Promoting the Knowledge of Source Syntax in Transformer NMT

Thuong-Hai Pham, Dominik Macháček, Ondřej Bojar
{pham,machacek,bojar}@ufal.mff.cuni.cz
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Prague, Czech Republic

Motivation
- if you learn syntax, you can translate better
- use expert linguistic knowledge (linguistic annotations) to enhance unsupervised NMT
- apply it to high-resource language pairs, not only low-resource

Simple Alternating Multi-Task
- primary task: MT
- secondary task: parsing of source side (several variants)
- and for comparison, dummy non-linguistic tasks:
  Source words: We have a three-colored cat.
  CountSrcWords: 6
  EnumSrcWords: W W W W W
  CopySrc: We have a three-colored cat.

- single encoder-decoder, the task is determined by a special token on source
- the trainer alternates between the tasks, 1:1 ratio, shuffled

Results:
- the model learns the secondary tasks very fast and well
- linguistic secondary task helps (MT quality better than the dummy referential task)
- cost for multi-tasking probably higher than benefit, the overall MT quality drops

Dependency Interpretation of Self-Attention
- interpret self-attention weights as probabilities of being a father in dependency tree
- modify the training objective to learn both MT and parsing of source

Results:
- the model learns parsing very fast and well
  - Good news: promoting syntax in one self-attention head improves the MT quality
  - Bad news: promoting a dummy "diagonal parse" improves MT the same way

Experiment details: CS->EN, subset of CzEng 1.7, 5.2M sent. pairs, dev and test set from CzEng.
Tensor2Tensor 1.5.6, word level (no subword segmentation).

The research was partially supported by the grants 19-26934X (NEUREM3) of the Czech Science Foundation and H2020-ICT-2018-2-825460 (ELITR) of the EU.