Rekha: A Reference Based Machine Translation Evaluation Metric Using Linguistic Knowledge and Contextual Embeddings

Nisheeth Joshi[1,2*], Pragya Katyayan[1,2]

{pragya.katyayan@outlook.com, nisheeth.joshi@rediffmail.in}
[1] Department of Computer Science, Banasthali Vidyapith, Rajasthan, India
[2] Centre for Artificial Intelligence, Banasthali Vidyapith, Rajasthan, India

Abstract. Since the beginning of machine translation (MT) research, MT evaluation has been an area of interest of researchers. In literature, one can find more papers on MT evaluation than on machine translation itself. This paper describes the work done on developing our MT evaluation metric which incorporates linguistic as well as word embeddings for the evaluation of MT outputs. We have studied the performance of our metric on some English to Indian language machine translation systems. For this study, a comprehensive corpus was also developed which considered sentences based on different constructs. It was found that the proposed metric provides good results which are comparable with human (evaluation) judgments.

Keywords: Machine Translation Evaluation, Automatic Evaluation, MT Evaluation Metric, Linguistic Evaluation, Word Embeddings.

1 Introduction

Since 1950s, research on machine translation (MT) has been one of the major research areas in artificial intelligence (AI). Along with the advancement of research in MT, MT evaluation has also been one of the forerunners in this area. In literature, one can find more papers on MT evaluation then on MT itself.

Since the beginning, as the technology advanced, so did the usage of that same technology in MT Evaluation which made the process more refined and robust. Initially efforts in MT development were very naïve, thus evaluating such system was very easy. Simple comprehensibility measures could access the quality of MT engine outputs. The first of the comprehensive evaluation task was the report of the ALPAC[1] (Automatic Language Processing Advisory Committee) committee which was released in 1960s. The report assessed that, "The current state of the art in MT is far

[1] http://www.hutchinsweb.me.uk/ALPAC-1966.pdf

ICIDSSD 2022, March 24-25, New Delhi, India

DOI 10.4108/eai.24-3-2022.2319019

from adequate and needs a lot of improvements. The translations produced by machines in no-where near the human levels and in most cases are non-sense translations. This leads to extra human efforts in post-editing (correction of translated text)." This report led to a lot of controversies which ultimately led to abandonment of research funding by the US government. This was the time when natural language processing (NLP) tools were developed which eventually led to the advancement of MT research. This was the era when development of rule-based (transfer based) machine translation systems was started. This can also be considered as the first revolution in MT research.

Alongside MT development, MT evaluation was also conducted. At that point in time, it was mostly human evaluation. A lot of measures for testing of MT outputs were incorporated which looked at fluency and adequacy of the translation. Some others looked at semantic adequacy of the MT outputs. Over the years, the MT system developers realized, that relying entirely on human evaluation is not feasible, as it is very time consuming and, in some cases, quite expensive too. Thus, the system developers started working on developing automatic measures for MT evaluation.

The initial breakthroughs mostly looked at automatic error analysis with metrics like word error rate (WER) and position independent error rate (PER) being developed. This all changed with the second revolution when corpus-based machine translation (CBMT), more specifically example-based machine translation (EBMT) and later on statistical machine translation (SMT) started gaining wide acceptance as the state of the art in machine translation. At this time, a lot of automatic evaluation-based matrices were developed, with BLEU (BiLingual Evaluation Under Study) (Papineni et al., 2002) becoming the standard MT evaluation metric. This was very interesting time in MT research in general and MT evaluation in particular as this metric (BLEU) accelerated the process of developing automatic MT evaluation metric. over the years, a plethora of MT evaluation metrics were developed, with Meteor (Lavie and Agarwal, 2007; Lavie and Denkowski, 2009), TER (Sover et al., 2006) and TERp (Sover et al., 2009) being the forerunners.

The third revolution of MT research saw the dawn of neural machine translation (NMT). This was the time when MT evaluation also saw overhauling in the state-of-the-art research in the evaluation arena. This raises a question in one's mind as to what is the need this overhauling? The simple layman's answer to this is that the NMT systems being developed, produce such high-quality translations that their results are at times not properly evaluated. In this context, we present, a metric that we have developed, which along with linguistic knowledge also takes into the account, the current best practices in deep learning-based NLP research.

The rest of the paper is arranged as: Section 2 briefly describes the work done in machine translation evaluation. Section 3 describes our proposed work. It first outlines the experimental setup and then describes ours proposed MT evaluation metric. Sec-

tion 4 discusses the evaluation of our MT evaluation. Finally, section 5 concludes the paper.

2 Literature Survey

During the initial years of MT research, evaluation was mostly done by humans aka manual evaluation. The first evaluation strategies were developed by Miller and Beebe-Center (1956) and Pfafflin (1965). These strategies were extended by Slype (1979) who used them for evaluation of a rule based commercial MT system (SYSTRAN) where multiple translations were provided to human evaluators and were asked to rank the same. Further, the evaluators were also asked to correct the translations which they think were not proper. Thus, by this mechanism, post editing effort was also analyzed. This was one of the initial breakthroughs in MT Evaluation as it changed the general perspective of people towards evaluation of MT systems. In 1980, a detailed evaluation for English-French MT system was performed where MT outputs and post-edited MT outputs were given to human evaluators (Falkedal, 1991). Although these evaluations were very comprehensive, but they took a lot of time. To some extent, this hampered the MT development process. Thus, MT system managers started looking for alternate measures for rapid evaluation of their MT systems which were being developed.

Since the beginning of the 21st century, automatic evaluation as become the first choice of MT system managers as it could provide rapid results. word error rate (WER) was one of the first metrics that was used in MT evaluation (Nießen et al., 2000). This metric was adopted from the automatic speech recognition (ASR) tasks where it become the de-facto standard for evaluation of such systems. Another similar metric was position independent error rate (PER) which calculated the mismatches in the translation (Tillmann et al., 1997). These metrics were more concerned identifying errors instead of provided an accuracy score. BLEU (Papineni et al., 2002) was the first metric which provided an objective score in this context. This metric became the de-facto standard for MT evaluation. Even today, every paper which discusses their development strategy, compares their results with the baseline systems using BLEU. An improved BLEU score is reported which shows the performance improvement of their MT system. Another such metric was NIST (Doddington, 2002). The difference between BLEU and NIST was that BLEU was a precision-oriented metric while NIST was a recall-oriented metric. Turian et al. (2003) developed an MT evaluation metric using F-Score which used both precision and recall for computation. Their metric compared reference translation and MT output by finding the maximum matches.

Snover et al. (2006; 2009) developed metrics which again looked at the efforts required for post editing an MT output. The metric computed the no of shifts (insertion, deletion, substitution operations) to make MT output exactly as reference translations. Denkowski and Lavie (2011; 2014) improved the Meteor metric which also used modified F-Score for computation of the accuracy of MT output.

From the start of the second decade of the 21st century, people started to look at alternate measures of MT evaluation. Measures were we do not need to provide the reference translations. Thus, MT evaluation metrics were developed which used machine learning techniques to identify the quality if MT outputs. Gammon et al. (2005) first showed the feasibility of this approach. They used support vector machine (SVM) classifier for this. Although they could produce good results. For another 5 years, no much work was done in this area. Specia et al. (2010) performed the first comprehensive study in this area and also organized several shared tasks which popularized this approach. In Indian context, Joshi et al. (2014; 2016) performed a similar study for reference free evaluation of English-Hindi MT systems.

Since the advancement of deep learning and NMT systems, a need was felt was develop metrics which could perform better evaluation and could correctly ascertain the quality of MT outputs. Thus most of the metrics started using contextual embeddings for MT evaluation. COMET (Rei, 2020) is one of the metrics in this area which used neural features. $\text{BERT}_{\text{SCORE}}$ (Zhang et al., 2020) is another such metric which uses embedding similarity for computation of its final score. RoBLEURT (Wan et al. 2021) is another such metric which uses neural embeddings for computation.

3 Rekha: An Automatic MT Evaluation Metric

3.1 Experimental Setup

For the development of MT evaluation metric, we needed some linguistic tools. Since, we wish to develop a metric which can do evaluation across levels (Lexical, Syntactic, Semantic, Contextual), we did our experiments on some of the English Indian Language Machine Translation Systems where Gujarati, Hindi, Marathi, Odia and Urdu were the target language. We evaluated the outputs of MT engines provided by Google[2] and Microsoft[3] for our study. Table 1 shows the language coding and Table 2 shows the MT engines used in the study. Thus, for example, L1 and E1 means for language pair L1 i.e English-Gujarati we registered the outputs of MT engine E1 i.e. Google Translator.

Table 1. List of Language Pairs

Language Pair No.	Lanuage Pair
L1	English-Gujarati
L2	English-Hindi
L3	English-Marathi
L4	English-Odia
L5	English-Urdu

Table 2. List of MT Engines Used

Engine No.	MT Engines

[2] https://translate.google.com/
[3] https://www.bing.com/Translator

E1	Google Translator
E2	Microsoft Translator

The first linguistic tool/resource we needed was a stemmer for the language we need to work upon. Thus, we have used the stemmers developed at our lab[4]. Next, we needed word embeddings. For this we have used the ones that are freely available through fasttext (Joulin et al., 2016; Bojanowski et al., 2017).

Further, we needed a parallelly aligned gold corpus, using which we can get outputs of MT engines and which can be compared with gold reference translations. Since, we wanted to perform comprehensive evaluation of our evaluation metric, we had developed a corpus which considered sentences across different constructs. The list of constructs that were used in our study are summarized in table 3.

Table 3. Constructs used in Corpus.

S.No.	Construct
1	Simple Construct
2	Infinitive Construct
3	Gerund Construct
4	Participle Construct
5	Appositional Construct
6	Initial Adverb
7	Coordinate Construct
8	Copula
9	Wh Structure
10	Relative Clause
11	Discourse Construct

3.2 Working of Rekha

In order to measure the effectiveness of MT output, we present an automatic MT Evaluation matric; REKHA (Robust Evaluation through Knowledge Harnessing Approach). Our metric performs evaluation by looking at the sentence across different levels. These are:

Lexical Level: Here, we look at word and word groups (n-grams) for matching the outputs of MT engines with reference translations by creating a window of n-grams and matching the longest common subsequence.

[4] https://www.copyright.gov.in/frmStatusGenUser.aspx (with Diary No. 15487/2017-CO/SW and copyright no. SW-15487/2017.

Syntactic Level: Here, we look at shallow syntactic level by matching the remaining lexicons in the sentences (both MT output and reference translation) by stemming and evaluating for a match.

Semantic Level: The remaining lexicons of both the classes of sentences are then matched by looking at their word embeddings. If the two words in MT output and reference translations are found to be equivalent of each other then it is considered as a match.

The working of the metric is shown using the following example. Since, it is not feasible to show examples for each language, we are showing the working using Hindi MT outputs and reference translations.

English Sentence: Ram gave Rahim a bouquet of flowers as a gift which was taken from the flora of Jim Corbett National Park.

Reference Translation: राम ने रहीम को उपहार के रूप में फूलों का एक गुलदस्ता दिया जो जिम कॉर्बेट राष्ट्रीय उद्यान के वनस्पतियों से लिए गए थे।

MT Output (Hypothesis): राम ने रहीम को उपहार के रूप में फूलों का एक गुलदस्ता दिया जो जिम कॉर्बेट नेशनल पार्क की वनस्पतियों से लिया गया था।

The first step was to remove the punctuations in the both reference translation and MT output. Next, we divided the sentences in 4-gram combinations. For example, the 4-grams of MT output would be:

[राम ने रहीम को]	[में फूलों का एक]	[जिम कॉर्बेट नेशनल पार्क]
[ने रहीम को उपहार]	[फूलों का एक गुलदस्ता]	[कॉर्बेट नेशनल पार्क की]
[रहीम को उपहार के]	[का एक गुलदस्ता दिया]	[नेशनल पार्क की वनस्पतियों]
[को उपहार के रूप]	[एक गुलदस्ता दिया जो]	[पार्क की वनस्पतियों से]
[उपहार के रूप में]	[गुलदस्ता दिया जो जिम]	[की वनस्पतियों से लिया]
[के रूप में फूलों]	[दिया जो जिम कॉर्बेट]	[वनस्पतियों से लिया गया]
[रूप में फूलों का]	[जो जिम कॉर्बेट नेशनल]	[से लिया गया था]

Similarly, the reference translation was also divided in the 4-gram combination. Then the 4-grams of both, reference and MT outputs were matched. Out of the 21 4-grams, 13 were matched. The lexicons of the 4-grams were removed from both reference and MT output. The remaining lexicons were matched for tri-grams, by dividing them in the same format as shown for 4-grams. Here, none of the tri-grams were matched. Thus, the lexicons were divided for bi-gram matches. Here, out of the seven bi-grams only one ([वनस्पतियों से]) bi-gram was matched. After removing the lexicons of this bi-gram. The remaining lexicons were divided in uni-grams. Here, again none were matched. At this point, lexical matching was completed.

Next, the remaining lexicons of both, reference translations and MT outputs were stemmed and their root words were matched. Here, out of the six lexicons, 4 (की, लिया, गया, था) were matched. After removing these, the remaining two lexicons were matched for semantic similarity using word embeddings. Here, the remaining two lexicons were also matched.

Once the matching process was completed, the scores were calculated. First, we calculated the clip count for the n-gram (equation 1) and then the modified precision for the n-gram matching was calculated as shown in equation 2. Next, we calculated brevity penalty as shown in equation 3 where c is the length of MT output and r is the length of reference translation. Finally, the score for Rekha was computed by multiplying brevity penalty with the weighted harmonic mean of precision. For this study all the weights were kept equal, but we can change them from case-to-case basis which provides flexibility in the metric.

$$Clip_{count} = \min(Hyp_{count}, Ref_{count}) \tag{1}$$

$$Precision\ (P)_n = \frac{\sum_{c \in \{candidates\}} \sum_{n-gram \in c} Clip_{count}(n-gram)}{\sum_{c' \in \{candidates\}} \sum_{n-gram' \in c'} Clip_{count}(n-gram')} \tag{2}$$

$$Brevity\ Penalty\ (BP) = \begin{cases} 1 & if\ \ c > r \\ e^{(1-r)/c} & if\ c \leq r \end{cases} \tag{3}$$

$$Rekha = BP\ \times \exp(\sum_{n=1}^{N} w_n \log(P_n) \tag{4}$$

4 Evaluation of MT Evaluation Metric

We compared our results with human evaluation and found that the results of the metric were comparable to human evaluation in all language pairs. For this, we employed a human annotator (one each for each language pair) for evaluation of MT outputs based on ten parameters as described by Joshi et al. (2013). There parameters were based on:

1. Comparing Gender and Number of the Noun(s) of Source and Translated sentence.
2. Translation of tense in the sentence.
3. Translation of voice in the sentence.
4. Identification of the Proper Noun(s).
5. Use of Adjectives and Adverbs corresponding to the Nouns and Verbs.
6. Selection of proper words/synonyms (Lexical Choice).
7. Sequence of phrases and clauses in the translation.

8. Use of Punctuation Marks in the translation.
9. Fluency of translated text and translator's proficiency.
10. Maintaining the semantics of the source sentence in the translation.

These parameters helped the human annotator to maintain objectivity and provide precise judgement. Overall, we tested our system for 1000 sentences. These 1000 sentences were divided into 10 documents of 100 sentences each. We evaluated our metric across segments viz sentence, document, and system. At sentence level, each evaluated sentence's result was compared with human evaluated result. The same thing was done at document level were the average score of 100 sentences of that document was taken. Finally at system level, the average score of 10 documents was considered. Table 4 shows the results of Engine E1 and Table 5 shows the results of Engine E2. Further the results of both the systems were corelated with human judgements using Pearson correlation and were found to be highly corelated. Table 6 shows the results of correlation between the results of human evaluation and the results produced by Rekha.

Table 4. Results of MT Evaluation for Engine E1

Language-Pair	Rekha	Human Evaluation
L1	0.6284	0.6201
L2	0.638	0.6247
L3	0.518	0.4946
L4	0.3898	0.3933
L5	0.4742	0.464

Table 5. Results of MT Evaluation for Engine E2

Language-Pair	Rekha	Human Evaluation
L1	0.6465	0.6284
L2	0.7094	0.6380
L3	0.6175	0.5180
L4	0.5832	0.4631
L5	0.4925	0.3520

Table 6. Correlation Between Rekha and Human Evaluation for E1 and E2

Language-Pair	E1	E2
L1	0.929	0.984
L2	0.960	0.981
L3	0.902	0.963
L4	0.910	0.952
L5	0.954	0.952

5 Conclusion and Future Work

In this paper we have shown the development of our metric which evaluates the results of MT outputs using linguistic resources and word embeddings. The results produced by this study are very encouraging. This suggests that this MT evaluation metric produces robust results which are at par with human evaluation. This gives our metric an advantage over traditional human evaluation as our metric is much faster as compared to human evaluations.

While doing the meta evaluation, we found that the metric still lacks some capabilities like analyzing the paraphrases. Thus, as a direct extension to this study, we would like to include paraphrase matching to our metric, so that it could produce better results and in turn better correlations. Further, we would also like to extend this metric for other languages. First, we shall add the capabilities of evaluating all the Indian Languages and then the languages of the rest of the world.

Acknowledgements

This work is supported by the funding received from SERB, GoI through grant number CRG/2020/004246.

References

1. Denkowski, M., & Lavie, A. (2011). Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In Proceedings of the sixth workshop on statistical machine translation (pp. 85-91).
2. Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In Proceedings of the ninth workshop on statistical machine translation (pp. 376-380).
3. Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proceedings of the second international conference on Human Language Technology Research (pp. 138-145).
4. Falkedal, K. (1991). Evaluation methods for machine translation systems: An historical overview and critical account. Report to Suissetra, ISSCO, Geneva.

5. Gamon, M., Aue, A., & Smets, M. (2005). Sentence-level MT evaluation without reference translations: Beyond language modeling. In Proceedings of the 10th EAMT Conference: Practical applications of machine translation.
6. Joshi, N., Mathur, I., Darbari, H., & Kumar, A. (2015). Incorporating Machine Learning Techniques in MT Evaluation. In Advances in Intelligent Informatics (pp. 205-214). Springer, Cham.
7. Joshi, N., Mathur, I., Darbari, H., & Kumar, A. (2016). Quality estimation of english-hindi machine translation systems. In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies (pp. 1-5).
8. Lavie, A., & Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Proceedings of the second workshop on statistical machine translation (pp. 228-231).
9. Lavie, A., & Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation. Machine translation, 23(2), 105-115.
10. Miller, G. A., & Beebe-Center, J. G. (1956). Some psychological methods for evaluating the quality of translations. HARVARD UNIV CAMBRIDGE MA PSYCHO-ACOUSTIC LAB.
11. Nießen, S., Och, F. J., Leusch, G., & Ney, H. (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In LREC.
12. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Associa-tion for Computational Linguistics (pp. 311-318).
13. Pfafflin, S. M. (1965). Evaluation of Machine Translations by Reading Comprehension Tests and Subjective Judgments. RCT, 1(32.7), 28-4.
14. Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. arXiv preprint arXiv:2009.09025.
15. Snover, M. G., Madnani, N., Dorr, B., & Schwartz, R. (2009). Ter-plus: paraphrase, seman-tic, and alignment enhancements to translation edit rate. Machine Translation, 23(2), 117-127.
16. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of transla-tion edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers (pp. 223-231).
17. Specia, L., Raj, D., & Turchi, M. (2010). Machine translation evaluation versus quality estimation. Machine translation, 24(1), 39-50.
18. Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997). Accelerated DP based search for statistical translation. In Eurospeech.
19. Turian, J. P., Shea, L., & Melamed, I. D. (2006). Evaluation of machine translation and its evaluation. NEW YORK UNIV NY.
20. Van Slype, G. (1979). Evaluation of the 1978 version of the SYSTRAN English-French automatic system of the commission of the European communities. The Incorporated Linguist, 18(3).
21. Wan, Y., Liu, D., Yang, B., Bi, T., Zhang, H., Chen, B., ... & Chao, L. S. (2021). RoBLEURT Submission for WMT2021 Metrics Task. In Proceedings of the Sixth Conference on Machine Translation (pp. 1053-1058).
22. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.