Université Paris Cité

How Reliable is MT Evaluation?

Guillaume Wisniewski
guillaume.wisniewski@u-paris.fr

October 2022

1 Re-evaluating the role of BLEU in machine translation research

BLEU is the de facto evaluation metric used in MT. It is, for instance, used by Meta AI in their "No Language Left Behind" (NLLB) initiative to claim that they achieve "an improvement of 44% BLEU relative to the previous state-of-the-art, laying important groundwork towards realizing a universal translation system." [3]. [1]

The goal of this lab is to understand the inner working of BLEU and show how using this value can easily result in wrong analysis : a thorough analysis of the result of NLLB [2] conclude that "many of Meta AI claims made in NLLB are : unfounded, misleading, and the result of a deeply flawed evaluation."

1. Implement the BLEU metric.
2. Using the WMT'15 test sets, [2] evaluate the performance of MBART and MARIANMT. These two models can be easily used with the HuggingFace API. [3] What can you conclude.

As noticed by [1], BLEU places no explicit constraints on the order that matching n-grams occur in. It is therefore possible, given a sentence, to generate many new sentences with at least as many n-gram matches by permuting words around *bigram mismatches*.

3. Explain on an example why such permutations will never decrease the BLEU score.
4. Given a sentence with n words and b bigram mismatches, how many sentences can you generate with this principle. Compute the number of sentences you will obtain on the WMT'15 test set.
5. Why this result question the use of BLEU as an evaluation metric.

1. [3] describes a wonderful work about collecting parallel corpora in 200+ languages and training a state-of-the-art MT systems on them. It is an extremely interesting reading that I warmly recommend to you.
2. https://statmt.org/wmt15/translation-task.html
3. https://huggingface.co/docs/transformers/model_doc/mbart and https://huggingface.co/docs/transformers/model_doc/marian

In addition to this flaw in its very conception, the practical implementation of BLEU poses many problems related in particular to the tokenization.

6. sacreBLEU [4] is an implementation of BLEU that aims to provide "hassle-free computation of shareable, comparable, and reproducible BLEU scores". Evaluate the two previous systems using sacreBLEU. What can you conclude ?
7. Using sacreBLEU and your own implementation of BLEU compute the score achieved :
   — when considering the "raw" translation hypotheses and references ;
   — when the translation hypotheses and references have been tokenized in subword units ; [5]
   — when the translation hypotheses and references have been tokenized in characters (this amounts to adding a space between each character of the references and of the translation hypotheses).

   How can you explain these results ?

Références

[1] Chris CALLISON-BURCH, Miles OSBORNE et Philipp KOEHN. "Re-evaluating the Role of Bleu in Machine Translation Research". In : *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy : Association for Computational Linguistics, avr. 2006, p. 249-256. URL : https://aclanthology.org/E06-1032.

[2] Benjamin MARIE. *Science Left Behind*. http://blog.benjaminmarie.com/science-left-behind.html. Accessed : 2022-10-19.

[3] NLLB TEAM et al. *No Language Left Behind : Scaling Human-Centered Machine Translation*. 2022. DOI : 10.48550/ARXIV.2207.04672. URL : https://arxiv.org/abs/2207.04672.

4. https ://github.com/mjpost/sacrebleu
5. You can use either one of the tokenizer provided by HuggingFace or train your own model, for instance on `Europarl` data.