A Better Gauging Model for the Evaluation of Automatic Machine Translation of English – Hindi Language

Pooja Malik, Mrudula Y, Anurag Singh Baghel

Abstract: The problem of language translation has prevailed in society for so long. However, up to some extent the problem is being reduced by the online available machine translation systems like Google, Bing, Babelfish, etc. But with the emergence of these Machine Translation Systems, there arises the problem of their validation. Can we trust on such translation systems blindly? Is there no scope of improvement? Are these Machine Translation systems not prone to errors? The answer to all these questions is No. So, for this purpose, we need a mechanism that can test or assess these Machine Translation systems. In this paper, we have proposed an algorithm that will evaluate such Machine Translation systems. Our algorithm is being compared with a very well-known BLEU algorithm that works very well for non-Indian languages. The accuracy of the designed algorithm is evaluated using the standard datasets like Tides and EMILLE.

Index Terms: ***Automatic Machine Translation Evaluation, Automatic Machine Translation Evaluation of English-Hindi Language Pair, Automatic Evaluation Metrics for Machine Translation***

I. INTRODUCTION

Machine Translation (MT) is one of the oldest fields in the area of Natural Language Processing. However, we have come across a far long in this field, but still we have not reached to the point where there is no further scope of improvisation. Even today, we also came across such kind of difficulties where people are still fighting with the problem of language translation. Most of them are the people who are engaged in agriculture field as such people are having very little or no knowledge of any other language apart from their mother tongue. Therefore, it is a very challenging task for such people to understand the meaning of different words or sentences, which are mostly written in English. A wide variety of Machine Translation systems are online available, but the evaluation of their accuracy and feasibility must be necessarily performed to scale them as good or bad. In simple words, we can say *Machine Translation Evaluation* (MTE) is the comparison or an assessment of the output of various MT Engines based on various criteria. It is not a simple task as it involves the lexical matching between different sentence pair.

Manuscript published on 30 August 2019.
*****Correspondence Author(s)

Pooja Malik, School of Information, Communication and Technology, Gautam Buddha University, Greater Noida, India.
Mrudula Y, Computer Science Department, University at Buffalo, Buffalo, New York, United States of America.
Anurag Singh Baghel, School of Information, Communication and Technology, Gautam Buddha University, Greater Noida, India.

There are two standard methods of MT Evaluation a) Automatic Evaluation and b) Human Evaluation. Evaluation of the MT outputs through humans is a time consuming task and is not objective in nature [1]. In addition to this, the same source sentences may attain different evaluation scores depending on the knowledge-base of different human evaluators. Therefore, there is an urgent need of automatic evaluation systems, which are faster and unbiased. A number of automatic evaluation systems have been developed for years like BLEU [2], METEOR [3], GTM [4], AMBER [5], BLLIP [6], etc. BLEU is the most extensively used evaluation metric which works well for a variety of languages like German to English, French to English, Dutch to English, Hungarian to English, Chinese to English and other European languages. A big question before researchers is to check whether BLEU and its modified versions can be used to evaluate Machine Translations in Hindi and other Indian languages so well as it is evaluating Machine Translations in other languages. We give emphasis to Hindi because it is our mother tongue and most widely spoken language in India and our main focus is on the people working in agriculture field and in the Judiciary field because people working in these fields have very less or no knowledge of the English language. The primary goal of our research is to find how much applicable BLEU and its modified versions are for evaluation of English to Hindi Machine Translations.

In this paper, we have proposed an algorithm that can evaluate MT systems for English to Hindi language pair. The algorithm is developed and tested using sentences of agriculture and judiciary domain. The algorithm is being compared with a very well-known BLEU algorithm and is giving more accurate results in terms of the efficiency of MT systems. It can be very helpful for the users to evaluate the machine translated texts from English to Hindi and the end-users of this product can be Indians Farmers, Low level judiciary Employees, Employees in Agriculture Department and Content Developers who, now, can better understand the meaning of the English text.

The organization of the paper is as follows: Section 2 discusses the related work. Section 3 explains the work proposed with the framework and its algorithms. In section 4, the experimental evaluation is conducted and the analysis is thoroughly discussed and Section 5 concludes the proposed work.

II. RELATED WORK

In the last few years, a number of automatic evaluation metrics have been proposed. The first few proposals for automatic evaluation that appeared in the 90's were given by Thompson [7] and Shiwen [8]. MT systems and MT evaluation systems were regarded as expert systems by Shiwen [8]. Thompson [7] used dynamic programming for matching string-to-string distance between clauses. The same concept was applied by Tillman [9] and Vidal [10] for machine translation. Automatic evaluation measures when compared with the human evaluation methods are cost-effective, reusable, less time-consuming and more objective in nature. Automatic evaluation methods play a crucial role in the system development cycle by allowing fast evaluations on demand. These are useful in error analysis, system comparison and system optimization [11].

BiLingual Evaluation Understudy (BLEU) was proposed by IBM [2]. The geometric mean (GM) of the modified n-gram precision is calculated. It does not use recall. However, to compensate for recall, it uses brevity penalty (BP). To check for sentences that are too short and have a high precision score, BP is considered. BLEU is known to be one of the best and the most commonly used metric for MT evaluation. However, it performs poorly while evaluating outputs from rule-based systems as compared with the statistical system outputs. All matched words are weighed equally in BLEU [12]. The brevity penalty used in BLEU score heavily penalizes small differences in length between the reference and the system translations. The BLEU metric was developed for document-level or system-level evaluations, and at times results in zero scores for sentence-level evaluations [13]. BLEU when computed at the sentence level often correlates poorly with human judgment as it computes a geometric mean of n-gram precisions. Therefore, in order to overcome this problem several smoothing techniques [14] [15] have been proposed. All smoothing techniques improved sentence-level correlations (τ) over no smoothing [16].

The shortcomings of the BLEU metric have motivated the development of a number of other metrics. These metrics try to improve on the BLEU baseline correlation with human judgments [17]. The limitations of BLEU were studied in detail by Song [18] and Tiedemann [19]. The limitations of BLEU for English-Hindi were studied by Ananthakrishnan [20]. A mixed set of approaches was used to overcome these limitations and simplify the BLEU metric. The approach used precision (P), recall (R), F-measure (F), combined P, R, F using arithmetic and geometric means, brevity penalty, clipping counts and P, R, F for 1-4 grams. *NIST* is one of the closest metrics to BLEU and is also based on n-grams. Unlike BLEU, that treats all n-grams equally, NIST assigns more weight to n-grams that are more informative i.e., it weights those n-grams that occur less frequently more heavily, according to their information value. The NIST metric uses arithmetic mean instead of geometric mean to combine results from levels up to 5-grams. It has also modified the brevity penalty to minimize the impact of small length variations [21]. As opposed to the BLEU metric, NIST does not provide an upper bound on the score, which makes comparisons between NIST evaluations very difficult.*General Text Matcher (GTM)* measures the similarity between texts. The similarity measure is based on precision and recall. The precision and recall are calculated using "maximum matching". The "Maximum Match Size" (MMS) of a bitext is the size of any maximum matching for that bitext [22]. An MMS prevents double counting of words. GTM outperforms BLEU and NIST in correlation with human scores [23], no matter what the available number of reference translations is.*Recall-Oriented Understudy for Gisting Evaluation (ROUGE)* [24] is quite similar to the BLEU metric, but unlike BLEU which is precision oriented, ROUGE is recall oriented. However, its variants use recall as well as precision to generate the scores. It was originally developed for evaluation of text summaries generated by machines, but can also be used for evaluating the MT output. A number of variations of ROUGE metric that have been developed are ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU [24] [14]. These variants use n-gram-based co-occurrence statistics, longest common subsequence (LCS) statistics, weighted LCS and skip-bigrams.*Translation Error/Edit Rate* (TER) was developed by Snover [25]. TER is an edit-distance based metric and can be defined as the minimum number of edits needed to change the candidate translation so that it exactly matches the reference translation. Here edits can include insertion, deletion, and substitution of single words and shift of word sequences. All the edit operations have equal cost, and punctuation tokens are treated as normal words and incorrect capitalization of a word is counted as an edit. A variant of TER metric called as TER-plus [26] is based on paraphrasing, used semantic and alignment enhancements in order to compute the translation edit rate. A human-based version of TER was proposed by Snover [25], called *Human-targeted Translation Edit Rate* (HTER). This is a promising method as it does not limit the assessor for comparing sentences with a given reference. However, it has a possible drawback; its results depend on the experience and training of the assessors to perform a minimal number of editions. *Metric for Evaluation of Translation with Explicit ORdering (METEOR)* was developed at CMU [3]. Some variants of METEOR are described in Lavie and Agarwal [27], Denkowski and Lavie [12] [28]. This automatic evaluation metric is based on unigram matching between the candidate translations and reference translations. Unigrams are matched on the basis of their stemmed forms, surface forms, and meanings. METEOR computes a score on the basis of unigram precision and unigram recall. It also uses a measure of fragmentation that captures the ordering of matched words in the candidate translation with respect to the reference translation. METEOR consistently outperforms BLEU in correlation with human judgments. More reference translations may be helpful [29]. Denkowski and Lavie [30] proposed a language-specific evaluation via *METEOR Universal*. The metric extracts paraphrase tables and function word lists from the bitext used to train MT systems. It also uses a universal parameter set that is learned from human judgments. *METEOR Universal* significantly outperforms the baseline BLEU on Russian (WMT13) and Hindi (WMT14). *Broad Learning and Adaptation for Numeric Criteria (BLANC)* is a family of evaluation metrics for MT, which are dynamic and trainable [31]. Different correlation measures are used to automatically learn models for adequacy and fluency.

It also uses ACS (All common skip-n-grams) algorithm that computes a weighted sum of all common skip-n-grams by estimating overlap between the reference and candidate translations. BLEU and ROUGE have been considered as special cases of BLANC. BLANC is more flexible in comparison with the other existing metrics [31].

III. PROPOSED WORK

In this paper, we have proposed a new algorithm for the evaluation of machine translated texts from English to Hindi, which includes some advance features like synonym replacement and shallow parsing which first replaces the unmatched tokens of the candidate sentences with their synonym words which are present in reference sentences.

So, once the unmatched words are replaced with their synonym words, the total count of matched unigrams will get drastically increased. Moreover, the possibility of increase in the number of matched unigrams may be high but it might not increase the number of matched bi-grams, tri-grams and so on.

It might also be possible that after synonym replacement, the semantics of the sentence may get wrong. So for this, we have incorporated another feature of shallow parsing in which after replacing the words with their synonyms, the proper dependency of all the words on the replaced word is checked and is corrected accordingly.

By performing shallow parsing after synonym replacement, the incorrect semantics of the sentences can be corrected and this new algorithm gives better results as compared to results before Shallow Parsing and Synonym Replacement.

The overall evaluation system is described in Fig. 1.

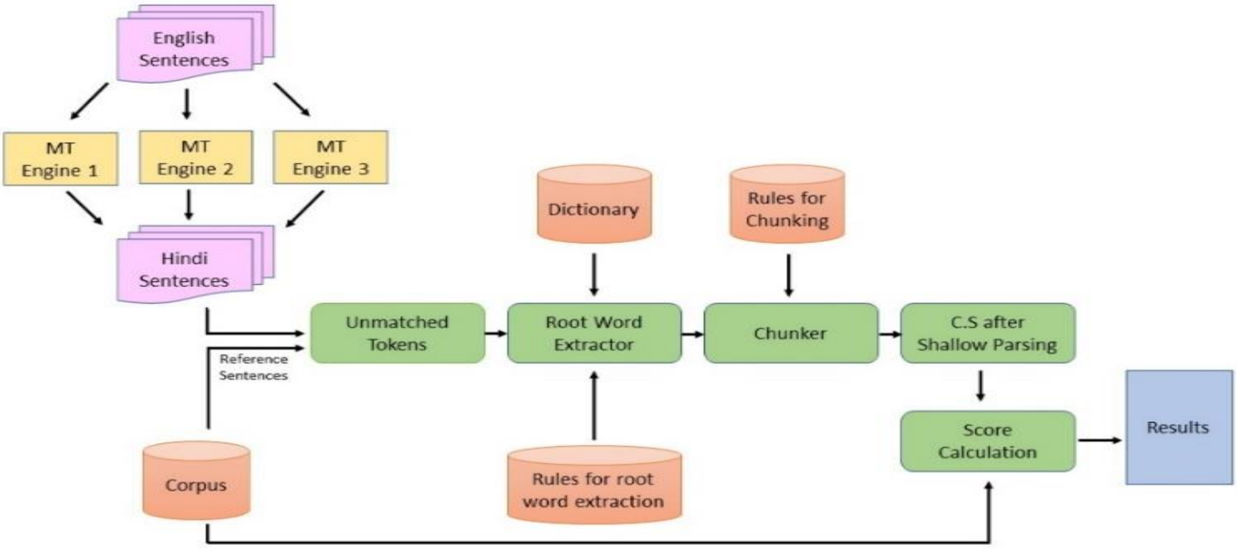


Fig. 1. The Framework of the Evaluation System

So, the proposed work can be better explained in different modules:

A.Fetching Input from the dataset

We have used two standard datasets – one from the agriculture domain (Tides) and another one from Judiciary Domain (EMILLE) for performing experiments and validating results. Both of the mentioned datasets are parallel corpus in which sentences are available in both English and Hindi Languages. First of all, we have translated given English sentences by using three different Free Online Machine Translation (FOMT) Systems – Google, Bing and Babelfish and stored all the translated sentences in three different XML files as CS1 (Google), CS2 (Bing) and CS3 (Babelfish) respectively. The given respective Hindi translations in the corpus are stored in another XML file and the same reference file is used for evaluating all three candidate files. So, we are evaluating three FOMT systems simultaneously by using the given Hindi translations in the corpus as references.

Furthermore, we have also implemented five different metrics (BLEU, METEOR, GTM, AMBER and BLLIP). The same has been discussed in one of our paper [32] where the scores are calculated for 200 sentences (100 from agriculture and 100 from judiciary) so as to see how the different metrics score different sentences translated using the same MT systems. Table I shows the scores of the same sentences for different metrics.

So after analyzing the scores of different metrics [32], we concluded that in many of the sentences, the score is lower down because of the use of either synonym of the same word or due to the different form of writing the same word. So, in our next module, we replaced the words with their synonyms so as to get matched with their references. The algorithm for n-gram matching is given in Table I.

Table I. Algorithm of n-gram matching

```
n-gram Matching()
{
Tokenize both Candidate and Reference sentences
Store Candidate tokens in $C_k$ and Reference tokens in $R_k$
Add all matched words between $C_k$ and $R_k$ into set CS and subtract them from $C_k$ and $R_k$
Add the unmatched words into set RS
If $R_k$ and $C_k$ are empty at the end, then calculate BLEU score else run Synonym Replacement()
}
```

B. Root Word Extraction Module

Hindi is morphologically rich language. Root Word Extraction is the first step towards any Indian language processing task. Root Word will be extracted on the basis of some rules known as suffix removal rules and with the help of a Hindi lexicon which are stored in the database. First, it will find the set of words that are present in candidate sentence, but not in references and also those words which are in references but not in the candidate. Then on the basis of some rules like suffix removal rules and with the help of dictionary stored in the database, it finds out the root words of all the unmatched inflected words.

Table II. Rules for Root Word Extraction

Id	Suffix	Replacement
1	ियों	ी
2	ियों	ि
3	ियाँ	ी
4	ियाँ	ि
5	ाओं	ा
6	ाएँ	ा
7	ता	No Replacement
8	ती	No Replacement
9	ते	No Replacement
10	ओं	No Replacement
11	ओं	ा
12	े	ा
13	े	No Replacement
14	ों	No Replacement
15	ी	No Replacement
16	ियाँ	No Replacement
17	ियों	No Replacement

It finds all the properties of these words which are required as their root word, number, gender and category, etc.

Once the Root Word is extracted, the further processing will be done on the Root Word itself, whether it will be synonym replacement or shallow parsing which will be our next step towards a better evaluation system for English- Hindi Machine Translation.

C. Synonym Replacement Module

Once the root word is extracted, it is checked that whether the root word of the unmatched words of the reference sentence is the synonym of the unmatched word of the candidate sentence on the basis of synset table and a reference link which is also stored in the database that contains all the possible synonyms of a Hindi word. After this, the unmatched words in the candidate sentences are replaced by their synonyms and finally the score is calculated using BLEU metric. The algorithm of Root Word Extraction and Synonym Replacement is given in Table III.

Table III. Algorithm of Root Word Extraction and Synonym Replacement

```
Synonym Replacement()
{
Add all unmatched root words of CS and RS into two separate sets t1 and t respectively
For all non-root words in CS and RS, obtain the root word by finding the applicable suffix rule from the database as given in Table II
Add all obtained root words into the t1 and t
Find the synonym for each word in t1 that matches the corresponding word in t
Replace the found synonyms in the Candidate sentence
}
```

D.Shallow Parsing Module

Shallow parsing is necessary for checking order of the words in Hindi sentence. A chunker or shallow parser identifies the simple noun, verb groups and simple adjectival and adverbial phrases in running text. It uses the basic knowledge of words provided by the lexical database to find out the different types of binding among the words of the sentence and then replacing them by suitable form. Shallow Parsing is necessary for checking correctness of the semantics of the Hindi sentence after replacing any word in the sentence by its most suitable synonym or similar word so that it gets matched with the reference sentence. A chunker or shallow parser identifies simple or non-recursive noun phrases, verb groups and simple adjectival and adverbial phrases in running text. The algorithm of Shallow Parsing is given in Table IV.

Table IV. Algorithm of Shallow Parsing

```
Shallow Parser()
{
Perform POS-Tagging on newly obtained Candidate sentence
Check for Adjective-Noun Modifier-Modified relationship in each chunk and modify the adjective according to the noun
Check for presence of ka/ke/kii karak in each chunk and modify them according to the successor noun.
Check for Verb-Noun Modifier-Modified relationship in each chunk and modify the verb according to the noun
Calculate Final Score
}
```

It uses the basic knowledge of words provided by the lexical database to find out the different types of binding among the words of the sentence and then replacing them. After replacing synonym it makes suitable changes in candidate sentence so that the semantics of sentence remain correct. It will take input from Synonym Replacement module and database.

Table V. Rules for chunking

Id	Rules
1	adj(noun)+pps
2	(adj)+noun?(noun)
3	(adjnoun)(pps)(adjnoun)
4	(nounpps)*(nounadj)?(noun)
5	(nounpps)*(nounpps)(adj)?(noun)
6	(nounpps)*(noun(adj)?pps)(adj)?(noun)
7	(adjnoun)(pps)(adjnoun)(pps)+
8	((((qad)*adj(c)?)*(qad)*adjnounpps(adj)?noun(c)?)*)(((qad)*adj(c)?)*(qad)*adjnounpps(adj)?noun)
9	((qad)*adj(c)?)*(qad)*adjnounpps(adj)?noun
10	(verbnounverb)
11	(nounverb)(verb)+
12	(adjnounverb)(verb)+
13	(nounppsadverbverb)(verb)+
14	(adj)+(nounpps)(verb)+
15	noun(kk)?
16	((hv)*(c)?)(kk)

E. Calculation of Final Score

After performing synonym replacement and shallow parsing of the candidate sentences, the final score of the candidate sentences are calculated by applying BLEU metric given by IBM.

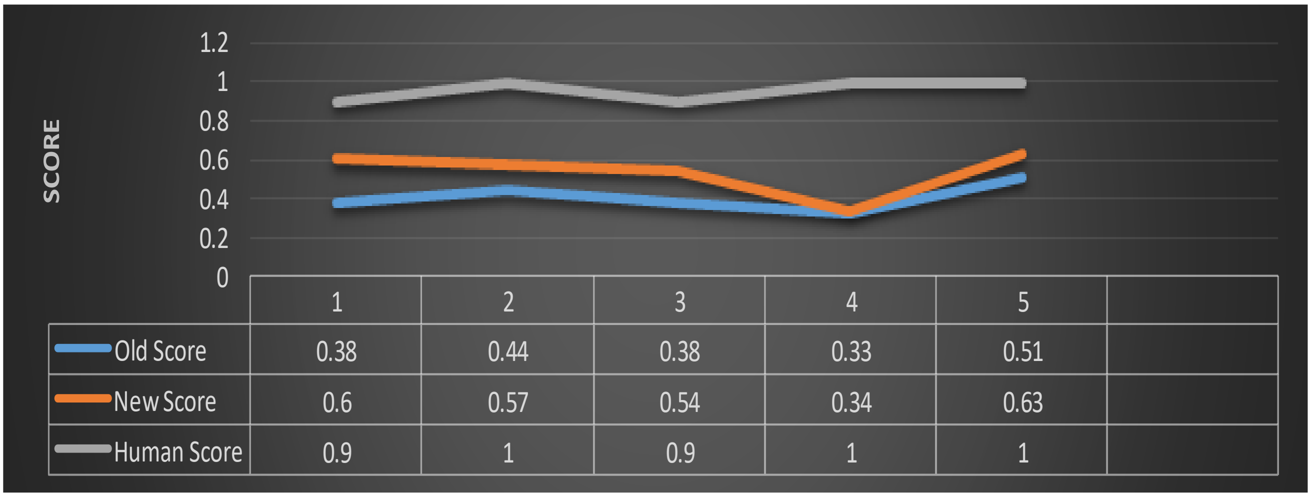
IV. RESULTS AND COMPARISON WITH EXISTING METHODS

The existing BLEU Algorithm does not considers synonyms while evaluating the candidate sentences which effects the final score of the evaluation of the candidate sentences. But now, as we have incorporated synonym replacement module in our new algorithm, which replaces the unmatched words in the candidate sentences with their synonyms and also correct the semantics of the sentences which got disturbed due to synonym replacement module. For more clarification, we have shown the results in the form of tables and graphs. The resultant scores are shown in Table VI.

Table VI. Comparison between Scores of Old Bleu Metric and New Metric

S. No.	Candidate Sentence	Reference Sentence	Candidate Sentence after performing Synonym Replacement and Shallow Parsing	Old BLEU Score	Score after Synonym Replacement	New Score after Parsing
1.	हाइड्रोपोनिक्स मृदा रहित बागवानी की एक विधि हैं ।	हाइड्रोपोनिक्स मिट्टी रहित बागवानी करने की एक विधि हैं ।	हाइड्रोपोनिक्स मिट्टी रहित बागवानी की एक विधि हैं ।	0.37608	0.6086	0.6086
2.	इस पद्धति में, पौधों को रासायनिक पोषक समाधान में उगाया जाता हैं ।	इस विधि से पौधे रासायनिक पोषक घोलों में उगाए जाते हैं ।	इस विधि में, पौधे को रासायनिक पोषक समाधान में उगाया जाता हैं ।	0.44126	0.5492	0.5708
3.	कंप्यूटर प्रबंधन के लिए एक महत्वपूर्ण उपकरण बन गए हैं ।	खेत-प्रबंधन के लिए संगणक आवश्यक उपकरण बन गए हैं ।	संगणक प्रबंधन के लिए एक आवश्यक उपकरण बन गए हैं ।	0.3800	0.5412	0.5412
4.	आज भी, भारत में किसान बैंकों से कृषि ऋण की अपनी आवश्यकताओं का केवल 15 प्रतिशत ही प्राप्त कर पाते हैं।	आज भी भारत के कृषक अपनी आवश्यकताओं का केवल 15 प्रतिशत कृषि ऋण ही बैंकों द्वारा प्राप्त कर सकते हैं ।	आज भी, भारत में कृषक बैंकों से कृषि ऋण की अपनी आवश्यकताओं का केवल 15 प्रतिशत ही प्राप्त कर पाते हैं।	0.3312	0.3364	0.3364
5.	क्षेत्र विस्तार के कारण 1950 के दशक में भारत का अनाज उत्पादन बढ़ा ।	क्षेत्र के विस्तार के कारण 1950 के दशक में भारत में खाद्यान्न उत्पादन बढ़ गया ।	क्षेत्र विस्तार के कारण 1950 के दशक में भारत का खाद्यान्न उत्पादन बढ़ ।	0.5073	0.6264	0.6264

Now for clarity, we have done a graphical comparison between these two scores of Old BLEU and Modified BLEU. Results of Graphical Representation are shown in Fig. 2.



	1	2	3	4	5	
Old Score	0.38	0.44	0.38	0.33	0.51	
New Score	0.6	0.57	0.54	0.34	0.63	
Human Score	0.9	1	0.9	1	1	

Fig. 2. Graphical Representation of the Comparison of Old Bleu Score, New Metric Score and Human Score

V. CONCLUSION

Based on the experiments performed and the results obtained, we can conclude that synonyms play a very important role in machine translations. As we have seen that, the systems using synonyms in the translated texts are getting low scores when evaluated using already existing algorithms. However, their scores improved drastically just by considering the synonyms and thereby correcting the semantics of the sentences accordingly. So, the synonym replacement module followed by the semantic correction module of the translated texts yields better scores after evaluation. Furthermore, there is no need of using large number of reference translations now as only few reference sentences will be enough to compare them with the candidate translations and the wrong semantics of the sentences will also be corrected using shallow parsing.

REFERENCES

1. A. Kalyani, H. Kumud, S. P. Singh, A. Kumar and H. Darbari, “Evaluation and Ranking of Machine Translated Output in Hindi Language using Precision and Recall Oriented Metrics,” International Journal of Advanced Computer Research, vol. 4, no. 14, pp. 54-59, March 2014.
2. K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 2002.
3. S. Banerjee and A. Lavie, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments,” in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005.
4. A. Kalyani, H. Kumud, S. P. Singh and A. Kumar, “Assessing the Quality of MT Systems for Hindi to English Translation,” International Journal of Computer Applications, pp. 41-45, March 2014.
5. B. Chen and R. Kuhn, “AMBER: A Modified BLEU, Enhanced Ranking Metric,” Proceedings of the 6th Workshop on Statistical Machine Translation, pp. 71-77, July 2011.
6. M. Pozar and E. Charniak, “Bllip: An Improved Evaluation Metric for Machine Translation,” Brown University Master Thesis, 2006.
7. H. S. Thompson, “Automatic Evaluation of Translation Quality: Outline of Methodology and Report on Pilot Experiment,” Proceedings of the Evaluators’ Forum, pp. 215-223, April 1991.
8. Y. Shiwen, “Automatic evaluation of output quality for Machine Translation systems,” vol. 8, no. 1-2, pp. 117-126, March 1993.
9. C. Tillman, S. Vogel, H. Ney, A. Zubiaga and H. Sawaf, “Accelerated DP Based Search for Statistical Translation,” in Fifth European Conference on Speech Communication and Technology, Eurospeech'97, Rhodes, Greece, 1997.
10. E. Vidal, “Finite-state speech-to-speech translation,” in IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, 1997.
11. J. Giménez and L. Màrquez, “Linguistic features for automatic evaluation of heterogenous MT systems,” StatMT '07 Proceedings of the Second Workshop on Statistical Machine Translation, pp. 256-264, June 2007.
12. M. Denkowski and A. Lavie, “METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support for Five Target Languages,” in Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, 2010.
13. C. Callison-Burch, M. Osborne and P. Koehn, “Re-evaluating the Role of BLEU in Machine Translation Research,” in 11th Conference of the European Chapter of the Association for Computational Linguistics., 2006.
14. C.-Y. Lin and F. J. Och, “Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics,” in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 2004.
15. P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero and L. Heck, “Learning deep structured semantic models for web search using clickthrough data,” in Proceedings of the 22nd ACM international conference on Information & Knowledge Management, San Francisco, California, USA, 2013.
16. B. Chen and C. Colin, “A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU,” in Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, Maryland USA, 2014.
17. O. Karolina, “A novel dependency-based evaluation metric for Machine Translation,” Dublin, 2008.
18. X. Song, T. Cohn and L. Specia, “BLEU deconstructed: Designing a better MT evaluation metric,” International Journal of Computational Linguistics and Applications, vol. 4, pp. 29-44, December 2013.
19. A. Smith, C. Hardmeier and J. Tiedemann, “Climbing Mount BLEU: The Strange World of Reachable High-BLEU Translations,” In Baltic Journal of Modern Computing (BJMC), Special Issue: Proceedings of the 19th Annual Conference of the European Association of Machine Translation (EAMT), vol. 4, 2016.
20. A. R, P. Bhattacharyya, S. Mukundan and R. M. Shah, “Some Issues in Automatic Evaluation of English-Hindi MT: More Blues for BLEU,” in Natural Language Processing ICON, 2007.
21. G. Doddington, “Automatic evaluation of machine translation quality using n-gram co-occurrence statistics,” in Proceedings of the second international conference on Human Language Technology Research, 2002.
22. I. D. Melamed, R. Green and J. P. Turian, “Precision and Recall of Machine Translation,” in Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers, 2003.
23. J. P. Turian, L. Shen and I. D. Melamed, “Evaluation of machine translation and its evaluation,” in In Proceedings of the MT Summit IX, New Orleans, USA, 2003.

24. C.-Y. Lin, “ROUGE: a Package for Automatic Evaluation of Summaries,” in Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain, 2004.
25. M. Snover, B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul, “A study of translation edit rate with targeted human annotation,” in Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, Cambridge, 2006.
26. M. G. Snover, N. Madnani, B. Dorr and R. Schwartz, “TER-Plus: paraphrase, semantic, and alignment enhancements to Translation Edit Rate,” Springer Science, vol. 23, no. 2-3, pp. 117-127, September 2009.
27. A. Lavie and A. Agarwal, “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. Association for Computational Linguistics, 2007.
28. M. Denkowski and A. Lavie, “Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems,” in Proceedings of the sixth workshop on statistical machine translation. Association for Computational Linguistics, 2011.
29. A. Lavie, “Evaluating the output of machine translation systems,” in AMTA Tutorial, 2010, p. 86.
30. M. Denkowski and A. Lavie, “Meteor Universal: Language Specific Translation Evaluation for Any Target Language,” in In Proceedings of the ninth workshop on statistical machine translation, 2014.
31. L. V. Lita, M. Rogati and A. Lavie, “BLANC: Learning Evaluation Metrics for MT,” in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, 2005.
32. P. Malik and A. S. Baghel, “A Summary and Comparative Study of Different Metrics for Machine Translation Evaluation,” in 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2018.

AUTHORS PROFILE

Pooja Malik has completed M. Tech in Computer Science and Engineering from Banasthali Vidyapeeth, Rajasthan, India. Before that, she had completed B. Tech in Computer Science and Engineering from Uttar Pradesh Technical University (now AKTU), Lucknow, India. At present, she is conducting her research in Computer Science Department at School of Information and Communication Technology, Gautam Buddha University, Greater Noida, India. Her Research Areas include Artificial Intelligence, Natural Language Processing, Machine Translation and Machine Translation Evaluation.

Mrudula Y has completed her B. Tech. (Computer Science) from Shiv Nadar University, Greater Noida. Currently, she is pursuing her MS (Computer Science) from University at Buffalo, New York. Her Research Areas include Artificial Intelligence and Machine Learning.

Dr. Anurag Singh Baghel is currently Assistant Professor of Computer Science and Engineering at Gautam Buddha University. He has also served as the Head of the Department of Computer Science and Engineering at Gautam Buddha University. His research interests are in the areas of Metaheuristics and applications, Software Engineering, and Big Data. Prior to joining Gautam Buddha University, he has worked as Lecturer in Banasthali University, Rajasthan, India from 2004 to 2011. He earned his D. Phil in 2010 from University of Allahabad, Allahabad, India.

Retrieval Number: I8965078919/19©BEIESP
DOI: 10.35940/ijitee.I8965.0881019
Journal Website: www.ijitee.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)