International Journal of Computer Science and Software Engineering
Volume 1, Number 1 (2015), pp. 49-57

http://www.irphouse.com

Evaluation of Performance Metric of Automatic Machine Translation

Vandana Sahaya* and Pardeep Singh

Computer Science & Engineering Department
National Institute of Technology Hamirpur, H.P. India
vandanasahaya16@gmail.com, avagaman@gmail.com

Abstract

Machine Translation (MT) is an automated translation from source language to target language using computer software with the same meaning and similar construction. Evaluation of any Machine Translation (MT) system is an important step to improve its accuracy. This paper evaluates the translation quality of MT systems. The evaluation of any translation system can be done in two ways, Machine Translation (MT) and human translation. Human evaluation not only provides the rank of the different MT system, but also analysis of the evaluation process at the higher level. We measured the correlation human judgments with automatic evaluation metrics. It is concluded that METEOR does not support Hindi by default, as it requires Hindi specific tools for computing stem words, synonym, etc. In this paper, we represent some of the automatic metrics and their advantages and disadvantages for the MT system.

Keywords— Automatic Evaluation machine, Human Evaluation, Machine Translation, Natural Language Processing.

I. INTRODUCTION

Translation is defined as the act of transmitting the source text language into the target text language taking into consideration Linguistic and cultural differences. In Arab World Translation, for instance, is known as "the process of understanding before explaining" that means, before starting the translation of any text the translator should have a clear understanding, semantically and culturally speaking of that source text so that the person should be able to express the real intended meaning of the target language. The competition towards establishing more business with different parts of

the world incited advanced countries in technology to look for quick and easy ways for communication. Hence, there emerged a kind of translation known as Machine Translation for the process of translation was carried out by machines. Machine Translation was an efficient way of translation and it saves both time and money, a Very large article and documents were easily translated in less time with a low amount of money.

The main task of Machine Translation is "To analyse the structure of each term or phrase within the text to be translated (source text). It fragments this structure into elements that can be very easily translated, and constitute a term of the same structure in the target language." This whole action of translation is done automatically i.e. by machines. But it does not mean that humans are totally absent from this translation process, without human translation is not at all possible because in some the case of Machine Translation is limited in terms of the vocabulary provided by their programmed dictionaries. In that case, the role of human translators is manifested in what is known as the process of pre-editing of the intended source text to be translated, and post-editing of the translated version provided by the Machine Translation (MT). If you replace human translation completely from Machine Translation would certainly face failure for, because of a simple reason, there is no Machine Translation that is capable of interpretation. For instance, it is only the human translator who is able to keep the same effect left by the source text in the target text. In this regard the automatic translation has proved its weakness, most of the time, when compared with a human translation. In case of correlation BLEU [1] showed high correlation with human judgments [2] and is still used as a standard automatic evaluation metric. BLEU and the closely related NIST [3] metric have been extensively used for comparative evaluation of the various MT systems developed under the DARPA TIDES research program, as well as by other MT researchers METEOR, an automatic metric for Machine Translation evaluation that is based on a generalized concept of unigram matching between the human-produced reference translations and machine produced translation.

Section 2, presents the study of related work in Automatic MT evaluation and discusses the contribution of each metric to the achievement of the final result. Section 3, presents the importance of human translation based on adequacy and fluency. In section 4, presents the difference between various automatic MT metrics.

II. INTRODUCTION AUTOMATIC EVALUATION

A metric is a measurement and Machine Translation uses a metric that evaluates the quality of the output. Therefore, to correlate with human judgment of quality any metric must assign quality scores. If the human scores high, a metric should score high translations, and if human assigns a low score then give low scores. The benchmark for assessing automatic metrics is human, because in any translation output humans are considered as the end-users.

The evaluation of any metrics is correlated with human judgment. Evaluation is generally done at two levels, first is at the sentence level, for a set of translated sentences [4], the scores are calculated from the metric and for the same sentences correlated

against human judgment. And the second is at the corpus level [5], where aggregates of the scores over the sentences for both metric judgments and human judgments, and then these aggregate scores are correlated. There are various automatic evaluation Metrics which are given below.

A. BLEU

BLEU was one of the first metrics which highly correlates with human judgments of quality. The most commonly used automatic evaluation metrics are BLEU and NIST. The main idea behind the metric is that "the closer output of a MT is to a professional human translation, the better it is". The scores for individual segments are calculated by the metric, generally sentences—then take averages of the whole corpus for a final score. At the corpus level, the score has been given to correlate highly with human judgments of quality.

To compare a candidate translation against multiple reference translations this metric uses a modified form of precision. Using only one reference BLEU performs less well. BLEU supports multiple references, which makes it hard to obtain an estimate of recall. Therefore, recall is replaced by the Branch Penalty, but state that Branch Penalty is a poor substitute for recall [6]. Reference [6, 7, 8] include recall in their metrics and get a better correlation with human judgments compared with BLEU. The main issue [9] in BLEU computes the same modified precision metric using n-grams. Another problem with BLEU scores is that they tend to favour short translations, which can produce very high precision scores, even using modified precision.

$$\text{Precision (P)} = \frac{\text{Number of words from the candidate that are found in the reference}}{\text{Total Number of words in the candidate}} \quad (1)$$

$$\text{Recall (R)} = \frac{\text{Number of words from the candidate that are found in the reference}}{\text{Total Number of words in the candidate}} \quad (2)$$

In BLEU metric we calculate the Geometric mean of the test corpus, modified precision scores and then multiply the result by an exponential Brevity penalty factor.

First, we compute the geometric average of the modified n-gram precisions [P_n] using n-grams up to length N and positive weights W_n summing to one. We calculate the Branch Penalty BP,

$$BP = \begin{cases} 1 \text{ If } c > r \\ e^{(1-\frac{r}{c})} \text{ If } c \leq r \end{cases} \quad (3)$$

Where c is the length of the candidate translation and r is the effective reference corpus length then,

$$BLEU = BP.\exp(\sum_{n=1}^{N} WnlogPn) \quad (4)$$

In our baseline, we use N= 4 and uniform weights W_n= 1/N.

B. NIST

NIST [11] Metric is based on the BLEU metric, but with some alterations. Where BLEU simply calculates n-gram precision adding equal weight to each one, this metric also calculates how informative a particular n-gram is. That is to say when a correct n-gram is found, the rarer that n-gram is the more weight it is given. For example, if the bi-gram matches correctly, it gets lesser weight than the correct matching of bi-gram interesting calculations, as this is less likely to occur. In case of the calculation of the penalty NIST also differs from BLEU, so far small variations in translation length do not affect much the overall score.

Information weights were computed using N-gram counts over the set of reference translations, according to the following equation:

$$info\ (w_1 \ldots\ldots w_n) = \log_2 \left(\frac{the\ \#\ of\ occurances\ of\ w_1 \ldots . w_{n-1}}{the\ \#\ of\ occurances\ of\ w_1 \ldots w_n} \right) \quad (5)$$

NIST's formula for calculating the score is

$$= \sum_{n=1}^{N} \left\{ \sum_{\substack{all\ w_1 \ldots w_n \\ that\ co-occur}} Info \left(\frac{w_1 .. w_n}{\sum_{\substack{all\ w_1 .. w_n \\ in\ sys\ output}} (1)} \right) \right\} . \exp \left\{ \beta \log^2 \left[\min \left(\frac{L_{sys}}{L_{ref}}, 1 \right) \right] \right\} \quad (6)$$

Where

β is chosen to make the brevity penalty factor (BPF) = 0.5. When the # is number of words in the system output is 2/3rd of the average number of words in the reference translation, $N = 5$ and averaged over all reference translations = the average number of words in a reference translation $Lref$ = the number of words in the translation being scored

C. METEOR

The METEOR [12] metric is designed to overcome the some of the deficiencies of the BLEU and NIST, do not correlate well with human judgments at the sentence level, even when they correlate well over large test sets [13]. The METEOR metric is based on the weighted Harmonic mean of uni-gram precision and uni-gram recall. The significance of recall in evaluation metrics is calculated by METEOR metric. This metric provides better correlation than the precision because this metric is based on recall, cf. BLEU and NIST.

METEOR also includes some other features like synonymy matching, where instead of matching only on the exact word form; the metric also matches on synonyms, not found in other metrics, For example, the word "good" in the reference rendering as "good" in the translation counts as a match. METEOR metric also includes a steamer, which lemmatizes words and matches on the lemmatized forms. The basic unit of evaluation the algorithm first creates alignment [14] between the two strings first is Translation string and the other is candidate string. The alignment is a set of mappings between unigrams. The mapping can be thought as a unigram between in a string map to the other string that means mapping the unigram of the candidate string

to reference string. Every unigram of candidate string maps to zero or one unigram in the reference string. If there are two type alignments with the same number of mappings, the alignment is chosen with the fewest crosses, that is, with fewer intersections of two mappings. The score is computed as follows: Unigram precision P is calculated as:

$$\text{Precision (P)} = \frac{\text{Number of words from the candidate that are found in the reference (m)}}{\text{Total Number of words in the candidate } (W_t)} \quad (7)$$

Where m is the no of matching words in candidate translation to the reference translation. W_t is the number of unigrams in the translation string. Unigram recall R is computed as:

$$\text{Recall (R)} = \frac{\text{Number of words from the candidate that are found in the reference}}{\text{Total Number of words in the reference}} \quad (8)$$

Where m is the no of matching words in candidate translation to the reference translation. W_r is the number of unigrams in the reference string. Precision and recall are combined using the Harmonic mean in the following fashion, with recall weighted 9 times more than precision:

$$\text{Fmean} = \frac{10PR}{R+9P} \quad (9)$$

The penalty has the effect of reducing the F_{mean} by up to 50% if there are no bigram or longer matches.

$$M = F_{mean}\,(1 - P) \quad (10)$$

TABLE I. WORD MACHER IN METEOR METRICS

Module	Candidate	Reference	Match
exact	good	good	yes
stemmer	good	good	yes
synonymy	well	good	yes

D. ORANGE

ORANGE [15] is a method for comparing metrics without using human judgments. The metrics to be compared are used both for references and MT output (n best lists). ORANGE is calculated as the average rank of the preferences P then best list. ORANGE does not use any extra human involvement, but it uses the existing human references but not human evaluations.

III. HUMAN EVALUATION

The main purpose of the manual is to ensure rigor, consistency and transparency across independent evaluations, and enhance the effectiveness and quality of work. The oldest use of human judges [16] is to assess a translation's quality. Even though the human evaluation is time-consuming, but it is the most reliable method to compare different systems such as rule-based and statistical systems. The outputs of the programs were compared to human translations and evaluated based on three components. The first component was fluency, also called intelligibility that measures the discrepancy between the output and an English speaker's mental model of fluent English. The second was adequacy, which measured the degree to which the meaning expressed in the human translation was present in the MT output. The last component was in formativeness, also called fidelity, which examines the amount of information needed to present in the output.
Instruction for Evaluators to evaluate
Read the target language Translated output first judge each sentence for its comprehensibility rate on the scale 1-5.

A. Normalizing the judgments

The human judges were presented with the following definition of adequacy and fluency, but no additional instructions [17].

TABLE II. ADEQUACY AND FLUENCY FOR MANUAL EVALUATION

S.No.	Adequacy	Fluency
5	all meaning	flawless english
4	most meaning	good english
3	much meaning	nonnative english
2	little meaning	diffluent
1	None	incomprehensible

B. Evaluation Method

If scoring is done for N sentences and each of the N sentences is given a score as above, the two parameters are as follows:
Comprehensibility = (Number of sentences with the score of 2, 3, or 4) /N (11)

$$\text{Fluency} = \sum_{i=1}^{N} \frac{Si}{N} \quad (12)$$ Where S_i is the score for i^{th} sentence.

IV. DIFFERENCE BETWEEN DIFFERENT METRICS OF MT

TABLE III. DIFFERENCE BETWEEN BLEU, NIST, METEOR AND ORANGE

Automatic Metrics	**Stands For**	**Core Process**	**The Formula Used In Metrics**	**Advantage**	**Limitations**
BLEU	Bilingual evaluation understudy	BLEU is a system that automatically evaluates he output of MT engines by comparing between the one or more candidate translation (human translation) based on n-gram.	(1), (2), (3), & (4).	The strength of BLEU's is that it correlates highly with human judgments by averaging out individual sentence judgment errors over a test corpus rather than attempting to divine the exact human judgment for every sentence: quantity leads to quality	Meaningless sentence-level score. Only exact matches. Lack of recall. Geometric averaging of n-grams. Admits too much variation by using higher order n-grams for fluency and grammatically.
NIST	National institute of standards and technology	NIST metric is based on the BLEU metric, but with some alterations.	(5) & (6).	NIST score correlates better than the BLEU score on all of the corpora.	NIST gives the poorer performance for the higher values of n may be due to poor estimation of n-gram Likelihoods.
METEOR Hindi	Metric for evaluation of translation with explicit ordering	METEOR is a system that automatically evaluates he output of MT engines by comparing between the one or more candidate translation (translated by human).	(7), (8), (9) & (10).	METEOR does not rely on totally on word order. It is a unigrams matching. METEOR is having flexible word matching, allowing for morphological variants and synonyms to be taken in account, including linguistic tools like Hindi morphological analyser, Hindi word net. METEOR uses and emphasizes recall in addition to precision	To train METEOR-Hindi with a large amount of high-quality data and using features like to paraphrase match to achieve better correlation.

V. CONCLUSION

- Train METEOR-Hindi, on a large amount of high-quality data and find optimum values of weightages for various parameters:-
    - Till now the penalty and METEOR score calculated based on the separate set of empirical test data. It is required to optimize the formula by training them on separate data set and that will best correlates with the human judgment.
- More Effective Use of Multiple Reference Translations:-
    - Our current metric uses multiple references to give the best result or to correlate with human judges. It is required to be explored the idea in such a way to improve our matching reference. Recent work by provides the mechanism for producing semantically meaningful additional "synthetic" references from a small set of real references.
- Use Semantic Relatedness to Map Unigrams;-
    - So far it has been experimented with exact mapping, synonym, however the use of semantic correlation to match unigrams that have similar meanings.

REFERENCES

[1] Papineni, "BLEU a method for automatic evaluation of Machine Translation," In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp no. 311-318, 2002.

[2] C. Burch, C. Koehn, P. Monz, C. Post, M. Soricut, and R. Specia, " Findings of the 2012 workshop on statistical Machine Translation," In Proceedings of the Seventh Workshop on Statistical Machine Translation, 2012.

[3] Doddington, "The NIST automated measure and its relation to IBM's BLEU," In Proceedings of LREC Workshop on Machine Translation Evaluation, Human Evaluators Meet Automated Metrics, Gran canaria, Spain, 2002.

[4] X. Cohn, "Regression and ranking based optimization for sentence level MT evaluation," In Proceedings of the Sixth Workshop on Statistical Machine Translation, pp no. 123-129, 2011.

[5] K. Papineni, S. Roukos, T. Ward, J. Henderson and F. Reeder, "Corpus-based comprehensive and diagnostic MT evaluation Initial Arabic, Chinese, French, and Spanish Results," In Proc. of the International Conference on Human, 2000.

[6] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A Method for Automatic Evaluation of MT," In Proc. of ACL02, pp no. 311–318, 2002.

[7] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," In Proc. of ACL WIEEMMTS, 2005.

[8] Liu and C. Dahlmeier, "Tesla: Translation evaluation of sentences with linear-programming-based analysis," In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR, pp no. 354-359, 2010.

[9] A. krishnan, P. Bhattacharyya, M. Shah and M. Ritesh, "Some Issues in Automatic Evaluation of English-Hindi MT: More Blues for BLEU," In the Proceedings of 5th International Conference on Natural Language Processing (ICON), Hyderabad, India, 4-6 January, 2007.

[10] Li, "Results of the NIST Machine Translation evaluation," In Machine Translation Workshop, 2005.

[11] A. Lavie and M. Denkowski, "The METEOR metric for automatic evaluation of Machine Translation," Language Technologies Institute Carnegie Mellon University pittisburgh, 2009.

[12] A. Lavie and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments," In Proc. Of WMT07, pp no. 228–230, 2007.

[13] D. Liu and D. Gildea, "Stochastic iterative alignment for Machine Translation evaluation," Department of Computer Science university of Rochester, pp no. 536-539, 2005.

[14] C.Y. Lin and F. J. Och, "ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation," Information Sciences Institute University of Southern California, 2004.

[15] M. Snover, N. Madnani, B. Dorr and R. Schwartz, "Fluency, Adequacy, or HTER Exploring Different Human Judgments with a Tunable MT Metric," In Proceeding. of WMT09, pp no. 259– 268, 2009.

[16] Bharati, Akshar, Vineet Chaitanya, and Rajeev Sangal, "Local Word Grouping and its Relevance to Indian Languages," in Frontiers in Knowledge Based Computing (KBCS90), V.P. Bhatkar and K.M. Rege (eds.), Narosa Publishing House, New Delhi, 1991, pp. 277-296.

[17] PVS. Avinesh and G. Karthik, "Part-of-speech tagging and chunking using conditional random fields and transformation based learning," 2007.