Beam Search Strategies for Neural Machine Translation

1 Main Idea

The standard beam search strategy for Neural Machine Translation (NMT) is for the decoder to predict the target sequence word-by-word and maintain a fixed amount of potential word candidates to predict at each step.

   The drawbacks of this approach are that it is less adaptive, because:

- Some candidates might not be as good as the current best

- Good candidates might not be considered because they missed out on the threshold for candidate inclusion marginally. Addressing this by naively increasing beam search size will result in slowing down the decoder.

$$beam\_size \propto model\_accuracy$$

$$beam\_size \propto \frac{1}{model\_performance}$$

The paper proposes a more flexible decoder strategy, by pruning the search graph, and reducing the number of candidates with the same partial hypothesis (shared past).

2 Method

-

3 Observations

-

4 Open Questions

-

References