EVALUATING MACHINE TRANSLATION EVALUATION'S NIST METRIC FOR ENGLISH TO HINDI LANGUAGE MACHINE TRANSLATION

NEERAJ TOMER*

DEEPA SINHA*

Abstract

The present research work aims at studying the Evaluation of Machine Translation Evaluation's NIST Metric for English to Hindi for tourism domain. The main objective of MT is to break the language barrier in a multilingual nation like India. 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.

Keywords: MTE- Machine Translation Evaluation
MT- Machine Translation,
EILMT – Evaluation of Indian Language Machine Translation.

* Centre for Apaji Institute of Mathematics & Applied Computer Technology (AIM & ACT) Banasthali University BanasthaliNiwai Jaipur Rajasthan.

Introduction:

The present research work aims at studying the "Evaluation of Machine Translation Evaluation's NIST 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.

India is a highly multilingual country with 22 constitutionally recognized languages. Even though, 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. Therefore, 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. 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 is not works 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.

Materials and Methods: 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 samples are taken randomly from the different resources like websites, pamphlets etc. Each output sentence was score 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 be consider the following MT engine in our study-

- Anuvadaksh

Objective: The main goal of this work is to determine how well a variety of automatic evaluation metrics correlated with human judges. A secondary goal is to determine for which the correlation of automatic and human evaluation is particularly good or bad. The other specific objectives of the present work are as follows.

1. To design and develop the parallel corpora for deployment in automatic evaluation of English to Hindi machine translation systems.
2. Assessing how good the existing automatic evaluation metric NIST, will be as MT evaluating strategy for evaluation of EILMT systems by comparing the results obtained by this with human evaluator's scores by correlation study.
3. To study the statistical significance of the evaluation results as above, in particular the effect of-
    - size of corpus
    - sample size variations
    - increase in number of reference translations

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:

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

Description of Corpus:

Domain	Tourism
Source Language	English
Target Language	Hindi
No. of Sentences	1000
No. of Words	23000
No. of Human Translation	3
No. of MT Engine	1

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 raw corpus we selected to use in our study is collected from different resources like websites, pamphlets etc. Then 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. For e.g.

Source Sentence: Antarctica is welcoming more tourist-orientated cruises and ferries to the region every year, and facilities are continually developing, with more accommodation, culinary and travel options available.

Candidate Sentence: *अन्टार्कटिका अधिक पर्यटक उन्मुख समुद्र विहार नौकाएँ को स्वागत कर रहा है और जिला प्रत्येक वर्ष और सुविधाएँ को घाटों अधिक आवास, खानपान और पर्यटन विकल्पों उपलब्ध के साथ निरन्तर विकास कर रहे हैं।*

Reference Sentences:

R1- *प्रत्येक वर्ष अण्टार्कटिका अधिकाधिक पर्यटकोन्मुख यात्रियों तथा समुद्री जहाजों का अपने क्षेत्र में स्वागत करता है और भोजन व्यवस्था तथा वैकल्पिक यात्रा की अधिक सुविधा के साथ यह लगातार विकास कर रहा है।*

R2- *अंटार्कटिका पर्यटकों से चमकती हुई समुद्री यात्राओं तथा घाट को प्रत्येक बर्फ प्रदेश में स्वागत करता है तथा लगातार सुविधाऐं विकसित हो रही हैं, साथ ही ज्यादा सुख सुविधा, पाकशाला संबंधी तथा यात्रा के विकल्प उपलब्ध हैं।*

R3- *प्रत्येक वर्ष अन्टार्कटिका, अपने प्रांत में और अधिक पर्यटक संबंधित समुद्री जहाजों और यात्राओं का स्वागत कर रहा है, तथा ठहरने, खाने और यात्रा के अधिक उपलब्ध विकल्पों के साथ सुविधाएं निरंतर विकसित हो रहीं हैं।*

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. Every sentence was assigned a grade in accordance with the following four point scale for adequacy.

	Score
• Ideal	1
• Acceptable	.5
• Not Acceptable	.25
• If a criterion does not apply to the translation	0

Automatic Evaluation by NIST: We used NIST evaluation metric for this study. This metric is specially designed for English to Hindi. NIST metric, designed for evaluating MT quality, scores candidate sentences by counting the number of n-gram matches between candidate and reference sentences. NIST 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 to 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 by which 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. Increases in correlation indicate that the automatic systems are more similar to a human in ranking the MT output.

The present research is the study of statistical evaluation of machine translation evaluation's NIST 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.

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%. To reach at decision, we have to set up a hypothesis and compute p-value to get final conclusion.

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 using different no. of references. These sentences are translated by Anuvadaksh or we may say that these are the output of Anuvadaksh machine translation engine.

S. No.	NIST Scores			
	Human Eval.	one no. of reference	two no. of references	three no. of references
1	1	.1761	.7738	.7738
2	.75	.2708	.3089	.3089
3	.35	0	0	0
4	1	.249	.4509	.4509
5	1	.1751	.1548	.1548

Table: Human Evaluation and NIST Metric scores

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.

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 NIST metric means human evaluation with NIST for sample size 20, 60,100,200,300,500 and 1000.

Sample Size	ρ values		
	one no. of reference	two no. of references	three no. of references
20	.066	.101	.072
60	.003	.058	.058
100	.015	.063	.063
200	.148	.101	.088
300	.126	.079	.079
500	.173	.154	.154
1000	.163	.154	.163

Table: Correlation (ρ) values

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.

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. 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 quantile from the standard normal distribution.

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	0.0001	0.0001	0.0001
300	0.0001	0.1446	0.1762
500	0.1814	0.1402	0.1492

1000	0.166	0.1492	0.166

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

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.

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. Results are as follows;

Results: In the domain tourism there is significance difference between the average evaluation score of human with NIST at 5% level of significance and for the sample size.

We see that as we increase the number of references there is improvement in our results. In Table the correlation value for NIST is .148 and .126 these values are for sample size 200 and 300 and for one no. of reference which is significant at 5% level of significance. A similar result is seen in the case of sample size 20, 60, 100 and 200 for two and three no. of references. From the analysis on the basis of z-test used for the significance test of human evaluation and automatic evaluation we obtain the following important point; in the domain tourism there is significance difference between the average evaluation score of human with NIST at 5% level of significance and for the some sample sizes.

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, or rather, can be expected to be judged of high quality by human evaluators.

References:

1. T. Neeraj (2012): "Evaluating Machine Translation (MT) Evaluation Metrics for English to Indian Language Machine Translation", Ph.D. Thesis, Banasthali University, Banasthali.
2. Neeraj Tomer and Deepa Sinha (2012): "Evaluating Machine Translation Evaluation's BLEU Metric for English to Hindi Language Machine Translation", in The International Journal of Computer Science & Application, Vol-01-NO-06-Aug-12, 48-58.
3. Rao, Durgesh (2001): "Machine Translation in India: A Brief Survey", National Centre for Software Technology Gulmohar Road 9, Juhu, Mumbai 400049, India, 21-23.
4. http://en.wikipedia.org/wiki/History_of_machine_translation
5. http://en.wikipedia.org/wiki/Evaluation_of_machine_translation
6. Andrew FINCH, Eiichiro SUMITA, Yasuhiro AKIBA (2004): "How Does Automatic Machine Translation Evaluation Correlate With Human Scoring as the Number of Reference Translations Increases?" ATR Spoken Language Translation Research Laboratories, 2-2-2 Hikaridai "Keihanna Science City" Kyoto, 619-0288, Japan, 2019-2022.
7. Alex Kulesza, Stuart M Shieber (2004): "A Learning Approach to Improving Sentence-Level MT Evaluation", Division of Engineering and Applied Sciences Harvard University 33 Oxford St. Cambridge, MA 02138, USA, 75-84.
8. Philipp Koehn (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.
9. George Doddington (2002): "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics", In Proceedings of the Second Conference on Human Language Technology (HLT-2002). San Diego, CA. 128-132.
10. Paula Estrella, Andrei Popescu-Belis , Maghi King (2007): "A New Method for the Study of Correlations between MT Evaluation Metrics", ISSCO/TIM/ETI University of Geneva 40, bd. du Pont-d'Arve 1211 Geneva, Switzerland, 35-43.
11. Coughlin, D. (2003) "Correlating Automated and Human Assessments of Machine Translation Quality" in MT Summit IX, New Orleans, USA, 23–27.
12. Sanjay Kumar Dwivedi, Pramod Premdas Sukhadeve (2010): "Machine Translation System in Indian Perspectives" Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India, J. Computer Sci., 6 (10): 1111-1116.

13. Deborah Coughlin, (2003): "Correlating Automated and Human Assessments of Machine Translation Quality", In Proceedings of MT Summit IX. New Orleans, 63-70.
14. Donaway, R.L., Drummey, K.W., and Mather, L.A., (2000): "A Comparison of Rankings Produced by Summarization Evaluation Measures", In Proceedings of the Workshop on Automatic Summarization, 69-78.
15. Akiba, Yasuhir Taro Watanabe, Eiichiro Sumita, (2002): "Using Language and Translation Models to Select the Best among Outputs from Multiple {MT} System", Proceeding of Colong, 8-14.
16. Andrei Popescu-Belis (2002): "An experiment in comparative evaluation: humans vs. computers", ISSCO/TIM/ETI, University of Geneva, 55-64.
17. Jesus´ Angel Gimenez Linare (2008): "Empirical Machine Translation and its Evaluation", Artificial Department the Languages Systems Informatics University, 27-38.
18. Vannatta Rachel: "Statistics in Education Course Packet", EDFI 641. Online available http://personal.bgsu.edu/~ rvanna/packetspring 09.pdf.
19. Feifan Liu, Yang Liu (2008): "Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries", the University of Texas at Dallas Richardson, TX 75080, USA, 201-208.
20. Philipp Koehn (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.
21. Sanjay Kumar Dwivedi, Pramod Premdas Sukhadeve (2010): "Machine Translation System in Indian Perspectives", Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India.
22. Bohan Niamh, Breidt, Volk Martin, (2000): "Evaluating Translation Quality as Input to Product Development", 2nd International Conference on Language Resources and Evaluation, Athens.

For above all calculation we used following sentences and all these are taken from following websites:

http://www.fernhoteljaipur.com/rajasthan-train-travel.htm

http://www.spectrumtour.com/rajasthan-tourism/jaipur-travel.htm

http://www.travel-in-rajasthan.com/rajasthan-travel/sisodia-rani-bagh.htm

http://www.kkroyalhotel.com/location_travelinfo.php

http://www.mapsofindia.com/jaipur/gardens/kanak-vrindavan-gardens.html

English Sentences:

1. Government Central Museum was constructed in 1876 when Prince of Wales has visited India and opened to public in 1886.
2. Government Central Museum has a rich collection of ivory work, textiles, jewellery, carved wooden objects, miniature paintings, marble statues, arms and weapons.
3. Sisodiya Rani-Ka-Bagh was built by Sawai Jai Singh II for his Sisodiya Queen.
4. The Jal Mahal is a picturesque palace built for royal duck shooting parties.
5. Kanak Vrindavan is a popular picnic spot in Jaipur.

Candidate Sentences (output of Anuvadaksh):

1. सरकारी केंद्रीय संग्रहालय को 1876 में निर्मित किया गया था जब वेल्स का राजकुमारों भारत भ्रमण किया है और 1886 में जनता को आरंभ किये
2. सरकारी केंद्रीय संग्रहालय के पास हाथीदांत कार्य , वस्त्रों , आभूषण , नक्काशीदार लकड़ी की वस्तुएँ का समृद्ध संग्रह , सूक्ष्म चित्र चित्रों , संगमरमर मूर्तियाँ , अस्त्रों और शस्त्रों है
3. सिसोदिया रानी का बाग उसके सिसोदिया रानी के लिए सवाई जय सिंह II द्वारा निर्माण किया गया था
4. जल महल शाही बत्तख आखेट दलों के लिए निर्माण किया गया चित्रात्मक महल है
5. कनक वृंदावन जयपुर में प्रसिद्ध विहार स्थल है

Reference 1:

1. गवर्नमेण्ट सेन्ट्रल म्यूजियम 1876 में, जब प्रिंस ऑफ वेल्स ने भारत भ्रमण किया,बनवाया गया था और 1886 में जनता के लिए खोला गया ।

2. गवर्नमेण्ट सेन्ट्रल म्यूजियम में हाथीदांत कृतियों, वस्त्रों, आभूषणों, नक्काशीदार काष्ठ कृतियों, सूक्ष्म चित्रों संगमरमर प्रतिमाओं, शस्त्रों और हथियारों का समृद्ध संग्रह है।
3. सवाई जय सिंहII ने अपनी सिसोदिया रानी के लिए सिसोदिया रानी का बाग बनवाया।
4. जलमहल शाही बतख शिकार गोष्ठियों के लिए बनाया गया एक सुंदर महल है।
5. कनक वृंदावन जयपुर में एक लोकप्रिय विहार स्थल है।

Reference 2:

1. राजकीय केन्द्रीय संग्रहालय का निर्माण सन् 1876 में जब प्रिंस ऑफ वेल्स भारत भ्रमण के लिए आये, तब करवाया गया तथा आम लोंगों के लिए यह 1886 में खोला गया।
2. राजकिय केन्द्रीय संग्रहालय में हाथी दांत से निर्मित वस्तुओं, गहनों, लकड़ी की नक्काशीदार चीजों, छोटी-छोटी तस्वीरों, संगमरमर की मूर्तियो तथा अस्त्र-शस्त्र का नायाब संग्रह है।
3. सवाई जयसिंह द्वितीय ने सिसोदिया रानी के लिए सिसोदिया रानी का बाग का निर्माण करवाया था।
4. जलमहल एक जीवन्त महल है जिसका निर्माण शाही लोगों के लिए पक्षी आखेट के लिए करवाया गया।
5. कनक वृन्दावन जयपुर का एक बहुत ही मशहुर भ्रमणीय स्थल है।

Reference 3:

1. केन्द्रीय सरकारी संग्रहालय 1876 में बनवाया गया था जब वेल्स के राजकुमार भारत के दर्शन को आये थे; और 1886 में जनता के लिए खुला था।
2. केन्द्रीय सरकारी संग्रहालय में, हाथीदांत के कार्य का, वस्त्रों, गहनों, नक्काशीदार लकड़ी की वस्तुओं, सूक्ष्म चित्रकारी, संगमरमर की मूर्तियों, अस्त्रों और शस्त्रो का अच्छा संग्रह है।
3. सवाई जयसिंह द्वितीय ने अपनी सिसोदिया रानी के लिए सिसोदिया रानी का बाग बनवाया था।
4. शाही बतख मारने वाले दलों के लिए निर्मित जल महल एक देखने योग्य महल है।
5. कनक वृंदावन जयपुर में एक लोकप्रिय पिकनिक स्थल है