ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940
www.ijsem.org
Vol 14, Issuse.3 Sep 2023

EVALUATION OF A TRANSLATION WITHOUT THE NEED OF A REFERENCE TEXT USING A MODEL OF USER BEHAVIOR

Dr. shabih Nafis Ahmed, Quassim university, Buraidha, Saudi Arabia.

Abstract:

Research on how to judge the quality of a machine translation system's output is crucial. The translation is largely evaluated based on its linguistic qualities using the conventional approach of translation assessment without reference. on this research, we take into account the time and effort used by users throughout the post-editing process and suggest a novel approach to evaluating translation quality that is grounded on a model of user behavior. Extract the decision knowledge of a user's behavior by tracking and recording the process from post-editing of machine translation to the formation of the final translation; use the knowledge as an indicator of translation evaluation; and finally, assess the quality of the machine translation by combining it with a language model. The experimental findings reveal that the proposed technique is comparable to or even better than the BLEU method with one reference in terms of the Spearmen rank order correlation coefficient with human assessment when no references are available.

I Introduction:

Indicators used to gauge the quality of machine translation include fluency, accuracy, and adequacy; for fluency and accuracy, only the translation itself is taken into account, while adequacy refers to how well the translation expresses the original meaning on a semantic level. Manual assessment has the drawbacks of being expensive and complicated to operate, as well as being subjective and making it hard to establish objective and fair standards. When a reference is provided (the gold standard response), the machine will analyze the n-gram similarity between the translation and the reference to determine how well it did its job. The research demonstrates that this assessment method is frequently utilized in machine translation evaluation and has a high correlation with manual evaluation. This technique is restricted by the need for authoritative sources. However, in fact, the machine translation system's output has no reference, therefore evaluating translations without a reference has been a topic of study. Errors of the same kind tend to occur for the same machine translation system. If the user (or translator) of machine translation can gradually collect these error types during the translation post-editing task, learn and summarize some useful

ISSN2454-9940
www.ijsem.org
Vol 14, Issuse.3 Sep 2023

rules, and then apply these rules to forecast the operation cost in translation and evaluate the translation, then the user (or translator) can achieve knowledge accumulation and effective use, and efficiency can be improved. This research proposes a reference-free way of assessing translation quality based on patterns of user behavior; the approach is accurate at predicting future translation mistakes and correlates well with human judgment. The main benefit of this approach is that it dynamically learns the error types and the decision-making knowledge of operations during the post-editing of the translation, using this information to build a user behavior model; using the model to predict the possible error types in the translation; and finally, using operation cost as an indicator to evaluate the translation. In this research, we employ a machine learning technique to train a translation assessment model, which then incorporates a language model and a model of user behavior to make decisions on the quality of a translation automatically. Knowledge-based assessment methods are considerably more similar to human evaluation criteria, as shown by experiments.

II LITERATURE SURVEY:

1. a Method for Automatic Evaluation of Machine Translation:

Extensive but costly human reviews of machine translation are necessary. Human assessments may take a long time and require unrecoverable human effort. We present a technique for evaluating automated machine translation that is fast, cheap, language-agnostic, and strongly correlated with human review, all while incurring low marginal costs each run. We describe our technique as an automated alternative to highly trained human judges for situations when fast and frequent assessments are required.

2. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics:

Research and development in the field of human language technology acknowledges that evaluation is a very useful forcing function. Due to the high cost, high time commitment, and difficulty of incorporating human judgment into the MT research agenda, assessment has not been a very effective technique in the field. However, IBM presented an automated MT assessment approach at the July 2001 TIDES PI conference in Philadelphia, which may give quick feedback and direction in MT research. Their approach, which they label a "evaluation understudy," involves comparing the statistical properties of small word sequences (word N-grams) from MT output to those of expert reference translations. A translation is deemed more accurate the more of these N-grams it shares with the reference translations. The notion is so simple, but so brilliant. More importantly, IBM demonstrated a robust relationship between these machine-generated rankings and human evaluations of translation quality.1 So, DARPA gave NIST a mandate to create an MT assessment facility based on IBM's research. NIST has released this tool, and it will be used as the standard by which TIDES MT studies are judged moving forward.

Since machine translation technology is being studied, several machine translation evaluation

ISSN2454-9940
www.ijsem.org
Vol 14, Issuse.3 Sep 2023

methods have emerged, such as reference-based and terminology-based approaches. Three widely-used techniques are BLEU, NIST, and WER [1-3]. For each candidate sentence T" a modified n-gram precision is computed with regard to its reference sentences R, in BLEU [1], which is based on the n-gram accuracy of the geometric mean approach. It penalized for brevity to prevent missing translations. NIST [2] presented a metric with comparable properties to BLUE, however with a different shortness and the substitution of information weight for n-gram accuracy and the arithmetic mean for the geometric mean. WER The Levenshtein distance minimal word error rate technique is the basis for [3]. Jones employs the harmony of the syntax tree as a gauge of translation quality [4], in contrast to the prevalent practice of reference-free translation assessment based on syntactic structure and test sets. Sentence perplexity according to a trigram language model of a training corpus, source sentence length, and some translation features, such as the number and size of mappings, are used to train a classifier and simulate human scoring in artificial commentary translation sentences [5]. The quality control of the patent summary corpus proposed in [6] uses support vector machines to extract syntactic features, including the error template of pas and the error syntactic template, and then uses these features to train a classifier to assess the quality of the translation. In most cases, linguistic attributes are utilized to assess translation with or without a reference, and the user's post-editing cost is only considered in exceptional cases. In reality, from a user's perspective, machine translation is impacted directly by the cost of post-editing processes from the translation to the final product. The higher the quality of the translation, the cheaper the cost of the post-editing processes should be. Despite [7]'s suggestion of Translation Edit Rate (TER), most machine translation does not include references. The impact of using the cost of this operation to assess the machine translation will be closer to the effect of using human evaluation criteria if we are able to get information from user post-editing records and then use it to forecast the likely post-editing operations in machine translation. Thus, in this paper, we extract a helpful decision-making template, and build a user behavior model for predicting the operation cost of the translation, and it is used as an indicator of translation evaluation, from the process from post-editing the translation to forming the final translation.

III. Translation Evaluation Mechanism without Reference Based on User Behavior Model:

Modeling User Behavior: Step 3.1 Construction While the original intent of the user behavior model was to improve the efficiency of an aided translation system, we are now using it to assess the quality of the translation itself. This study improves knowledge utilization by constructing a model of user behavior, learning user decision-making information dynamically from user's dominating post-editing, and then using this knowledge to the assessment of the translation. The steps involved in creating the user behavior mode are shown in Figure 1.

User Behavior Mapping:

For optimal classification results, feature selection is crucial in pattern recognition. While we have been able to extract a significant number of rules describing user behavior from the user behavior records, using these rules as features directly will lead to data sparsity and inefficient trials. Thousands of features are mapped into the 5 dimensions

ISSN2454-9940
www.ijsem.org
Vol 14, Issuse.3 Sep 2023

(modification, insertion, deletion, substitution, reordering) without taking into account the linguistic phenomena that underlie them. For example, two feature templates, [a pain thing] 7 [a painful thing] and [filler with] 7 [fill with], use the same editing operation modification to denote. The success of the experiments validates the technique.

Translation Evaluation Method Based on User Behavior Model:

In this research, a translation forecasting model is trained using the support vector machine (SVM) [10] approach, and the translation is assessed in isolation. Users rated the quality of the translation from 0 (worst) to 1 (best) in increments of 0.1 (see [8] for a detailed explanation of this metric), and we used their ratings to determine a cutoff point between good and poor translations. Then, train a classifier to predict translation quality with the use of SVMlight [11]'s tools, where the radial basis function (RBF) kernel function classifier is used.

Experimental Corpus:

When using the Google machine translation system to automatically translate from Chinese to English, the average number of words per sentence is 32.9. The source text for this experiment is a collection of 48 Chinese articles (primarily based on financial, news, science and technology exposition, review, etc.).We sectioned the translation into six pieces, gave each portion to a separate user, and tracked their progress as they post-edited the translation.

IV. Conversion of Classification Value to User Evaluation Score:

Since there are only two possible outcomes when using the aforementioned user behavior model to evaluate translation, excellent and poor, and the classifier accuracy is low, the following approach is used to get an evaluation value. The fact that SVMs are not probabilistic classifiers is one obstacle. SVMs make their categorization calls based on a decision boundary established by the support vectors they learn to recognize throughout the training process. Unseen test instances are put through a decision function to establish where on the decision boundary they fall. While the decision function results typically just reflect class assignments for the cases, we utilized them to generate confidence ratings for evaluating user behavior modes in this experiment. Each value produced by the decision function is transformed into an evaluation score using formula (4) to create a score from the SVM classifier.

Experimental Settings:

The following procedures are developed to put our assessment to the test: Using the training user behavior model as a guide, we select 572 Chinese sentences from the Internet that cover the same domains, and use them as a source language at Google, yahoo, and Microsoft's translation engines to translate Chinese into English. Four people will rate each Chinese sentence outcome. Both the sentence's fluency and its adequacy in English are graded independently, from 0 (worst) to 5 (best), in increments of 1. We calculated an overall assessment score by averaging the sentence's fluency and appropriateness with its length. If three translators rank a group translation of a single Chinese phrase in the same order, we utilize those sentences as our test corpus for the experiment. Finally, we narrowed our data set to only 232 Chinese sentences to be

ISSN2454-9940
www.ijsem.org
Vol 14, Issuse.3 Sep 2023

translated (the average length of a Chinese sentence translated by each system was 31, 28, and 30 words, respectively).

Experimental Results:

We apply the resulting n-gram and user behavior model classifiers to the test corpus, calculate the evaluation scores using formula 4, and rank the accuracy of each set of three English translations. Spearman's rank correlation coefficients [12] are used to assess the degree to which these techniques are correlated with human evaluations. Results from the experiment are summarized in Table 3; it is assumed that there is only a single reference used in the translation, and BLEU is used to rank and score each set of outputs. The findings are the same as [9], with a low correlation since translating from Chinese to English is more challenging than translating from other European languages to English.

Experimental Summary:

The template of the most common mistake kinds has a high usage rate from the perspective of user behavior. Test data usage analysis reveals that replacement is the most often used and hence most efficient process. It will take a considerable amount of time to build a user behavior template since it is dynamically created and accumulates over time. This is the primary restriction on the applicability of these experimental findings. There is a little discrepancy between this approach and the BLEU assessment with a single reference, although the latter is more often employed and more in line with practice than the former.

V. Conclusion and Future Directions:

Three factors—fluency, accuracy, and adequacy—need to be taken into account for automated review of machine translation; however, adequacy pertains to semantic evaluation, which is a challenging area of study. This study proposes a user behavior model for evaluating translations, and describes how to extract decision-making information from the user post-editing record to create such a model. The experimental findings demonstrate the superiority of assessment based on human knowledge. At the same time, the procedure is utilized to establish an assessment score between 0 and 1, allowing us to assign a value to each phrase. BLEU (with a single reference) achieves almost the same result as human assessment. Step two involves expanding the amount of feature templates and generalizing the decision-making templates gathered from users' behaviors. The experiment does not take adequacy into account; in the following stage, we will take into account alignment between the source and target languages, semantic assessment, etc., to build a more robust evaluation procedure.

References:

[1] PAPINENI, Kishore A., Salim ROUKOS, Todd WARD and Wei-ling ZHU (2002): BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of ACL 2002, pp.II-18.

[2] Doddington. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics[R].NIST Research Report, 2002.

ISSN2454-9940
www.ijsem.org
Vol 14, Issuse.3 Sep 2023

[3] Sonja Niellen, Franz 1. Och, Gregor Leusch, and Hermann Ney. An Evaluation Tool for Machithe Translation: Fast Evaluation for MT Research. In Proc. of the Second lnt. Conf, on Language Resources and Evaluation, 2000. pp.39-45.

[4] Douglas A lones, Gregory M Rusk. Toward a scoring function for quality-driven machine translation. In: Proceeding of COLlNG-2000. 2000.

[5] Quirk, Christopher(2004): Training Sentence-Level Machine Translation Confidence Measure. In Proceedings of LREC 2004.

[6] Wei Ning, Xuelei Miao, Yonghua Hu , Duo li, Dongfeng Cai. The Fourth of China machine translation Symposium. pp: 196-203, 2008

[7] Matthew Snover,Bonnie Door, Richard Schwartz, Linnea Micciulla, and John Makhoul. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of AMTA-2006, Cambridge, USA.,2006

[8] Nie&#xfffd;en, S., F. J. Och, G. Leusch, H. Ney. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In Proceedings of LREC, Athens, Greece, 2000.

[9] Leusch,G., N.Ueffing, and H.Ney. A novel string-to string distance measure with applications to machine translation evaluztion. In Proceedings of MT Summit IX, New Orleans, U.S.A 2003

[10] V. Vapnik. The Nature of Statistical Learning Theory. Springer,N.Y., 1995