ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 8, October 2012

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

Neeraj Tomer[1]
AIM & ACT
Banasthali University Banasthali
Jaipur, India

Deepa Sinha[2]
Department of Mathematics
South Asian University
New Delhi, India

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 NIST 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

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 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.

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. 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 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.

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 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-

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.

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.

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 metrics NIST, 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.
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	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. For e.g.

Source Sentence: Guru Shikhar is the highest peak on the Mount Abu which provides an excellent view of the whole town.

Candidate Sentence: पहाड़ी अबु जो प्रदान सर्वोत्तम दृश्य का सभी नगर पर गुरु शिखहर उच्चतम शिखर

Reference Sentences:

R1: गुरूशिखर माऊण्ट आबू की सबसे ऊँची चोटी है जहां से पूरे शहर का उत्तम दृश्य देखने को मिलता है।

R2: गुरू शिखर माउण्ट आबू की सबसे ऊंची चोटी है जो पूरे शहर का अत्यंत उत्कृष्ट दृश्य प्रदान करती है।

R3: गुरू शिखर माउण्ट आबू पर सबसे ऊँचा शिखर है जो पूरे शहर का उत्कृष्ट दृश्य प्रदान करता है।

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 four point scale for adequacy.

AUTOMATIC EVALUATION BY NIST METRIC

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 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. Increases in correlation indicate 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%. 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 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.

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.

Table 1: Human Evaluation and NIST Evaluation scores

S. No.	NIST Score			
	Human Eval.	one no. of reference	two no. of references	three no. of references
1.	0.75	0	0.0792	0.0792
2.	0.5	0.2221	0.2511	0.2511
3.	0.75	0.0695	0.1508	0.1508
4.	0.75	0.1464	0.1797	0.1797
5.	0.75	0	0.1394	0.1394

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.

Table 2: Correlation (ρ) values

Sample Size	ρ values		
	one no. of reference	two no. of references	three no. of references
20	.110	.360	.360
60	-.075	.017	.017
100	.060	.048	.048
200	.407	.287	.407
300	.376	.274	.274
500	.304	.245	.245
1000	.274	.245	.274

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 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	0.1977	0.1814	0.1977
300	0.2358	0.2033	0.2033
500	0.2643	0.2266	0.2266
1000	0.2451	0.2266	0.2451

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.

For above calculation we used following sentences:
English Sentences:

1. The best way to experience the real magic of Thar Desert, Rajasthan is with the help of Desert Safari.
2. Camel safari could be the best choice for the adventure-seeking tourists moving around the interiors of Thar Desert, witnessing the cities and historical ruins of the city.
3. Sam Sand Dunes is a perfect tourist destination known for the sunset and sunrise point.
4. Manvar Desert is best for the outdoor adventures and excitement.
5. Bikaner is located to the north of Rajasthan popularly known as a camel country.

Candidate Sentences (translated by ManTra):

1. महसूस करने के लिए योग्यता यात्रा का थार मरूस्थल , रजस्थन के साथ सहायता का सर्वोत्तम तरीका वास्तविक जादू
2. ऊँट यात्रा सबसे अच्छा पसंद साहस प्राप्त पर्यटकों के लिए थार मरूस्थल का आंतरिक करीब , सुधार कर रही हैं नगर का शहरों और ऐतिहासिक खंडहर देखरहे हैं
3. सम बालू बालू के टीले पूरा पर्यटक मंजिल प्रसिद्ध के लिए सूर्यास्त और सूर्योदय इशारा करते हैं
4. मांवर योग्यता सबसे अच्छा के लिए बाहरी साहसिक कार्यों और उत्तेजना
5. रजस्थन सामान्यतया प्रसिद्ध के रूप में ऊँट देश का बीकानेर स्थित तक उत्तर

RESULTS

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

In Table 2 (Correlation (ρ) values) correlation value for NIST is .110 and .360 these values are for sample size 20 and for one and two number of references which is significant at 5% level of significance. A similar result is seen in the case of sample size 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.

CONCLUSION

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. 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.

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.

REFERENCES

1. 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.
2. 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.
3. Andrei Popescu-Belis (2002): "An experiment in comparative evaluation: humans vs. computers", ISSCO/TIM/ETI, University of Geneva, 55-64.
4. 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.
5. Bohan Niamh, Breidt, Volk Martin, (2000): "Evaluating Translation Quality as Input to Product Development", 2nd International Conference on Language Resources and Evaluation, Athens.
6. Coughlin, D. (2003) "Correlating Automated and Human Assessments of Machine Translation Quality" in *MT Summit IX, New Orleans, USA*, 23–27.
7. Deborah Coughlin, (2003): "Correlating Automated and Human Assessments of Machine Translation Quality", In Proceedings of MT Summit IX. New Orleans, 63-70.
8. 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.
9. 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.
10. 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.
11. http://en.wikipedia.org/wiki/Evaluation_of_machine_translation
12. http://en.wikipedia.org/wiki/History_of_machine_translation
13. Jesus´ Angel Gimenez Linare (2008): "Empirical Machine Translation and its Evaluation", Artificial Department the Languages Systems Informatics University, 27-38.
14. 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.
15. 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.
16. 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.
17. Rao, Durgesh (2001): "Machine Translation in India: A Brief Survey", National Centre for Software Technology Gulmohar Road 9, Juhu, Mumbai 400049, India, 21-23.
18. 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.

19. Sanjay Kumar Dwivedi, Pramod Premdas Sukhadeve (2010): "Machine Translation System in Indian Perspectives", Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow, India.
20. Tomer Neeraj and Sinha Deepa, "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-2012, 48-58.
21. Tomer Neeraj and Sinha Deepa, "Evaluating Machine Translation Evaluation's modified BLEU Metric for English to Indian Language Machine Translation", in International Journal of Emerging Sciences (IJES).
22. Tomer Neeraj and Sinha Deepa, "Evaluation of Modified-BLEU Metric for English to Hindi Language Using ManTra Machine Translation Engine", in International Journal of Advanced Research in Electronics & Communication Engineering (*IJARECE*), Volume 1, Issue 4, October 2012, 103-108.
23. Tomer Neeraj and Sinha Deepa: "Evaluating Machine Translation Evaluation's NIST Metric for English to Hindi Language Machine Translation", paper accepted in The International Journal of Multidisciplinary Academy IJMRA for November-2012 issue.
24. Tomer Neeraj, "Evaluating Machine Translation (MT) Evaluation Metrics for English to Indian Language Machine Translation", Ph.D. Thesis 2012, Banasthali University, Banasthali.
25. Tomer Neeraj, Sinha Deepa and Piyush Kant Rai, "Evaluating BLEU Metric for English to Hindi Language Using ManTra Machine Translation Engine", is communicated in International Journal of Advanced Research in Computer Science-IJARCS.
26. Tomer Neeraj, Sinha Deepa and Piyush Kant Rai, "Evaluating Machine Translation Evaluation's F-Measure Metric for English to Hindi Language Machine Translation", in International Journal of Academy Research Computer Engineering and Technology-IJARCET, volume1, Issue 7, September 2012, 151-156.
27. Vannatta Rachel: "Statistics in Education Course Packet", EDFI 641. Online available http://personal.bgsu.edu/~rvanna/packetspring 09.pdf.

Author 1

Neeraj Tomer
tneeraj12@rediffmail.com
9460762117

Area of Interest:

- Machine Translation and Indian Language Technology
- Theoretical Computer Science and related technologies

Academic Qualification:

Ph.D (thesis submitted) in Computer Science, Banasthali University, Banasthali.
MCA, Maharishi Dayanand University, Rohtak 2005.
Master of Economics, Kurukshetra University Kurukshetra 1999.
Bachelor of Economics, Kurukshetra University Kurukshetra 1997.

Employment History:

Post graduate and graduate teaching at Mahatma Gandhi Institute of Applied Sciences, Jaipur as a lecturer from July 2003 to August 2006. As a Research Associate at Banasthali University Banasthali in 2007. As a lecturer at LCRT College of Education Panipat from August 2007 to July 2010. As an Assistant Professor at SINE International Institute of Technology, Jaipur from August 2010 to March 2012.

- Papers Published : 2
- In Press : 2
- Communicated : 3

Seminar and Conferences Attended: 5

Research Guidance:

Guided 3 students for their dissertation work at PG (M.Sc) level.

Future Plans: To grow academically

Author 2

Deepa Sinha
Associate Professor
Department of Mathematics
South Asian University
Akbar Bhawan
Chanakyapuri, New Delhi 110021 (India)
Cell No: 08744022273
deepasinha2001@gmail.com

Research Area: **Graph Structures**

Academic Qualification:

M.Sc., Ph. D. (University of Delhi), CSIR-NET (twice)

Future Plans: To grow academically

Achievements: CSIR_NET (qualified Twice)

Publications:

(a) Books: one
(b) Research Papers: 27

Conference/workshop/symposium attended: 39

Invited talks delivered: Ten

Papers presented: 23

Work experience: 16 years

Served several Institutions for graduate and post graduate courses, particularly Khalsa Girls Degree College (1996-1999), Delhi College of Engineering (1999-2004), Banasthali University (2004-2012).
Seven students got awarded their M. Phil. Degree under her supervision.
Three students got awarded their Ph.D. in the year 2011-2012.
Have refereed several research papers for National and international Journal of high impact factor like Discrete Applied Mathematics, Graphs and Combinatorics, International Journal of Physical Sciences etc.
Sessions chaired in the National/ International conferences: four