*, 1(4), 103-108.

[10]. Tomer N. and Sinha D. (2012): "Evaluating NIST Metric for English to Hindi Language Using ManTra Machine Translation Engine", International Journal of Academy Research Computer Engineering and Technology-IJARCET, 1(8), 365-369.

[11]. Tomer N. and Sinha D. (2012): "Evaluating Machine Translation Evaluation's NIST Metric for English to Hindi Language Machine Translation", The International Journal of Multidisciplinary Academy IJMRA 2(11), 359-371.