IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol-6, Special Issue-5, Jun 2018 E-ISSN: 2347-2693

A Study of metrics for evaluation of Machine translation

K. Sourabh[1], S. M Aaqib[2], V. Mansotra[3]

[1]Dept. of Computer Science, GGM Science College, Jammu, India
[2]Dept. of Computer Science, Amar Singh Science College, Srinagar, India
[3]Dept. of Computer Science and IT, University of Jammu, Jammu, India

Available online at: www.ijcseonline.org

Abstract: Machine Translation has gained popularity over the years and has become one of the promising areas of research in computer science. Due to a consistent growth of internet users across the world information is now more versatile and dynamic available in almost all popular spoken languages throughout the world. From Indian perspective importance of machine translation become very obvious because Hindi is a language that is widely used across India and whole world. Many initiatives have been taken to facilitate Indian users so that information may be accessed in Hindi by converting it from one language to other. In this paper we have studied various available automatic metrics that evaluate the quality of translation correlation with human judgments.

Keywords**:** Machine Translation, Corpus, bleu, Nist, Meteor, wer, ter, gtm.

I. Introduction

Machine translation is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another. As internet is now flooded with multilingual information for global community, research and development giving space to Machine Evaluation plays a major role in the field of Natural Language Processing. Many MT tools like Google translate, Bing, Systran SDL etc are providing online services to translate text from one language to another. Evaluation done manually (by humans) is the most reliable way for evaluating MT systems but it is subjective, expensive, time-consuming and involves human labor that cannot be reused. In general, evaluation can be understood as judgment on the value of a public intervention with reference to defined criteria of this judgment. Automatic MT evaluation metrics play a prominent role in the evaluation of MT systems. Many automatic measures have been proposed to facilitate fast and cheap evaluation of MT systems, the most widely used of which is BLEU, NIST METEOR etc. For Hindi language evaluation METEOR-Hindi is one of the promising metrics which has gained popularity. The measure of evaluation for metrics is correlation with human judgment. This is generally done at two levels, at the sentence level, where scores are calculated by the metric for a set of translated sentences, and then correlated against human judgment for the same sentences. In this paper we have studied various automatic metrics available which correlate with human judgment.

II. Manual Translation Evaluation

Evaluation plays a very important role in examining the quality of MT output. Manual evaluation is very time consuming and prejudiced, hence use of automatic metrics is made most of the times. Some parameters taken into consideration for manual evaluation are listed below

A. Rating

Judgments are based on encoded ranking scale from 1 as lowest and 5 as a highest (Koehn & Monz, 2006).[1] . The two main metrics used in this type of evaluation are

B. Adequacy

According to the (LDC) Linguistic Data Consortium, adequacy is defined meaning expressed in the target translation. Scale for meaning preservation is as follows

5: all meaning
4: most meaning
3: some meaning
2: little meaning
1: none

C. Fluency

The target is considered only, does not take the source into account; grammar, spelling, choice of words, and style are the main criteria. A typical scale used to measure fluency is based on

5: flawless
4: good
3: non-native
2: disfluent
1: incomprehensible

Ranking

Two or more translations are offered to the judges (typically from unlike MT systems) and are required to choose the best option. The judges must decide which errors have greater impact on the quality of the translation (Denkowski & Lavie, 2010). [2]

D. Post-Editing

In Post-editing tasks an attempt is made to measure the minimum amount of editing required by a human annotator to fix machine translation output. The most widely used post-editing measure is human-targeted translation edit rate (HTER) (Snover et al., 2006).[3]

III. Automatic Translation Evaluation

Due to the high costs, lack of repeatability, subjectivity, and slowness of evaluating machine translation output using human judgments; automatic machine translation evaluation metrics were developed. Automated MT evaluation metrics are fast, scalable, and consistent which makes them very efficient to use but most of the times not reliable. Automated MT metric needs to correlate quality with respect to human translator, and produce reliable results for similar translations of the same content. Automated measures judge the output (candidate text) of a MT system against reference text. There are a number of automatic MT evaluation metrics: WER, TER, BLEU, NIST, METEOR, GTM and the list go on

Mostly all automatic metrics are based one of the following methods to calculate scores.

- **Edit Distance Based**: Number of insertions, deletions and substitutions that are being made to change candidate into reverence are counted
- **Precision Based**: Total numbers of matched unigrams are divided by the total length of candidate
- **Recall Based**: Total number of matched unigrams is divided by the total length of reference
- **F-measure Based**: Both precision and recall scores are used collectively

Automatic machine translation evaluation started with the introduction of BLEU then followed by NIST, GTM, ROUGE, CIDEr METOER and many others like [4], Blanc Ter Rose Amber Lepor Port and Meteor Hindi. Few of the metrics are studied below.

A. Word error rate (WER)

derived from the Levenshtein distance, word error rate can be computed as: WER=S+D+I/N where N=S+D+C where *S* is the number of substitutions, *D* is the number of deletions, *I* is the number of insertions, *C* is the number of the corrects, *N* is the number of words in the reference (N=S+D+C) [5]

B. *Translation Error Rate (TER)*

An error metric for machine translation that measures the number of edits required to change a system output into one of the references. [Snover, M. (2006)].[3]

C. *GTM (General text Matcher)*

Turian et al the evaluation score is obtained by sharing corresponding words between MT output and mentioned output, Not only on precision and recall but it is also based upon harmonic mean of both, known as F-measure calculated as

$$F - Measure = \frac{2PR}{P + R}$$

D. *BLEU (Bilingual Evaluation Understudy)*

proposed by, Papineni in 2000 [6] the metric is one of the most popular in the field. The fundamental idea behind the metric is that "the closer a machine translation is to a professional human translation, the better it is" Papineni et al. (2002). N-grams in the candidate translation are matched with n-grams in the reference text, where 1-gram (unigram) is a token and a bigram assessment would be each word pair. The comparison is made despite of word order. BLEU is not perfect, but offers five convincing benefits: [7]

- Calculation is quick and inexpensive
- Easy to understand
- Language independent
- Correlates highly with human evaluation
- Widely adopted

N-gram precision, Clipping and Brevity Penalty are main components of BLEU [8].

BLEU uses tailored n-gram precision a brevity penalty is introduced to compensate difference in the length of candidate and reference translations. Since the precision of 4-gram is many times 0, the BLEU score is generally computed over the corpus than on the sentence level. [4] Many up gradations have been made on the basic BLEU like Smoothed BLEU, BLEU deconstructed etc. to offer enhanced results.

Score calculation method for Blue can be:

$$BLEU = BP.exp\left(\sum_{n=1}^{N} Wn \log Pn\right)$$

$$BP = \begin{cases} 1 & (if\ c > r) \\ e^{(1-\frac{r}{c})} & (if\ c \le r) \end{cases}$$

BP (Brevity Penalty), N is length of N grams used to compute Pn and P_n Modified n gram precision

NIST (National Institute of Standards and Technology)

[Doddington 2002] **NIST** A modification of BLEU has been adopted by NIST for MT. Based on the score of Bleu, attempt is made to compute particular n-gram's usefulness i.e. how informative an n-gram is candidate text by giving it more weight depending upon its rareness. Instead of n-gram precision the information gain from each n-gram is taken into account Additionally, BP(Brevity Penalty) calculation varies somewhat as little disparity in translation text length don't affect the general score as much as in BLEU. It uses Arithmetic mean rather than geometric mean [9]. Bad correlation on sentence level with respect to human judgment still remains a problem.

E. *The METEOR(Metric for Evaluation of Translation with Explicit Ordering)*

[Satanjeev Banerjee, Lavie (2005)][10] for evaluation of machine translation output is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. It can also make use of features such as stemming and synonymy matching which are not present in other metrics. Contrary to BLEU which aims to achieve good correlation with human judgment at the corpus level, METEOR was designed to produce a good correlation at the sentence or segment level.

METEOR addresses several limitations in IBM's BLEU metric. METOER supports not only matching between words that are identical, but can also match words that are simple morphological variants and synonyms of each other, The results reported by [banerjee et al] demonstrate that all of the individual components included within METEOR contribute to improved correlation with human judgments. In particular, METEOR is shown to have statistically significant better correlation compared to unigram-precision, unigram recall and the harmonic F1 combination of the two.

Score calculation in METEOR

$$Fmean = \frac{10\,PR}{R + 9P}$$

Where First unigram precision (P) is computed as the ratio of the number of unigrams in the system translation that are mapped (to unigrams in the reference translation) to the total number of unigrams in the system translation. Similarly, unigram recall (R) is computed as the ratio of the number of unigrams in the system translation that are mapped (to unigrams in the reference translation) to the total number of unigrams in the reference translation.

In order to compute this penalty, unigrams are grouped into the fewest possible chunks, where a chunk is defined as a set of unigrams that are adjacent in the hypothesis and in the reference. The longer the adjacent mappings between the candidate and the reference, the fewer chunks there are. A translation that is identical to the reference will give just one chunk. The penalty p is computed as follows $P = 0.5 * \left(\frac{\#Chunks}{\#UnigramsMatched}\right)^3$

The final score for a segment is calculated as M below. $\mathrm{M} = \mathrm{Fmean}(1 - \mathrm{P})$

F. *METEOR-Hindi*

Ankush Gupta et.al [11] developed METEOR-Hindi, an automatic evaluation metric for a machine translation system where the target language is Hindi. METEOR-Hindi is a modified version of the metric METEOR, containing features specific to Hindi. Appropriate changes are made to METEOR’s alignment algorithm and the scoring technique. METEOR, does not support Hindi by default, as it requires Hindi specific tools for computing synonyms, stem words, etc. additional modules listed below are added to METEOR to make well-organized for Hindi.

- Local Word Group (LWG) match
- Part-of-Speech (POS) and Clause match

METEOR-Hindi achieved high correlation with human judgments significantly outperforming BLEU.

IV. Problems with BLEU/NIST metric

Reported by [Xingyi Song et.al] [7**]** a short document or sentence, there is a high probability of obtaining zero tri-gram or 4-gram precision, which makes the overall BLEU score equal zero due to the use of geometric mean..BLEU shows good performance for corpus level comparisons over which a high number of ngram matches exist. However, at a sentence-level the n-gram matches for higher n rarely occur [12]. As a result, BLEU performs poorly when comparing individual sentences. [Xinlei Chen.et.al]

BLEU supports multiple references, which makes it hard to obtain an estimate of recall. Therefore, recall is replaced by the BP, but BP is a poor substitute for recall.

BLEU with only uni-gram precision has the highest adequacy correlation (0.87), while adding higher order n-gram precision factors decreases the adequacy correlation and increases fluency

Reported by [ankush etal] BLEU is not an appropriate metric for English-Hindi evaluation because of Meaningless Sentence-level Score, Only Exact Matches (morphological variants not considered, Lack of recall and Geometric Averaging of n-grams.

V. Conclusion

While well known, shortcomings have been noted in BLEU as of late, most remarkably the absence of solid sentence-level scores. Further, it isn't appropriate for assessment of English-Hindi MT frameworks in view of the properties of Hindi, for example, rich morphology and relative free word orderings. With a specific end goal to beat the shortcomings of BLEU, a few measurements were proposed, for example, METEOR, TER. METEOR is the most reasonable for assessment of English-Hindi MT, as it offers immense flexibility in encoding parameters that show nature of understanding the translated text.

Since automatic evaluation metrics do not always correspond to human judgment. This is necessary to resolve whether future algorithms are actually improving, or whether they are merely over fitting to a specific metric.

References

[1] Philipp Koehn, Christof Monz *,"Manual and Automatic Evaluation of Machine Translation between European Languages*" School of Informatics University of Edinburgh ,Department of Computer Science Queen Mary, University of London. Proceeding StatMT '06 Proceedings of the Workshop on Statistical Machine Translation Pages 102-121

[2] Michael Denkowski and Alon Lavie Language ,"*Choosing the Right Evaluation for Machine Translation: an Examination of Annotator and Automatic Metric Performance on Human Judgment Tasks*" Proceedings of the Ninth Biennial Conference of the Association for Machine Translation in the Americas https://www.cs.cmu.edu/~mdenkows/pdf/mteval-amta-2010.pdf

[3] Matthew Snover Bonnie Dorr Richard Schwartz, Linnea Micciulla, and John Makhoul, "*A Study of Translation Edit Rate with Targeted Human Annotation*" Proceedings of association for machine translation in the Americas, pp 223-231.

[4] Aditi Kalyani, Hemant Kumud Shashi Pal Singh Ajai Kumar," *Assessing the Quality of MT Systems for Hindi to English Translation*" International Journal of Computer Applications (0975 – 8887) Volume 89 – No 15, March 2014

[5] Klakow, Dietrich; Jochen Peters (September 2002). "*Testing the correlation of word error rate and perplexity". Speech Communication*. 38 (1-2): 19–28. doi:10.1016/S0167-6393(01)00041-3. ISSN 0167-6393

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu," *BLEU: a Method for Automatic Evaluation of Machine Translation* ". Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318

[7] Jason Brownlee "*A Gentle Introduction to Calculating the BLEU Score for Text in Python* ".November 20, 2017 in Natural Language Processing" Online https://machinelearningmastery.com/calculate-bleu-score-for-text-python/

[8] Xingyi Song and Trevor Cohn and Lucia Specia," *BLEU deconstructed: Designing a Better MT Evaluation Metric*" University of Sheffield Department of Computer Science Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)

[9] Doddington, George. (2002),"*Automatic evaluation of machine translation quality using n-gram co-occurrence statistics*".138-145. 10.3115/1289189.1289273

[10] Satanjeev Banerjee Alon Lavie ,"*METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments"* Institute Language Technologies Institute Carnegie Mellon University. Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005.

[11] Ankush Gupta and Sriram Venkatapathy and Rajeev Sangal ," *METEOR-Hindi : Automatic MT Evaluation Metric for Hindi as a Target Language*". Language Technologies Research Centre, IIIT-Hyderabad, Hyderabad, India. Proceedings of ICON-2010:8th International conference on Natural language processing, Macmillan Publishers, India.

[12] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam Saurabh Gupta, Piotr Dollar, C. Lawrence Zitnick ,"*Microsoft COCO Captions: Data Collection and Evaluation Server*". CoRR 2015 Vol: abs/1504.00325

Authors Profile

Dr. Kumar Sourabh, is currently working as Assistant Professor in the Deptt. Of Computer Sciences, G.G.M College Cluster University, Jammu. He did his doctorate on the topic "A study of Web Mining Tools for Query Optimization". Dr. Sourabh has worked as a faculty in University of Jammu, Agricultural University Jammu, MIET Engineering College Jammu. Dr. Sourabh has a teaching experience for over 8 years. His research interests include Information retrieval, Machine Translation, Natural Language Processing, Big Data Analytics. He also has a number of national and International Publications and Paper Presentations to his credit.

Dr. Syed Mutahar Aaqib, a Gold Medalist from the University of Kashmir, is currently working as Assistant Professor in the Deptt. of Computer Sciences, A S College Cluster University, Srinagar. He was awarded JRF-SRF by the Deptt. Of Science and Technology , Government of India and he did his doctorate on the topic "To Analyse Performance, Scalability and Security Mechanism of Apache Web Server viz-a-viz Contemporary Web Servers". Dr. Mutahar has worked as Scientist/Engineer-SB in National Informatics Centre, Government of India for over 5 years and is also a member of CSI and IEEE. His research interests include High Performance Computing, Machine Translation, Machine Learning and Big Data Analytics. He also has a number of International Publications and Paper Presentations to his credit.

Professor Vibhakar Mansotra, is a senior Professor in the Deptt. Of Computer Science and IT, University of, Jammu. He is also director of IT enabled services. He has produced six PhD students and acted as a guide for M.Tech students. Around five students are registerd with him for their PhD work. He is member of CSI, various academic as well as research bodies and editorial board member of various National and International Journals of high repute. His research interests include Information Retrieval , Machine Translation,Software re-engineering and Natural language processing. He also has a number of National and International Publications to his credit.