Proposed Algorithm

Overview

- Neural machine translation models rely on Beam Search to produce output words.
- Beam search has to sacrifice an existing decoding path to explore a new candidate.
- As the output sentence is decoded in left-to-right order, a discarded path cannot be recovered again. We call it monotonic constraint.
- In this work, we extend beam search to allow revisiting a discarded decoding path in the past.
- The algorithm is implemented with a single priority queue, which we refer to as single-queue decoding (SQD).

Score Function

- The score function adds two auxiliary penalties to log likelihood:
  \[
  \text{score}(y) = \frac{1}{|y|} \log p(y|X) + \alpha \text{PG}(y) + \beta \text{LMP}(y)
  \]

- The progress penalty (PG) encourages the algorithm to select longer candidates.
- The length matching penalty (LMP) punishes the hypotheses that tend to produce translations much longer or shorter than expectation.

Quantitative Results

- SQD algorithm generally improves the translation results.
- Length matching penalty (LMP) is only effective when a small beam size is used.
- Without PG, the translation performance hurts (not shown in the table).

 Insights

- SQD algorithm generally improves the translation results.
- Length matching penalty (LMP) is only effective when a small beam size is used.
- Without PG, the translation performance hurts (not shown in the table).

Have Fun

- Contact me: shu@nlab.ci.i.u-tokyo.ac.jp
- Deep Learning Monitor: https://deeplearn.org