Evaluating Text Output in NLP: BLEU at your own risk

Rachael Tatman (2019)

NLP reading club 2021/06/17

Key Points

- Metrics is very important to all machine learning tasks -- ideally it should be easy to calculate and one that we can optimise training against (if we're using it as loss function as well)
- BLEU score is a very common metric that people use for sequence to sequence task in NLP (e.g. machine translation),
- BLEU also comes with a lot of caveats -- it means quoting a BLEU score can be meaningless in certain cases
- Alternative metrics exists and before starting on any NLP problem, should carefully consider which metric makes most sense to optimise for
- Always try to involve actual humans in the final evaluation

The BLEU Score

BLEU: a Method for Automatic Evaluation of Machine Translation

What / Why

- automatic machine translation evaluation that is quick, inexpensive to evaluate (and therefore easy to incorporate into model training)
- Language-independent
- correlates highly with human evaluation (although this is debated)
- Used as a common metric in many research papers so easy to compare your results against existing baselines

How

- for each source input sentence, calculate the 1, 2, 3, 4-grams overlap between the output sentence and all the reference translations, then calculate the geometric mean of these
- Brevity penalty for outputs that are shorter than the shortest reference sentence
- Should only be applied over a corpus rather than individual sentences

Problems with BLEU

- doesn’t consider meaning
- doesn’t directly consider sentence structure
- doesn’t handle morphologically rich languages well
- doesn’t map well to human judgements

Alternative metrics

- NIST (weights n-grams based on their rareness)
- ROUGE (how many n-grams in the reference translation show up in the output)
- STM (compare parses and penalizes outputs with different syntactic structures.)
- METEOR (similar to BLEU, includes additional steps, like considering synonyms and comparing the stems of words, explicitly designed to use to compare sentences rather than corpora)
- TER (translation error rate) measures the number of edits needed to chance the original output translation into an acceptable human-level translation.
- hLEPOR is a metric designed to be better for morphologically complex languages. Considers things like part-of-speech (noun, verb, etc.) that can help capture syntactic information.

Alternative metrics

- RIBES, doesn’t rely on languages having the same qualities as English. It was designed to be more informative for Asian languages—like Japanese and Chinese—and doesn’t rely on word boundaries
- MEWR

Further resources

- The original BLEU paper
- Evaluating Text Output in NLP: BLEU at your own risk