BERTology for Machine Translation: What BERT Knows about Linguistic Difficulties for Translation

Yuqian Dai, Marc de Kamps, Serge Sharoff
University of Leeds

1. Contributions

- What does BERT know about syntactic dependencies before and after fine-tuning for Neural Machine Translation (NMT)?
- We use probing to detect changes in syntactic knowledge before and after fine-tuning. Prediction of some syntactic dependencies is negatively affected by fine-tuning.
- Which syntactic dependencies lead to a drop in translation quality across different languages?
- We recognise certain types of syntactic dependencies linked to low-quality machine translations.
- When our BERT-based NMT engine produces a non-sensical translation, does it "understand" the syntactic structure in this sentence?

2. Construction of NMT Engines

- Approximately 1.2 million (1.2M) parallel sentence pairs for Chinese, Russian, and German. BERT is fine-tuned as the encoder in the NMT engine.
- The BLEU scores were as follows:
  - Language Dataset BLEU
    - Zh → En UNPC 56.34
    - Ru → En UNPC 55.85
    - De → En Europarl 38.06

3. Probing task

- A syntactic dependency indicates the relationship between two words. BERT needs to predict the syntactic dependency of the current word without specifying the target word in the sentence.
- We test for theses three languages on the PUD and GSD corpora from the Universal Dependencies.

4. Translation Quality Estimation

- Using Quality Estimation (QE) to evaluate translation quality without reference translations. Choosing highest 20% score as high-quality translations and lowest 20% score as low-quality translations.
- Using chi-square test to detect whether there is a correlation between syntactic dependencies and the quality of machine translation.
- The null hypothesis ($H_0$) that translation quality and syntactic dependencies are unrelated is not valid. Instead, the alternative hypothesis ($H_1$) is accepted that translation quality is associated with syntactic dependency.
- Using Quality Estimation (QE) to evaluate translation quality without reference translations. Choosing highest 20% score as high-quality translations and lowest 20% score as low-quality translations.
- Using chi-square test to detect which syntactic dependencies are associated with low-quality translations.
- Syntactic dependencies ordered according to $\chi^2$ in the German examples.

5. Conclusions

- F1-score for detecting most Universal Dependencies by BERT decreases after NMT fine-tuning.
- Translation quality is associated with syntactic dependencies.
- The F1-score for those dependencies causing low-quality translations are relatively low, and often they decrease with training.
- BERT has a different syntactic dependency performance as a standalone monad than an NMT engine.