Improving BLEU evaluation metric for English to Hindi Machine Translations by introducing synonym and semantic correction

Under the guidance of
Mrs. Pooja Malik, Assistant Professor

MRUDULA Y (1410110251), B.Tech Computer Science

13th April, 2018

DEPARTMENT OF COMPUTER SCIENCE

SHIV NADAR UNIVERSITY

I. ABSTRACT

Evaluating machine translations by humans is a herculean task. Not only do, human evaluations take up a lot of time and are expensive, the evaluation method changes from person to person which introduces bias in the evaluation process. We focus on a widely used Automatic Evaluation metric called BLEU – Bi Lingual Evaluation Understudy. We propose to improve the BLEU metric for English-Hindi translations by introducing synonym and semantic corrections. We divided the methodology of the improvements into two stages: First is synonym correction and second is correcting the semantics in accordance with the correction done in the previous part. We tested the new BLEU metric with 15 sentences and the average of Old and New implementation of BLEU scores show that there is an improvement of 16.18% with the proposed new implementation of BLEU.

II. INTRODUCTION

With the emergence of several machine translators, evaluating the translations done by such machines/software is of paramount importance. For this purpose, Automatic Evaluation techniques are being used to generalize this process by using objective numerical scores as metric. This is, to avoid any intervention of human views and perceptions in the evaluation process. The main principle of any evaluation method that measures the translation performance of a machine is – 'the closer a machine translation is to a professional human translation, the better it is.'

In this project, we are concentrating on the use of BLEU (Bi Lingual Evaluation Understudy) evaluation metric for measuring the quality of English to Hindi translations. This metric uses n-gram comparison i.e. matching done between pairs of 'n' consecutive words of a candidate sentence (machine translated sentence) and several reference sentences (human translated sentence). This way the BLEU score tells us the correlation between a human and a machine translated sentence, and hence diligently follows the principle of evaluation methods. BLEU metrics, though one of the best known evaluation metric, has its share of disadvantages. The two main issues that we are going to focus on, in our work are: synonym consideration in n-gram calculation and grammatical error consideration in the translated corpus. When BLEU score is calculated in the conventional way, words that directly match in candidate and reference sentences are only considered, but we propose a modification in the calculation by also matching words with similar meanings (synonyms). As a consequence of such a modification, when the existing words are replaced with their synonyms in the candidate sentence, the grammar of the sentence is likely to change, which needs to be checked for.

Till now, all the metrics that have been proposed have been tested for Chinese-English [1, 2] and Arabic-English MT [2] systems. This work focuses on improving the **BLEU metric** and testing it for the **English-Hindi translations**.

Section 3 discusses how the BLEU, GTM and METEOR metric scores are calculated. Section 4 and 5 discuss the reason for choosing BLEU metric for improvement and drawbacks in BLEU metric for English-Hindi translations respectively. Section 6 talks about the methodology used for synonym replacement. Section 7 explains the semantic correction done after synonym replacement. Section 8 is the performance evaluation of improved BLEU metric. In the whole document, the candidate sentence and the reference sentence are simply addressed as candidate and reference respectively.

III. A DISCUSSION: BLEU, GTM AND METEOR

This section will briefly explain three important Automatic MT evaluation metric – BLEU, GTM and METEOR.

BLEU

BLEU which stands for BiLingual Evaluation Understudy is an Automatic MT evaluation metric proposed by IBM. It simply matches the n-grams in the candidate with the n-grams of the reference and counts the matches [1], where 'n' generally goes up to 4. These matches are independent of their position in the sentence [1]. To explain n-gram, let us take an example – when n equals 1, we match single words in the candidate with that of the references. When n becomes 2, we match two consecutive words in a sentence, starting from two words of the candidate and matching them with the pair of first two words of the reference. Then the second and third words of the candidate are considered and matched with those of the reference and so on. This process continues until the last two words in sentence are reached. In this similar method, 3-gram and 4-gram is also counted with the slight difference that three and four consecutive words respectively, are considered at a time. Here, "n" can take a maximum value equal to the length of the candidate. Once the n-grams are counted, the n-gram precision is calculated. Precision here means the proportion of the matched n-grams out of the total number of n-grams in the candidate [2]. Due to this n-gram precision calculation, the word order and word choice in the candidate are in place, but what about the length of the candidate. Here comes a quantity called the Brevity Penalty (BP). BP is used to penalize those candidates whose length is lesser than that of the average length of all the references. The formula for calculating BLEU score is as follows–

$$Bleu = BP.exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

Where, $\mathbf{p_n}$ is n-gram precision (no. of matched n-grams/total no. of n-grams in candidate)

$$BP = \begin{cases} 1, & c > r \\ e^{\left(1-\frac{r}{c}\right)}, & c \leq r \end{cases}$$

Where r = average length of all the references (in words)
c = length of the candidate (in words)
$\mathbf{w_n = 1/N}$, where N = the maximum value of "n" in n-gram being considered.

Formula 1: Calculating the BLEU score [1]

In the original paper [1], BLEU has been tested on Chinese-English MT systems.

GTM

GTM was proposed by New York University. It stands for General Text Matcher. This metric uses only precision (P) and recall (R) values to calculate the score (Formula 2).The metric counts the number of matched unigrams in the same way as in BLEU. Precision is defined as the proportion of the matched unigrams out of the total number of unigrams in the candidate and Recall is the proportion of the matched unigrams out of the total number of unigrams in the reference [2].

$$GTM = \frac{2PR}{P + R} \qquad where,$$

P is precision = total number of matched unigrams / total number of unigrams in candidate
R is recall = total number of matched unigrams / total number of unigrams in reference

Formula 2: Calculating the GTM score [6]

METEOR

METEOR is metric was developed at Carnegie Mellon University. It stands for Metric for Evaluation of Translation with Explicit ORdering. To explain this metric in simple words, the number of matched unigrams are counted in the same way as in BLEU. The precision (P) and Recall (R) values are calculated using these. Using the P and R values, F_{mean} value is calculated (Formula 3). Further, it counts the number of fragments that match in the candidate and reference. Fragment, here means the collection of consecutive words in a given sentence. For example we have a reference and candidate set as shown –

Reference: आधुनिक कृषि बहुत हद तक अभियांत्रिकी और प्रौद्योगिकी तथा जैविक और भौतिक विज्ञान पर निर्भर करती है ।

Candidate: आधुनिक कृषि अभियांत्रिकी और प्रौद्योगिकी पर और जैविक और भौतिक विज्ञान पर भारी निर्भर करती है ।

In this the number of matching fragments is 5 – आधुनिक कृषि, अभियांत्रिकी और प्रौद्योगिकी, जैविक और भौतिक, विज्ञान पर and निर्भर करती है.

The matching of these fragments is irrespective of their length. Using this, the number of chunks are calculated (Formula 3). Further this is used to calculate the Penalty and finally the METEOR score is obtained used the Fmean and Penalty values (Formula 3).

Meteor Score = F_{mean} * (1 – Penalty) where,

$$F_{mean} = \frac{10 * P * R}{9 * P + R} \ where,$$

P is precision = Total number of matched unigrams / Total number of unigrams in candidate
R is recall = Total number of matched unigrams / Total number of unigrams in reference
Penalty = 0.5 * (Number of chunks / Number of matched unigrams)^3 where,
Number of chunks = (Number of Fragments – 1) / (P – 1)

Formula 3: Calculating the METEOR score [2]

This metric has been tested for Chinese-English and Arabic-English MT systems.

IV. REASON FOR CHOOSING BLEU METRIC FOR IMPROVEMENT

Bleu metric is the most widely used metric [5, 4] because of it is simplicity, fastness and because it can be used as a target function in parameter optimization training procedures that are commonly used in state-of-the-art statistical MT systems. [4]. Moreover, it was seen in [3] that BLEU performed better than METEOR for Arabic - English translation.

We chose BLEU to be tested for English-Hindi translations. Fifteen Hindi candidate and reference sentence groups were evaluated using BLEU, GTM and METEOR, it was found that BLEU scores the least in the case of English-Hindi MT systems. (Fig 1) Hence, we wanted to improve the score by identifying the drawbacks and correcting them.

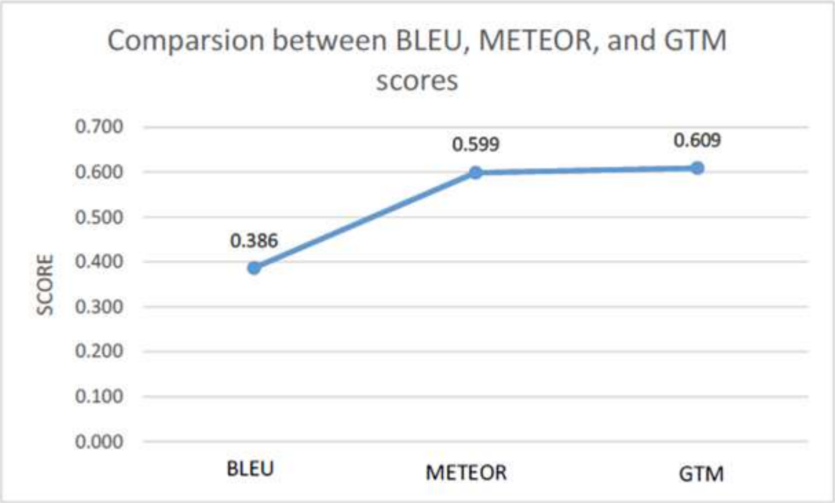

Fig 1: Comparison between the average values of BLEU, METEOR and GTM values for English to Hindi MT systems

V. DRAWBACKS OF BLEU METRIC ADDRESSED IN THIS WORK

Many words in Hindi have synonyms and hence any of those could be used in place of each other. BLEU only counts the exact matches and leaves out the words that mean the same as the word used in the reference. This causes wrong evaluation of the MT system. Hence, if the metric counts even those words whose synonyms have been used, the metric score will be improved, hence better evaluating the translation engine.

In the process of counting the synonyms of the unmatched words, we replace the existing words in the candidate with their synonyms. In Hindi, every word has some properties (like gender ('Ling' in Hindi), number ('Vachan' in Hindi)) associated with it. When a particular word is used in a sentence, other words related to it in the sentence also change due to the properties of this word. For example consider the following sentence set –

Candidate: दमन एवं द्वीप आपको ताज़गी भरी छुट्टियाँ देता है ।
Reference: दमन एवं द्वीप आपके अवकाश ताज़गी भरा बना देता है ।

The meaning of these sentences is absolutely the same. The main words that do not match in these two sentences are भरी and भरा , which are verbs and hence does not have a synonym. The other word that does not match are, छुट्टियाँ and अवकाश, which are nouns and synonyms of each other. But, replacing the one in the candidate with the the one in the reference, would make the whole sentence grammatically wrong as the gender (or 'Ling') of the word in the candidate (छुट्टियाँ) is female and that of the word in the reference (अवकाश) is male. To make it right, the verb shall also change from भरी (female) and भरा (male). This shows that on doing synonym replacement the sentence is prone to semantic errors.

Doing such a change in the candidate is justified because the MT system used the correct grammar for the word छुट्टियाँ which is भरी. This means, the MT system knows the word that needs to be used while using a particular word.

VI. IMPROVING BLEU SCORE: USING SYNONYM REPLACMENT (STAGE 1)

As discussed above, one of the major drawback of BLEU is that it considers only the words that identically match in the candidate and reference. But, Hindi is a vast language with plethora of words and most of the words have many synonyms which can be used in a sentence without changing its meaning. Hence, our claim is that, if we consider synonyms of the unmatched words in the candidate, the BLEU score will have significant improvement.

To implement this, we created a database of words and their synonyms. In the first iteration, the code checks the identical words only, and stores the unmatched words in an array. Later, the code finds the synonyms of these unmatched words in the database matches the synonym with the words in the reference. It replaces the word in the candidate with its synonym, if it finds one. Once the candidate is modified, the BLEU score is found out normally.

VII. IMPROVING BLEU SCORE: SEMANTIC CORRECTION AFTER SYNONYM REPLACEMENT (STAGE 2)

As discussed in Section 5, synonym replacement causes the semantics of a sentence (here, candidate sentence) to change. Hence, the usage of words that are related to the replaced synonym has to be changed. In this work, we have focused on the *main verbs* in the sentence that get affected due to the replacement.

To achieve this, the various grammatical aspects of the changed candidate sentence have to be known first. For getting this information, a technique called Parts-of-Speech Tagging (*POS Tagging)* is used. POS Tagging also called *Grammatical tagging* or *Word-Category Disambiguation*, is the process of assigning a part-of speech marker to each word in an input text [7].
The software that is used to do POS Tagging is generally referred to as POS Tagger. It basically lists out the linguistic knowledge contained in each word of a sentence. Linguistic knowledge includes grammatical category and grammatical features such gender, number, person etc [8]. In [8] they have proposed '*AnnCorra*', an Annotation of Corpora especially for Indian Languages, as existing ones were inadequate for Indian languages such as Hindi, Telugu etc. Annotated Corpora is the basic building block for constructing statistical models for automatic processing of natural languages. It is a widely used tool for investigators of Natural Language Processing (NLP) and related areas.[8].
The tool used in this project for POS tagging of Hindi sentences is the Hindi-POS-Tagger by [9]. The tool along with the POS tags also indicated the gender, number and root of each word in a sentence.

After the linguistic knowledge of the changes candidate sentence is known, the main verbs are identified. Their gender and number is matched with the replaced synonym. If they so not match, the form of these unmatched main verbs is changed as per the gender and number of the replaced synonym using a database of words that contains the usage of various words as per their gender and number.
This accomplishes the task of correcting the semantics of the candidate.

VIII. PERFORMANCE EVALUATION

To test the effectiveness of our new method of implementing BLEU metric, we ran the code for 15 candidate and reference sentence groups where each group consisted of one candidate and one reference sentence. We also calculated human evaluation scores as proposed in [10] to check the closeness of the new score with human determined score. This is important because the best machine translation is the one that is best understood by a human. Closeness of the automatic evaluation metric scores to human scores indicates that the metric is evaluating the translations better.

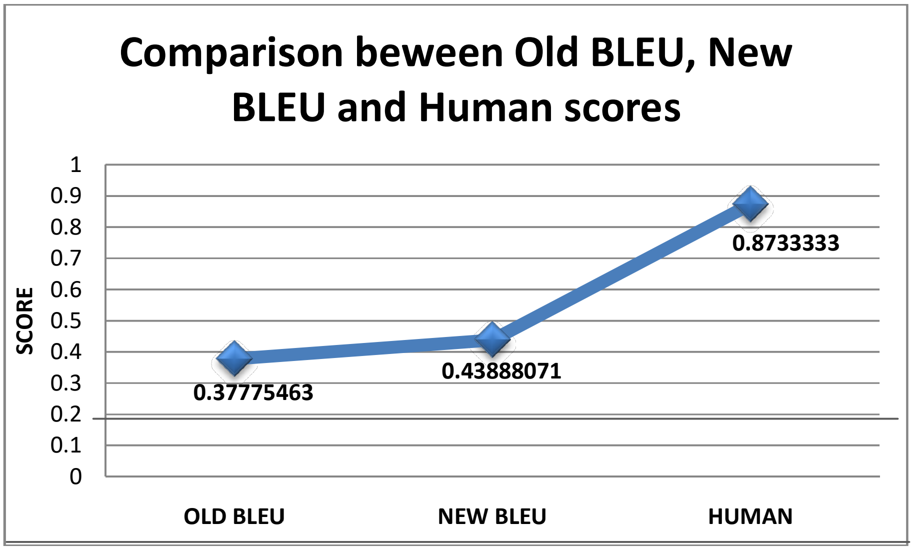

Fig 2: Shows the improvement in average BLEU score due to the new BLEU implementation.

Then we took the average of both the old and new BLEU scores. And plotted them on a graph (Fig.2).An improvement of 16.18% is seen in the BLEU score with this new method.

This method is surely limited to the size of our synonym database. The more extensive the synonym database the better will be the BLEU score. Another limitation to synonym replacement is that of the accuracy of the POS Tagger. POS Tagger basically uses machine learning algorithms to predict the linguistic knowledge of a word in a sentence. Hence, the accuracy of prediction matters. The Hindi-POS-Tagger used in this project is 91.31% accurate, which is very huge and hence does not pose much of an issue to our proposed new implementation of BLEU.

IX. NOMENCLATURE

BLEU	BiLingual Evaluation Understudy
GTM	General Text Matcher
METEOR	Metric for Evaluation of Translation with Explicit ORdering
MT	Machine Translation
NLP	Natural Language Processing
POS-Tagging	Parts-of-Speech Tagging

X. ACKNOWLEDMENTS

I would like to express my special thanks of gratitude to my guide and professor Mrs. Pooja Malik who for putting her faith in me and giving me the chance to work under her. She guided me at every step and created an environment that helped me to freely think in a directed manner towards completion of the project. I learned a lot about

research methods and practices as well as the field of Machine Translation Evaluation from her.

I am grateful to Dr. Rupamanjari Ghosh (Dean of Research), Dr. Debopam Acharya (Head of Computer Science Department) as well as the faculty, team members and administrative staff involved in organizing OUR for giving me this golden opportunity to do a research project which opened gates to many vast and interesting fields for me. Lastly, I would also like to thank the two most important people in my life, my father and mother, without whose support I would have never achieved anything in my life.

XI. REFERENCES

[1] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02). Philadelphia, PA. July 2002.pp. 311-318.

[2] Satanjeev Banerjee and AlonLavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 2005pp. 65-72.

[3] L. S. Hadla, T. M. Hailat, and M. N. Al-Kabi, "Comparative Study between METEOR and BLEU Methods of MT: Arabic into English Translation as a Case Study," International Journal of Advanced Computer Science and Applications (IJACSA), SAI Publisher, Vol. 6, No. 11, pp. 215-223, 2015.

[4] Abhaya Agarwal and AlonLavie. 2008. “METEOR,M-BLEU and M-TER: Evaluation metrics for high-correlationwith human rankings of machine translationoutput”. In StatMT workshop at ACL.

[5] Pooja Malik, Abhilasha Gupta, and Anurag Baghel. 2013. “Key Issues in Machine Translation Evaluation of English-Indian Languages”, International Journal of Engineering Research & Technology (IJERT)Vol. 2 Issue 10, October – 2013.

[6] A. Kalyani, H. Kamud, S. Pal Singh, and A. Kumar. “Assessing the Quality of MT Systems for Hindi to English Translation” In International Journal of, volume 89, 2014.

[7] Chapter 10: Parts-of-speech tagging, Speech and Language Processing. Daniel Jurafsky & James H. Martin.
Available:https://web.stanford.edu/~jurafsky/slp3/10.pdf

[8] Bharati, Akshar & Sharma, Dipti & Bai, Lakshmi & Sangal, Rajeev. (2018). AnnCorra : Annotating Corpora Guidelines For POS And Chunk Annotation For Indian Languages.

[9] Siva Reddy and Serge Sharoff. 2011. Cross language POS Taggers (and other tools) for Indian 638 languages: An experiment with Kannada using Telugu resources. In Proceedings of the IJCNLP 2011 workshop on Cross Lingual Information Access: Computational Linguistics and the Information Need of Multilingual Societies (CLIA 2011). Chiang Mai, Thailand.

[10] EuroMatrix, “Survey of Machine Translation Evaluation”. Available:www.euromatrix.net/deliverables/Euromatrix_D1.3_Revised.pdf