What is bandit machine translation?

“Guten Morgen” → “Good Morning”

Bandit Feedback

The UMD Neural Bandit MT Systems

● Adapts a translation system to a new domain;
● Receives a German sentence to translate, produces an English sentence, and only gets a scalar score as feedback;
● Extends a standard Neural MT system in two ways:
  a. Robust reinforcement learning techniques to learn effectively from the bandit feedback;
  b. Domain adaptation using data selection from a large corpus of parallel data.

Neural MT Architecture

Schematic view of the OpenNMT neural machine translation model. (Credit: Klein et al., 2017)

Reinforcement Learning

● The NMT model can be viewed as a Markov decision process operating on a continuous state space:
  a. States: the hidden vectors hdec generated by the decoder.
  b. Action space: the target language’s vocabulary.
● Advantage Actor-Critic (A2C) algorithm:

Receive Source Sentence

Update NMT and Critic Models

Sample Translation

Receive Reward

Domain Adaptation

● Cross-Entropy Difference: uses the cross-entropy difference \( H_I(s) - H_O(s) \) for scoring a given sentence \( s \), based on an in-domain language model \( LM_I \) and an out-of-domain language model \( LM_O \).
● Cross-Entropy:

\[
H(W) = -\frac{1}{n} \sum_{i=1}^{n} \log P_{LM}(w_i|w_1, \ldots, w_{i-1})
\]

Experiments

The results support the following conclusions:

1. Domain adaptation: data selection for domain adaptation alone improves translation quality by about 1.5 BLEU points.

2. Reinforcement Learning: on top of the domain adaptation, reinforcement learning (which requires exploration) leads to an initial degradation of about 3 BLEU points, which is recovered (on development data) after approximately 40k sentences of bandit feedback.

Comparing sampling, greedy decoding, and the A2C algorithm on the development data.

References