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

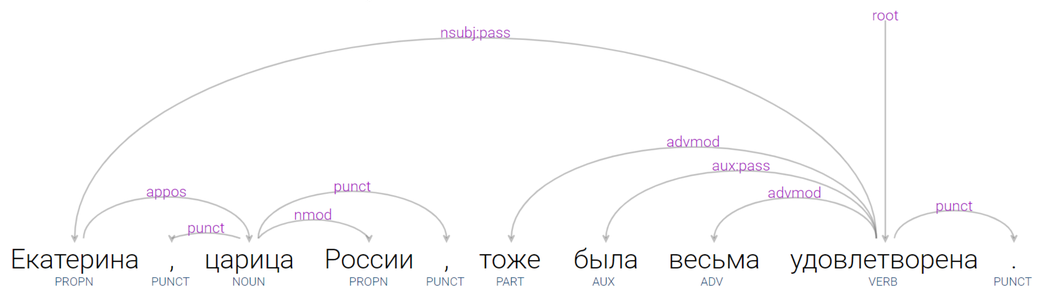
Yuqian Dai, Marc de Kamps, Serge Sharoff

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?



MT herat , her country ' s russian federation , was also very pleased .
Reference Catherine of Russia was also very satisfied.

"`appos`" linking *Catherine* and *the Queen of Russia* in the source sentence is a possible cause of problems in rendering the first half of this translation.

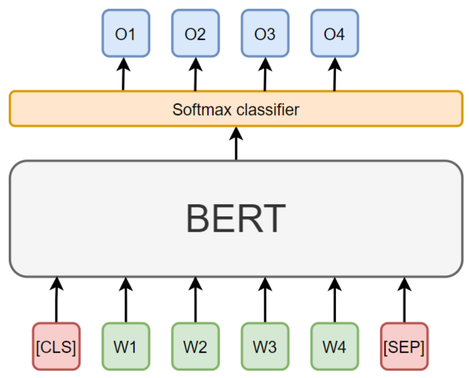
2. Construction of NMT Engines

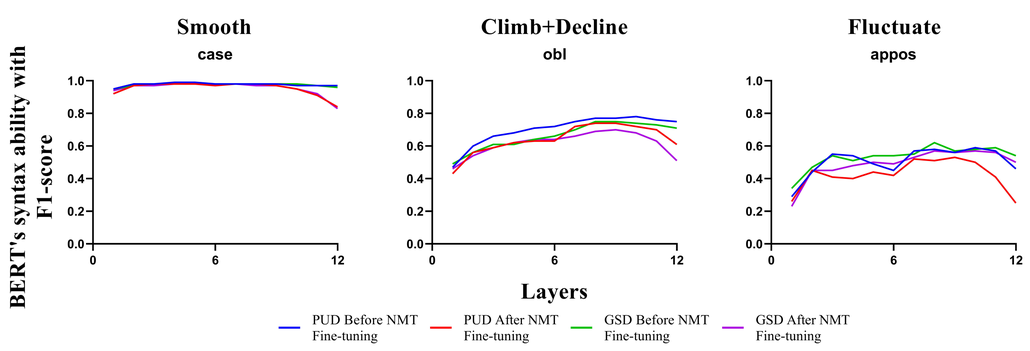
- Approximately 1.2 million (1.2M) parallel sentence pairs for Chinese, Russian and German. BERT is fine-tuned as the encoder in the NMT engine.

Language	Dataset	BLEU
Zh $\rightarrow$ En	UNPC	56.34
Ru $\rightarrow$ En	UNPC	55.85
De $\rightarrow$ 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.





- Most Universal Dependencies by BERT decrease after NMT fine-tuning.
- Syntactic dependencies are related to BERT's layers but are more likely determined by the working mechanism of BERT itself during the pre-training.
- BERT has already formed a syntactic pattern during the pre-training, and the fine-tuning of machine translation only changes the performance of the downstream task.

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.

	Zh$\rightarrow$En	Ru$\rightarrow$En	De$\rightarrow$En
High-quality range	0.808-0.891	0.845-0.917	0.822-0.896
Low-quality range	0.061-0.519	0.325-0.742	0.226-0.533

- 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

Languages	Dependencies	df	p-value	Test statistic	χ^2
Zh	32	31		43.77	171.4
Ru	29	28	0.05	41.34	154.9
De	30	29		42.56	182

- Using chi-square test to detect which syntactic dependencies are associated with low-quality translations.
- Syntactic dependencies ordered according to χ^2 in the German examples.

Quality Estimation Examples for German

	Dependency	Quality			F1-score	
		High	Low	χ^2	Layer-6	Layer-12
De	flat	0	9	-	0.25	0
	appos	10	96	739.6	0.42	0.25
	flat:name	6	61	504.1	0.52	0.39
	compound	25	72	88.3	0.47	0.51
	obl	212	327	62.3	0.63	0.61
	compound:prt	10	34	57.6	0.92	0.48
	case	324	459	56.25	0.97	0.84
	obl:tmod	14	34	28.5	0.55	0.62
	nmod:poss	39	67	20.1	0.96	0.85
	nsubj	241	308	18.6	0.73	0.68
	advcl	27	47	14.8	0.34	0.36

- Worst relations for all languages specific syntactic dependencies are likely to be a significant cause of low-quality translations.

Dependency Zh/Ru/De	Quality χ^2	F1-score Layer-6	F1-score Layer-12
appos	396.9/48.3/739.6	0.4/0.36/0.42	0.48/0.47/0.25
flat	242/900/-	0.7/0.22/0.25	0.74/0.18/0
flat:name	2704/168.7/504.1	0.68/0.82/0.52	0.78/0.86/0.39
obl	24.6/66.1/62.3	0.39/0.71/0.63	0.31/ 0.66/0.61
case	24.2/40/56.25	0.82/0.98/0.97	0.76/0.91/0.84

5. Conclusions

1. F1-score for detecting most Universal Dependencies by BERT decreases after NMT fine-tuning.
2. Translation quality is associated with syntactic dependencies.
3. Common low-quality translation problems occur in the context of specific syntactic dependencies: "appos", "flat", "flat:name", "obl", "case".
4. The F1-score for those dependencies causing low-quality translations are relatively low, and often they decrease with training.
5. BERT has a different syntactic dependency performance as a standalone monad than an NMT engine.

LREC, 20-24 June 2022, Marseille, France.