International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : 10.32628/CSEIT195258

Automated Evaluation of Language Translation

Akshaya U[1], Kalpana P[2]

[1]Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India.

[2]Assistant Professor, Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, Tamil Nadu, India.

ABSTRACT

Human evaluations of machine translation are expensive and extensive. Human evaluations can take a longer time to finish and involve human labour that can't be reused. We proposed a methodology of automated machine translation evaluation that is fast, inexpensive, and language-independent, that relates highly with human evaluation, and that has only little marginal cost initially. This reduces the cost needed for translation, human labour wastage and also the time. This will benefit the developers as it is inexpensive.

Keywords : Language Translation, BLEU, NLG

I. INTRODUCTION

Human evaluations of machine translation (MT) weigh many aspects of translation, including adequacy, fidelity , and fluency of the translation (Hovy, 1999; White and O'Connell, 1994). For the most part, these various human evaluation approaches are expensive (Hovy, 1999). Moreover, they can take weeks or months to finish. This is an intensive task because developers of machine translation systems need to monitor the effect of daily changes to their systems in order to maintain consistency in their system.We believe that MT progress stems from evaluation and that there is a logjam of fruitful research ideas waiting to be released from 1So we call our method the bilingual evaluation understudy, BLEU. the evaluation bottleneck. Developers would benefit from an inexpensive automated evaluation that is fast, language-independent, and correlates highly with human evaluation. This will reduce the manual effect of checking the machine translation and will be more effective and non erroneous than the manual checking of machine translation.We propose such an evaluation method in this paper.

The closer the machine translation is to a professional human translation, the better is its quality. This is the main idea behind our proposal. To judge the quality of a machine translation, one measures its closeness to a human translated one in the form of numeric metric. Thus, our MT evaluation system requires two ingredients: 1. a numerical "translation closeness" metric 2. a corpus of good quality human reference translations We fashion our closeness metric after the highly successful word error rate metric used by the speech recognition community, appropriately modified for multiple referenced translations which allows for legitimate differences in word choice and word order. The core idea is to use a weighted average of variable length phrase matches against the referenced translations. We have selected a promising baseline metric from this family.

II. RELATED WORKS

The path to a systematic picture of MT evaluation is long and intensive.While it is difficult to write a comprehensive overview of the MT evaluation literature, certain tendencies and trends should be mentioned. First, throughout the history of evaluation, two aspects – often called quality and fidelity – stand out. Particularly MT researchers often feel that if a system produces syntactically and lexically well-formed sentences after translation (i.e., high quality output), that does not distort the meaning (semantics) of the input (i.e., high fidelity), then the evaluation is good and sufficient.System developers and real-world users often add evaluation measures, notably system extensibility like how easy it is for a user to add new words, grammar, and transfer rules and coverage (specialization of the system to the domains of interest), and price. In fact, as discussed in (Church and Hovy, 1993), for some real world applications quality may take a back seat to these factors. Various ways of measuring quality have been proposed, some focusing on specific syntactic constructions like relative clauses, number agreement etc. (Flanagan, 1994), others are asking judges to rate each sentence as a whole on an N-point scale (White et al., 1992 1994; Doyon et al., 1998), and others automatically measuring the perplexity of a target text against a n-gram language model of ideal translations (Papineni et al., 2001). The amount of agreement among such measures has never been taken into account. Fidelity requires bilingual judges, and is usually measured on an N-point scale by having judges rate how well each portion of the system's output expresses the content of an equivalent portion of one or more ideal human translations (White et al., 1992 1994; Doyon et al., 1998). A proposal to measure the quality automatically is by projecting both system output and a number of ideal human translations into a vector space, and then measuring how far the system's translation deviates from the mean of the human ones, is an intriguing idea whose generality still needs to be proved (Thompson, 1992). In 2 similar vein, it may be possible to use the above mentioned perplexity measure also to evaluate fidelity (Papineni et al., 2001). The Japanese study of 1992 (Nomura, 1992; Nomura and Isahara, 1992), paralleling EAGLES, identified two sets of 14 parameters each: one that characterizes the desired context of use of an MT system, and the other that characterizes the MT system and its output. A mapping between the two sets of parameters allows us to determine the degree of correlation, and hence to predict which system would be appropriate for which type of user.The OVUM report includes usability, customizability, scalability, reusability application to total translation process, language coverage, terminology building, documentation, and others.

III. PROPOSED SYSTEM

Typically, there are many “perfect” translations of a given source sentence. These translations may vary from one another in word choice or in word order even if they use the same words. And yet humans can clearly distinguish a good translation from a bad one.It is clear that the good translation, Candidate 1, shares many words and phrases with these three reference translations, while Candidate 2 does not. We will shortly quantify this notion of sharing in Section 2.1. But first observe that Candidate 1 shares "It is a guide to action" with Reference 1, "which" with Reference 2, "ensures that the military" with Reference 1, "always" with References 2 and 3, "commands" with Reference 1, and finally "of the party" with Reference 2 (all ignoring capitalization). In contrast, Candidate 2 exhibits only a fewer matches, and that extent is less. It is clear that a program can rank Candidate 1 higher than Candidate 2 simply by comparing ngram matches between each candidate translation and the reference translations. Experiments over large collections of translations presented in Section 5 show that this ranking ability is a general phenomenon, and not an artifact of a few toy examples. The primary programming task for a BLEU implementer is to compare n-grams of the candidate with the n-grams of the reference translation and count the number of matches. These matches are position independent. The more the matches, the better is the translation. For simplicity, we first focus on computing unigram matches. 2.1 Modified n-gram precision The cornerstone of our metric is the familiar precision measure. To compute precision, one simply counts the number of candidate translated words (unigrams) which occur in any reference translation and then divides by the total number of words in the candidate translation. Unfortunately, MT systems can overgenerate "reasonable" words, resulting in improbable, but high-precision, translations like that of example 2 below. Here the problem is clear: a reference word can be considered exhausted after a matching candidate word is identified. We formalize this intuition as the modified unigram precision. To compute this, one first counts the maximum number of times a word occurs in any single reference translation. Next, one clips the total count of each candidate word by its maximum reference count,2 adds these clipped counts up, and divides by the total (unclipped) number of candidate words.

Modified n-gram precision

The cornerstone of our metric is the familiar precision measure. To compute the precision, one can simply count the number of candidate translation words (unigrams) which occur in any reference translation and then divide it by the total number of words in the candidates translation. Unfortunately, MT systems can overgenerate "reasonable" words, resulting in improbable, but high-precision, translations like that of example 2 below. Intuitively the problem is clear: a reference word can be considered exhausted after matching candidate word is identified. We formalize this intuition as the modified unigram precision. To compute this, one first counts the maximum number of times a word occurs in any single reference translation. Next, one clips the total count of each candidate word by its maximum reference count,2 adds these clipped counts up, and divides by the total (unclipped) number of candidate words.

Modified n-gram precision is computed similarly for any n: all candidate n-gram counts and their corresponding maximum reference counts are collected. The candidate counts are clipped by their corresponding reference maximum value, summed, and divided by the total number of candidate ngrams. In Example 1, Candidate 1 achieves a modified bigram precision of 10/17, whereas the lower quality Candidate 2 achieves a modified bigram precision of 1/13. In Example 2, the (implausible) candidate achieves a modified bigram precision of 0. This sort of modified n-gram precision scoring captures two aspects of translation: adequacy and fluency. A translation that is done using the same words (n-grams) as in the references tends to satisfy the requirement. The longer n-gram matches account for fluency.

Modified n-gram precision on blocks of text

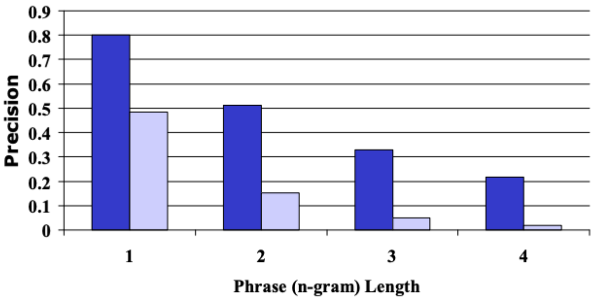
Although one typically evaluates MT systems on a corpus of entire documents, our basic unit of evaluation is the sentence. A source sentence may translate to many target sentences, in which case we abuse terminology and refer to the corresponding target sentences as a "sentence." We first compute the n-gram matches sentence by sentence. Next, we add the clipped n-gram counts for all the candidate sentences and divide by the number of candidate n-grams in the test corpus to compute a modified precision score, pn, for the entire test corpus.

$$p_n = \frac{\sum_{C\in\{Candidates\}} \sum_{n\text{-}gram\in C} Count_{clip}(n\text{-}gram)}{\sum_{C'\in\{Candidates\}} \sum_{n\text{-}gram'\in C'} Count(n\text{-}gram')}.$$

Ranking systems using only modified n-gram precision

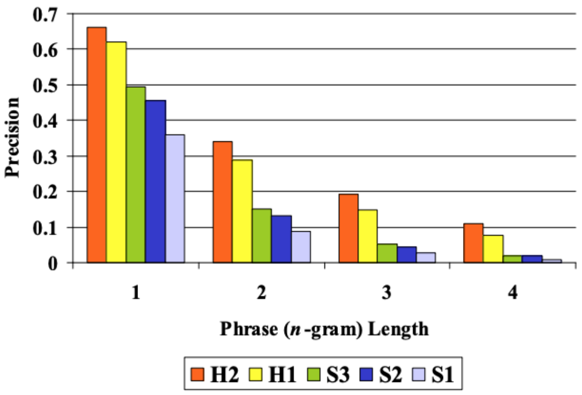
To verify that modified n-gram precision distinguishes between very good translations and bad translations, we computed the modified precision numbers on the output of a (good) human translator and a poor machine translation system using 5 reference translations for each of 136 source sentences.

Figure 1: Distinguishing Human from Machine



The strong signal differentiating human (high precision) from machine (low precision) is striking. The difference becomes stronger as we go from unigram precision to 4-gram precision. It appears that any single n-gram precision score can distinguish between a good translation and a bad translation. To be useful, the metric must also be reliable and must be able to distinguish between translations that do not differ so greatly in quality. Furthermore, it must also be able to distinguish between two human translations of differing quality. This latter requirement ensures the continued validity of the metric as MT approaches human translation quality. To this end, we obtained a human translation by someone lacking native proficiency in both the source (Chinese) and the target language (English). For comparison, we need the human translations of the same documents by a native English speaker. We also obtained machine translations by three commercial systems. These five "systems" — two humans and three machines — are scored against two reference professional human translations. The average modified n-gram precision results.

Figure 2: Machine and Human Translations



Each of these n-gram statistics implies the same ranking: H2 (Human-2) is better than H1 (Human1), and there is a big drop in quality between H1 and S3 (Machine/System-3). The S3 appears better than the S2 which in turn appears better than the S1. Remarkably, this is the same rank order assigned to these "systems" by human judges, as we discuss later. While there seems to be ample signal in any single n-gram precision, it is more robust to combine all these signals into a single number metric.

Combining the modified n-gram precisions

How should we combine the modified precisions for the various n-gram sizes? A weighted linear average of the modified precisions resulted in encouraging results for the 5 systems. However, as can be seen in Figure 2, the modified n-gram precision decays roughly exponentially with n: the modified unigram precision is much larger than the modified bigram precision which in turn is much bigger than the modified trigram precision. A reasonable averaging scheme must take this exponential decay into account; a weighted average of the logarithm of modified precisions satisfies this requirement. BLEU uses the average logarithm with uniform weights, which is equivalent to using the geometric mean of the modified n-gram precisions.5,6 Experimentally, we obtain the best correlation with monolingual human judgments using a maximum n-gram order of 4, although 3-grams and 5-grams give comparable results.

Sentence length

A candidate translation should not be too long or too short, and an evaluation metric should enforce this. To some extent, the n-gram precision already accomplishes this. N-gram precision penalizes spurious words in the candidate that do not appear in any of the reference translations. Additionally, modified precision is penalized if a word occurs more frequently in a candidate translation than its maximum reference count. This reward of using a word as many times as wanted and penalizes using a word more times than it occurs in any of the references. However, modified n-gram precision alone fails to enforce the proper translation length.

The trouble with recall

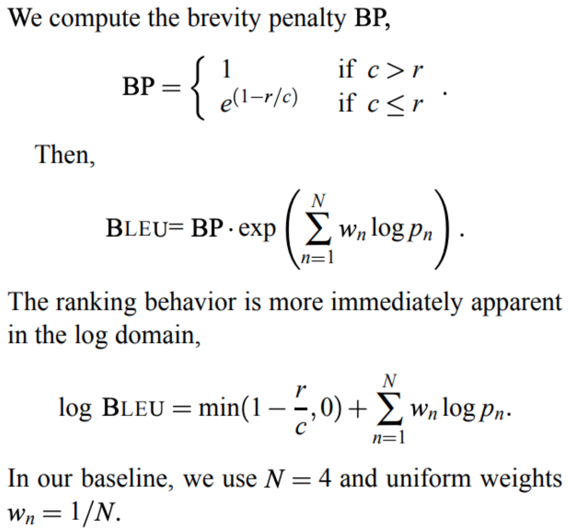
Traditionally, the precision has been combined with recall in order to overcome such length-related problems. However, this system considers multiple reference translations, each of which will use a different word choice to translate the same source word. Furthermore, a good candidate translation will use recall one of these possible choices, but not all. Indeed, recalling all choices leads to a bad translation.

Example 4:
Candidate 1: I always invariably perpetually do.
Candidate 2: I always do.
Reference 1: I always do.
Reference 2: I invariably do.
Reference 3: I perpetually do.

The first candidate recalls more words from the references, but is a bad method of translation than the second candidate. Thus, na¨ıve recall computed over the set of all reference words is not a good measure. Admittedly, one could align the reference translations to discover synonymous words and compute recall on concepts rather than words. But, given that reference translations vary in length and differ in word order and syntax, such a computation is complicated.

Sentence brevity penalty

Candidate translations longer than their references are already penalized by the modified n-gram precision measure: there is no need to penalize them again. Consequently, we introduce a multiplicative brevity penalty factor. With this brevity penalty in place, a high-scoring candidate translation must now match the reference translations in length, in word choice, and in word order. Note that neither this brevity penalty nor the modified n-gram precision length effect directly considers the source length; instead, they consider the range of reference translation lengths in the target language. We wish to make the brevity penalty 1.0 when the candidate's length is the same as any reference translation's length. For example, if there are three references with lengths 12, 15, and 17 words and the candidate translation is a terse 12 words, we want the brevity penalty to be 1. We call the closest reference sentence length the "best match length." One consideration remains: if we computed the brevity penalty sentence by sentence and averaged the penalties, then length deviations on short sentences would be punished harshly. Instead, we compute the brevity penalty over the entire corpus to allow some freedom at the sentence level. We first compute the test corpus' effective reference length, r, by summing the best match lengths for each candidate sentence in the corpus.



We compute the brevity penalty BP,

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} .$$

Then,

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

The ranking behavior is more immediately apparent in the log domain,

$$\log \mathrm{BLEU} = \min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n .$$

In our baseline, we use $N = 4$ and uniform weights $w_n = 1/N$.

IV. Conclusion

We believe that BLEU will accelerate the MT R&D cycle by allowing researchers to rapidly home in on effective modeling ideas. Our belief is reinforced by a recent statistical analysis of BLEU's correlation with human judgment for translation into English from four quite different languages (Arabic, Chinese, French, Spanish) representing 3 different language families (Papineni et al., 2002)! BLEU's strength is that it correlates highly with human judg8 Crossing this chasm for Chinese-English translation appears to be a significant challenge for the current state-of-the-art systems. ments by averaging out individual sentence judgment errors over a test corpus rather than attempting to divine the exact human judgment for every sentence: quantity leads to quality. Finally, since MT and summarization can both be viewed as natural language generation from a textual context, we believe BLEU could be adapted to evaluating summarization or similar NLG tasks.

V. REFERENCES

[1]. E. H. Hovy. 1999. Toward finely differentiated evaluation metrics for machine translation. In Proceedings of the Eagles Workshop on Standards and Evaluation, Pisa, Italy.

[2]. Kishore Papineni, Salim Roukos, Todd Ward, John Henderson, and Florence Reeder. 2002. Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results. In Proceedings of Human Language Technology 2002, San Diego, CA. To appear.

[3]. Florence Reeder. 2001. Additional mt-eval references. Technical report, International Standardsfor Language Engineering, Evaluation Working Group. http://isscowww.unige.ch/projects/isle/taxonomy2/

[4]. J.S. White and T. O'Connell. 1994. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In Proceedings of the First Conference of the Association for Machine Translation in the Americas, pages 193–205, Columbia, Maryland.

Cite this article as :

Anitha Nithya R, Saran A , Vinoth R, "Adaptive Resource Allocation and Provisioning in Multi-Service Cloud Environments ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 382-387, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT195253

Journal URL : http://ijsrcseit.com/CSEIT195253