BLEU

1 Definition of BLEU

BLEU (Bilingual Evaluation Understudy) is an automatic evaluation indicator commonly used in machine translation tasks to measure the similarity between the generated translation text and the reference translation text. BLEU is designed to be as consistent as possible with human evaluation, but it is an automated evaluation metric that allows for faster evaluation on large amounts of generated text.

1.1 N-gram

N-gram is a statistical language model. This model can express a sentence as a sequence of n consecutive words, and use the collocation information between adjacent words in the context to calculate the probability of the sentence, so as to judge whether a sentence is smooth or not.

1.2 precision

BLEU is a precision based metric. For a given sentence, the reference translation sentence can be expressed as R, and the translation sentence predicted by the model can be expressed as G, then the formula of precision is:

$$P_n(G,R) = \frac{\sum_k min(C_k(G), max(C_k(R))}{\sum_k C_k(G)}$$

Among them, n indicates the number of possible n-grams. BLEU needs to calculate the precision of translation 1-gram, 2-gram, ..., N-gram, and generally N is set to 4.

Given n, if w_k represents the possible n-grams of the k group in the sentence, $C_k(G)$ represents the number of occurrences of w_k in the generated translation sentence, $C_k(R)$ represents the number of occurrences of w_k in the reference translation sentence.
The precision is to compare the degree of overlap between the n-grams in the generated translation sentence and the reference translation sentence. The higher the degree of overlap, the higher the quality of the translation.

1.3 Calculation method of BLEU

The formula for BLEU is:

$$BLEU(G,R) = BP(G,R)exp(\sum_{n=1}^{N} W_n log P_n(G,R))$$

Among them, the possible value of N is in the range of integers between 1 and 4.
W_n represents the weight. In actual calculation, for all n, the value of W_n is $1/N$.
$P_n(G,R)$ is precision.
$BP(G,R)$ is the penalty factor, the formula is as follows:

$$BP(G,R) = \begin{cases} 1 & lc > lr \\ exp(1 - lr/lc) & lc \leq lr \end{cases}$$

Among them, lc is the length of the generated translation sentence, and lr is the length of the shortest reference translation sentence.

2 Advantages and Disadvantages of BLEU

2.1 Advantages of BLEU

- BLEU is easy and fast to calculate

- In some cases, BLEU is closer to human evaluation
- BLEU has a wide range of applications, and BLEU can be applied to tasks such as machine translation, text summarization, and speech recognition generation

2.2 Disadvantages of BLEU

- BLEU does not consider semantics, sentence structure and expressive accuracy
- BLEU cannot handle rich sentences well, and the evaluation accuracy will be interfered by common words
- BLEU is biased towards shorter translation results, and the evaluation accuracy of shorter translation results is sometimes higher
- BLEU does not take into account the case of synonyms or similar expressions, which may lead to the rejection of reasonable translations

References

https://aclanthology.org/P02-1040.pdf