An Analysis of Encoder Representations in Transformer-Based Machine Translation

Presenter: Xing Wang

- Experiment setup:

	#Training sentences
English → Czech	51.391.404
English → German	25.746.259
English → Estonian	1.064.658
English → Finnish	2.986.131
English → Russian	9.140.469
English → Turkish	205.579
English → Chinese	23.861.542

Table 1: Number of training instances used to train each system.

	newstest 2017	newstest 2018
English → Czech	18.11	17.36
English → German	23.37	34.46
English → Estonian	–	13.05
English → Finnish	15.06	10.32
English → Russian	21.30	18.96
English → Turkish	6.93	6.22
English → Chinese	23.10	23.75

Table 2: BLEU score for the newstest2017 and newstest2018 test data.

- Experiment setup:
  - use the full word forms, allowing us to evaluate and compare the internal representation on standard sequence labeling benchmarks tagged with gold labels on the full word forms
  - use a large vocabulary of 100K words per language

- Encoder Evaluation: Visualization
  - Experiment: focus only on attention weights with high scores that are visually interpretable
  - Findings:
    - 1) discovered four different patterns: paying attention to the word itself, to the previous and next word and to the end of the sentence
    - 2) long dependencies between words on higher layers
    - 3) whereas it tends to focus on local dependencies in lower layers

- Encoder Evaluation: Inducing Tree Structure
  - Experiment: evaluated the induced trees on the English PUD treebank from the CoNLL 2017 Shared Task http://aclweb.org/anthology/W18-5431
  - Findings:
    - 1) the models trained with larger dataset are able to induce better syntactic relationships
    - 2) while among high resource languages all models are in the same ballpark, without any specific correlation with BLEU score

- Encoder Evaluation: Probing Sequence Labeling Tasks
  - Experiment: evaluated the encoder representation through four different sequence labeling tasks: Part-ofSpeech (PoS) tagging, Chunking, Named Entity Recognition (NER) and Semantic tagging (SEM).

  - Findings:
    - 1) the syntax information, i.e., the PoS task, is encoded mostly in the first 3 layers
    - 2) while moving towards more semantic tasks, as NER and SEM we can see that in general the decoder needs more encoder layers to achieve better results
    - 3) the models encode the information about the sentence length in the first three layers, and then the information starts to vanish with an increase of the error rate

- Encoder Evaluation: Transfer learning
  - Experiment: used the encoder weights from one high resource language, i.e., English-German, to train a Transformer system for our low resource language pair, English-Turkish.

  - Findings: starting with a better encoder representation, taken from a high resource language pair, and then fine tuning the parameters on the low resource language achieves the best result, matching and corroborating previous findings on recurrent networks

- Overall
  - each layer has at least one attention head that encodes a significant amount of syntactic dependencies
  - lower layers tend to encode more syntactic information, whereas upper layers move towards semantic tasks
  - the information about the length of the input sentence starts to vanish after the third layer
  - attention can be used to transfer knowledge between high- and low-resource languages