Paper: *BLEU: a Method for Automatic Evaluation of Machine Translation*

Razwan Ahmed Tanvir

Date: 02/15/2022

Quote *"The main idea is to use a weighted average of variable length phrase matches against the reference translations"*

Overview Machine Translation is one of the most useful applications of natural language processing. Not only the development of the translation model, it is also important to ensure that the translated sentences are valid, convey the same meaning as the input and correlate with professional human translation. In this paper, the authors came up with a useful idea that can quantify the machine generated translation and provide a score to measure how close they are from a human translation. To calculate the score, the authors at first considered to measure the precision. However, the precision has a problem because the MT systems sometimes overgenerate some words which cause better precision measure for the worst translations. Therefore, the authors used a technique called modified n-gram precision where they measured the n-grams where n is 4 in the baseline model. Moreover, they clipped the maximum reference counts and the word count. In this way they ensured that the model only provide better precision measure if candidate sentence and reference sentences closely match. Furthermore, the authors introduced brevity penalty to ensure that the small candidate sentences should be penalized so that the candidate sentence does not get too long or too short. So, for this reason authors added a multiplicative brevity penalty factor.

BLEU provides a quick, inexpensive and language independent way of evaluating the quality of a machine translation. Moreover, this method highly correlates with the human professional translations. This research has shown several correlation illustrations to validate that BLEU can be used as a standard substitute if a human translator is not affordable.

Intellectual Merit *This research provides a useful solution to evaluating machine generated translations. BLEU scores can be used by the developers to judge their model's translation and using this method could save much time if they were to wait for a human evaluation of their translations. It has substantially advanced the natural language translation tasks. This research explored various areas of improving the BLEU scoring and introduced novel approach to address issues with scoring. The authors measured their success by two groups of translator one monolingual and one bilingual and showed that the proposed model outputs useful results.*

Broader Impact This research impacted the advancement of natural language translation tasks and provided a useful solution to evaluate machine generated translations. This method is widely used to evaluate the quality of machine generated translation. The researchers are previously from IBM. The lead author, Kishore Papineni is currently working as senior staff research scientist at Google.

Keywords Natural Language Processing, Machine Translation, Brevity Penalty, Human Evaluation

Discussion Questions

- BLEU matches the words in candidate sentence with the reference sentences. However, if the candidate sentence has synonym words then the generated sentence would have worst score. A newer model could be proposed which considers the synonyms in the candidate sentences.

- Authors mentioned that BLEU could be used to the summarizing of a text. However, BLEU only focus with the existing words in the candidate and reference sentences. I am not sure how well the summarizing task will perform.

Table 1: Grade deductions by section

Overview	Intellectual M.	B. Impact	Keywords	Questions	Is Online?