BLEU score

LING83600

1 Introduction

The gold standard for measuring machine translation quality is the rating of candidate sentences by experienced translators. However, automated measures are necessary for rapid iterative development. BLEU (Papineni et al. 2002) is the best-known automatic measure of translation quality. BLEU and related measures are used to automatically *evaluate* machine translation (MT) systems, as well as an *objective* for training MT systems.

2 Intuitions

- Much like the models underlying modern statistical translation, BLEU and variants assume that the goal of the translation task is to mimic a corpus of human-generated translations. The quality of a computer-generated translation (the *candidate*) is determined by the degree to which it contains the same information found in one or more human-generated translations (the *reference*). Let C and R be sets of word tokens.[1] Then, the *precision* of candidate C with respect to reference R:

$$P(C \mid R) = \frac{|C \cap R|}{|C|} . \tag{1}$$

  Thus defined, precision corresponds to the probability that a token in the candidate is found in the reference.

- Precision is traditionally paired with *recall*, the probability that a token in the reference is found in the candidate. However, in machine translation, the reference may consist of many translations of the same sentence, and the human translators may have chosen different target words (e.g., *boredom*, *weariness*, *discontent*) to translate a single reference word (e.g., *Weltschmerz*). Recall would reward a pathological candidate that chose to use **all** the synonyms of a source word:

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

[1] Standard preprocessing steps include case-folding and removing punctuation.

For this reason, BLEU does **not** employ recall. To discourage this pathological strategy, "... a reference word should be considered exhausted after a matching candidate word is identified." (Papineni et al. 2002:312). This leads to the definition of *modified precision*.

- Another pathological strategy to maximize precision is to generate a candidate translation which contains only the most likely target word. To discourage this—which also produces poor-quality translations—BLEU incorporates a strict *brevity penalty*.
- Token-level precision does not effectively measure translation *fluency*; i.e., all permutations of a candidate have the same token precision. Therefore, BLEU combines precision scores from higher order *n*-grams.

3 Definition

Let the *clipped count* of some *n*-gram x

$$cc(x \mid C,\ R) = \min(\sum_{c\,\in\,C} [c = x], \sum_{r\,\in\,R} [r = x]) \ . \tag{2}$$

where the two right-hand terms denote the frequency of occurrence of x in C and in R, respectively. Then, let the *modified precision* of bag C with respect to R be

$$P_n(C \mid R) = \frac{1}{|C|} \sum_{x\,\in\,\{C\,\cup\,R\}} cc(x \mid C,\ R) \ . \tag{3}$$

Here, the subscript n refers to the n-gram order; e.g., P_2 refers to bigram modified precision. Modified precisions from different-order n-grams of orders 1, ..., N (often $N = 4$) are combined using the geometric mean:

$$\begin{aligned} \mu_N &= \left(\prod_{n=1}^{N} P_n(C \mid R) \right)^{\frac{1}{N}} & (4) \\ &= \exp \left(\frac{1}{N} \sum_{n=1}^{N} \ln P_n(C \mid R) \right) \ . & (5) \end{aligned}$$

Then, we apply a brevity penalty defined as

$$BP = \begin{cases} 1 & \text{if } |C| > |R| \\ \exp\left(1 - \frac{|R|}{|C|}\right) & \text{if } |C| \leq |R| \end{cases} \ . \tag{6}$$

That is, the brevity penalty is 1 when the candidate is longer than the reference, and exponentially increasing when the candidate is shorter. The brevity penalty is usually applied at "corpus level", rather than with respect to individual candidate-reference pairs. Putting it all together,

$$\text{BLEU} = BP \cdot \exp\left(\frac{1}{N}\sum_{n=1}^{N} \ln P_n(C \mid R)\right) . \tag{7}$$

4 Model tuning

Och (2003) proposes a discriminative approach to statistical MT, in which a linear translation model is trained to directly minimize an loss (i.e., error) function, such as (negative) BLEU score. This technique is known as *minimum error rate training* (MERT). MERT can be used to (re)rank an *n*-best list of translation candidates produced by a generative translation model (e.g., Shen and Joshi 2005), or it can be used to train an end-to-end discriminative translation model.

5 Critiques

BLEU and variants have suffered a great deal of abuse from MT researchers; a non-exhaustive list of critiques follows.

- Some specific details of BLEU are not sufficiently motivated; e.g.:
  - Why is the brevity penalty exponential?
  - Why are modified precisions across *n*-grams aggregated using a geometric mean? Doddington (2002), for example, argues for aggregation using an arithmetic mean.
- Though efforts were taken to discourage *precision-hacking*, there are still many varieties of pathological candidates receiving abnormally high BLEU scores despite being intuitively poor translations:[2]
  - Swapping *n*-grams between two non-matching regions has no effect on BLEU score (Callison-Burch et al. 2006).

    Reference: *The dog bit the mailman.*
    Candidate 1: *The dog snapped at the mailman.*
    Candidate 2: *The mailman nipped at the dog.*

[2] This is particularly concerning given that the use of BLEU in MERT training, which surely increases the chance that these pathological candidates will be produced by an MT system. On a related note, management theorists have long observed that employee performance metrics quickly become useless or even harmful:

> Robert Austin, in his book *Measuring and Managing Performance in Organizations*, says there are two phases when you introduce new performance metrics. At first, you actually get what you wanted, because nobody has figured out how to cheat. In the second phase, you actually get something worse, as everyone figures out the trick to maximizing the thing that you're measuring, even at the cost of ruining the company. (Spolsky 2006)

On the other hand, the intellectual challenge of maximizing a complex metric like BLEU—no matter how flawed—might still be an effective way to reveal engineering talent (and to allocate research dollars).

  - Not all *n*-grams are created equal, but high-frequency *n*-grams are given as much credit as low-frequency (highly informative) *n*-grams (Doddington 2002).

    Reference: *Orejuela appeared calm as he was led to the American plane which will take him to Miami, Florida.*
    Candidate 1: *Appeared calm when he was taken to the American plane, which will to Miami, Florida.*
    Candidate 2: *To Miami, Florida as he was led is to the was take him which will.*

    Noting this, NIST has adopted a variant of BLEU which weights the clipped *n*-gram counts according to the frequency of each *n*-gram in a reference corpus.
  - The use of higher-order *n*-grams privileges translation *fluency* over translation correctness (Zhang et al. 2004).

- Some studies have found that BLEU is poorly correlated with human ratings. For example, in the 2005 NIST evaluation, the MT system with the highest BLEU score was only rated 6th (!) by human judges.

While BLEU has proved **unreasonably effective**, particularly for rapid system development, human ratings of translation quality are still the gold standard.

6 Implementations

- `sacreBLEU` (Post 2018): `https://github.com/mjpost/sacrebleu`
- PyTorch version: `torchnlp.metrics.bleu`
- NLTK version: `nltk.translate.bleu_score`
- NIST version: `mteval-v13a.pl` (search for it online)

References

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 249–256.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *HLT '02: proceedings of the second international conference on Human Language Technology Research*, 138–145.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 160–167.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191.

Shen, Libin, and Aravind K. Joshi. 2005. Ranking and reranking with perceptron. *Machine Learning* 60:73–96.

Spolsky, Joel. 2006. The Econ 101 management method. Post on Joel on Software. URL `http://www.joelonsoftware.com/items/2006/08/09.html`.

Zhang, Ying, Stephan Vogel, and Alex Waibel. 2004. Interpreting BLEU/NIST scores: How much improvement do we need to have a better system? In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, 2051–2054.