Evaluating F-Measure Metric Using ManTra Machine Translation Engine in Tourism domain for English to Hindi Language

Dr. Neeraj Tomer
Department of Computer Science
Karnal Institute of Technology and Management
Karnal, India

Abstract: The present research work aims at studying the Evaluation of Machine Translation Evaluation's F-Measure Metric for English to Hindi for tourism domain. This work will help to give the feedback of the Machine Translation engines. Evaluation of Machine Translation is required for Indian languages because the same Machine Translation systems 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 Machine Translation. The main objective of Machine Translation is to break the language barrier in a multilingual nation like India.

Keywords: MT – Machine Translation, MTE- Machine Translation Evaluation, EILMT –Evaluation of Indian Language Machine Translation, EMT – Evaluation of 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 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 "Evaluating F-Measure Metric Using ManTra Machine Translation Engine in Tourism domain for English to Hindi Language". 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 used ManTra MT engine for this work.

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.

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.

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 F-Measure, 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

Table 1: 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: The main markets of Udaipur are Palace Road, Hathi Pol, Bada Bazaar, Bapu Bazaar and Chetak Circle. Rajasthali, is the approved emporium of the Rajasthan government.

Candidate Sentence: महल रोड, हाथीपोल, बड़ा बाजार, बापू बाजार तथा चेतक सर्किल उदयपुर के मुख्य बाजार हैं। राजस्थली, राजस्थान सरकार द्वारा अनुमोदित ऐम्पोरियम है।

Reference Sentences:

R1: उदयपुर के मुख्य बाजार पैलेस रोड़, हाथी पोल, बड़ा बाजार, बापू बाजार और चेतक सर्किल हैं, राजस्थली राजस्थान सरकार द्वारा स्वीकृत ऐम्पोरियम है।

R2: उदयपुर के मुख्य बाजार पैलेस रोड़, हाथी पोल, बड़ा बाजार, बापू बाजार और चेतक सर्किल हैं। राजस्थली राजस्थान सरकार का स्वीकृति प्राप्त विक्रय केन्द्र है।

R3: महल रोड, हाथीपोल, बड़ा बाजार, बापू बाजार तथा चेतक सर्किल उदयपुर के मुख्य बाजार हैं। राजस्थली राजस्थान सरकार द्वारा अनुमोदित emporium है।

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. 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

IV. AUTOMATIC EVALUATION BY MODIFIED-BLEU METRIC

We used Modified-BLEU evaluation metric for this study. This metric is specially designed for English to Hindi. Modified-BLEU metric, designed for evaluating MT quality, scores candidate sentences by counting the number of n-gram matches between candidate and reference sentences. Modified-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. 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 Modified-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.

Table 2: Human Evaluation and F-Measure

Evaluation scores

S. No.	F-Measure Score			
	Human Eval.	one no. of reference	two no. of references	three no. of references
1.	.5	0.1155	0.1768	0.25
2.	7.5	0.1443	0.202	0.2121
3.	1	0.2211	0.2175	0.2163
4.	1	0.1538	0.1723	0.173
5.	7.5	0.1568	0.1903	0.2055

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.

The Spearman's rank correlation coefficient is given as (when ranks are not repeated)-

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

In equation (1) d is the difference between corresponding values in rankings and n is the length of the rankings. For correlation analysis we calculate the correlation between human evaluation and automatic evaluation (F-Measure) one by the Spearman's Rank Correlation method as shown in the table below.

Table 3: Correlation between Human Evaluation and F-Measure

	Human		F-Measure	
Spearman's rho	Human	Correlation Coefficient	1.000	.157
		Sig. (1-tailed)	-	.254
		N	20	20
	F-Measure	Correlation Coefficient	.157	1.000
		Sig. (1-tailed)	.254	-
		N	20	20

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

Table 4: Correlation (ρ) values

Sample Size	ρ values		
	one no. of reference	two no. of references	three no. of references
20	.157	.262	.228
60	.028	.092	.117
100	.099	.088	.126
200	.446	.436	.446
300	.464	.464	.455
500	.404	.357	.359
1000	.365	.353	.356

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. The test statistic is calculated as:

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

In equation (2) $\bar{x}_1$ 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.

For statistical analysis we calculate the Z-test one by one by the formula given above and the calculated as shown in the table below.

Table 5: p-values of output of ManTra 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	0.0281	0.0505	0.0281
300	0.0307	0.0307	0.0359
500	0.184	0.0475	0.0359
1000	0.0901	0.0409	0.0869

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. Shilpagram is designed on the concept of village with little emphasis on the modern concept.
2. The royal city Bikaner has a timeless charm like no other.
3. There is also a library inside the Lalgarh palace which has a large collection of sanskrit manuscripts.
4. The Bhandeshwar & Sandeshwar Temple were built by two brothers and are dedicated to Jain Teerthankar, Parsvanathji.
5. The Camel Festival is held in the month of January every year and is organized by the Department of Tourism, Art & Culture, and Rajasthan.

Candidate Sentences (translated by ManTra Machine Translation Engine):

1. शिल्पग्रम ने आधुनिक संकल्पना पर छोटा बल के साथ गाँव के संकल्पना पर डिजाइन किया जाता है
2. शाही नगर बीकानेर कालातीत मनोहरता जैसे कि कोई अन्य हैंF
3. लाल्गई राजभवन के अंदर पुस्तकालय भी जिसके बड़ा का संग्रहण संस्कृत पाण्डुलिपियाँ है
4. भंडेश्वार और संडेश्वार मंदिर ने दो भाई द्वारा निर्माण किये गये और जैन टीर्थकर, पर्सवनथ्जि के समर्पण किया जाता है
5. ऊँट त्यौ◌ेहार जनवरी प्रति वर्ष का महीना रोक रखी ता है और पर्यटन विभाग, कला और संस्कृति द्वारा, और राजस्थान संगठित किया जाता है

VI. RESULTS

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

In Table 4 (Correlation (ρ) values) correlation value for F-Measure is .365 and .356 these values are for sample size 1000, for one and three number of references which is insignificant at 5% level of significance and same result are seen for the sample sizes 200, 300 and 1000.

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.

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, I pronounce utmost of gratitude to Late Prof. Rekha Govil, Vice Chancellor, Jyoti Vidyapith, Jaipur Rajasthan.

REFERENCES

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. Tomer N., Sinha D. and Rai P.K. (2012): "Evaluating BLEU Metric for English to Hindi Language Using ManTra Machine Translation Engine", International Journal of Advance Research in Computer Science -IJARCS, 3(7), 318-322.
7. Tomer N. "Evaluation on Machine Translation" in National Conference on "Advancement in Information, Computer & Communication", organized by Department of Computer Science and Engineering and IT, Kautilya Institute of Technology and Engineering, Jaipur, technically sponsored by The Institution of Engineers (India) Indian Society for Technical Education and CSI.
8. Andrew FINCH, Eiichiro SUMITA, Yasuhiro AKIBA, "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, 2004, 619-0288, Japan, 2019-2022.
9. Deborah Coughlin, "Correlating Automated and Human Assessments of Machine Translation Quality", In Proceedings of MT Summit IX. New Orleans, 2003, 63-70.
10. Donaway, R. L., Drummey, K. W., Mather, L. A., "A Comparison of Rankings Produced by Summarization Evaluation Measures*",* Proceedings of the Workshop on Automatic Summarization, 2000, 69-78.
11. Feifan Liu, Yang Liu, "Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries", the University of Texas at Dallas Richardson, TX 75080, USA, 2008, 201-208.
12. Lin C. Y., E. hovy., "Manual and Automatic Evaluations of Summaries", In Proceedings of the Workshop on Automatic Summarization, post-Conference Workshop of ACL.2002, 45-52.
13. Papineni, K., Roukos, S., Ward, T., and Zhu, W. J., "BLEU: a method for automatic evaluation of machine translation" in ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, 2002 311–318.
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. S.Niessen, F.J.Och, G. Levsch, and H.Ney., "An Evaluation Tool for Machine Translation: Fast Evaluation for Machine Translation Research", In Proceedings of the Second International Conference on Language Resources and Evaluation. Athens, Greece, 2000, 39-45.
17. Stephan Minnis, "A Simple and Practical Method For Evaluating Machine Translation", Machine Translation 9: 133-149, 1994.
18. Yanli Sun, "Mining the Correlation between Human and Automatic Evaluation at Sentence Level", School of Applied Language and Intercultural Studies, Dublin City University, 2004, 47-50.

Author
Dr. Neeraj Tomer
Area of Interest:
- Machine Translation
- Indian Language Technology

Neeraj Tomer received Ph.D. from Banasthali University, Banasthali India in Computer Science. Earlier she did MCA and M.Sc. Computer Science, Maharishi Dayanand University, Rohtak. Masters of Economics from Kurukshetra University, Kurukshetra, Bachelor of Economics, Kurukshetra University Kurukshetra and since then she has been serving several Institutions for graduate and post graduate courses, particularly Banasthali University, Mahatma Gandhi Institute of Applied Sciences –JECRC Foundation, Jaipur.

