Promoting the Knowledge of Source Syntax in Transformer NMT

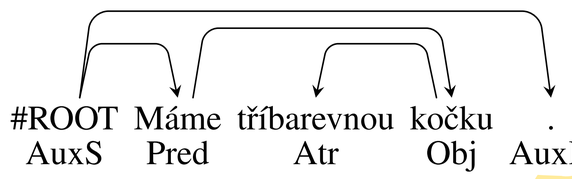
Thuong-Hai Pham, Dominik Macháček, Ondřej Bojar
{pham,machacek,bojar}@ufal.mff.cuni.cz
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Prague, Czech Republic

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)



Task ID tokens

MT source (cs)		Máme tříbarevnou kočku . #Translate
MT target (en)		We have a three-colored cat .
DepHeads	src	Máme tříbarevnou kočku . #DepHeads
	tgt	#ROOT kočku Máme #ROOT
DepLabels	src	Máme tříbarevnou kočku . #DepLabels
	tgt	#Pred #Atr #Obj #AuxK
DepHeads +DepLabels	src	Máme tříbarevnou kočku . #DHeadsLab
	tgt	#ROOT #Pred kočku #Atr Máme #Obj #ROOT #AuxK

- and for comparison, dummy non-linguistic tasks:

Source words	We have a three-colored cat .
CountSrcWords	6
EnumSrcWords	W 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

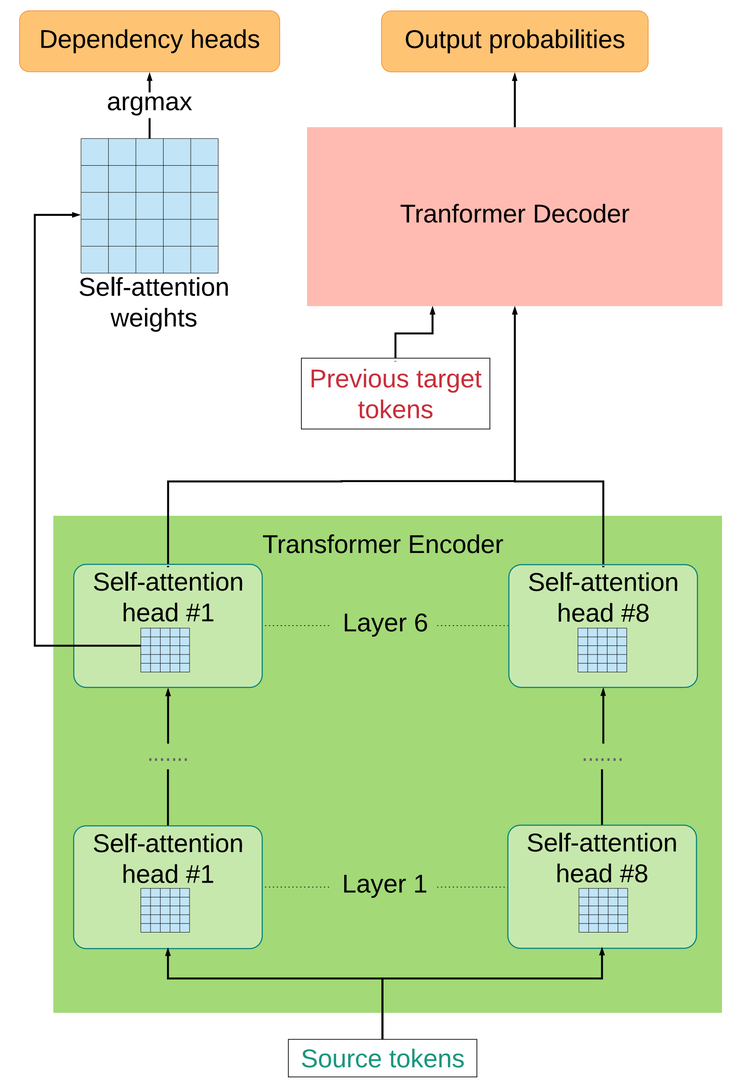
Model	BLEU dev	BLEU test
MT Baseline	**17.90**	**19.74**
MT TaskID	16.53	18.20
MT+DepLabels	16.52	17.87
MT+DepDheads	16.36	17.62
MT+CountSrcWords	15.70	17.51
MT+CopySrc	14.73	16.07
MT+DepHeads+DepLabels	13.62	15.45
MT+EnumSrcWords	12.16	14.04

Model	de2cs UAS	de2cs label acc	cs2en UAS	cs2en label acc
referential parser	**62.87**	*73.62*	**86.28**	83.38
MT+DepLabels	–	**75.40**	–	**85.01**
MT+DepHeads	*62.15*	–	*80.35*	–
MT+DepHeads+ DepLabels	54.98	68.44	80.01	*83.99*

DE->CS, Europarl+OpenSubtitles, 8.8M sentence pairs,
test: news2018, dev: news2011. Results at 600k training steps.

Dependency Interpretation of Self-Attention

Our extension of the basic Transformer

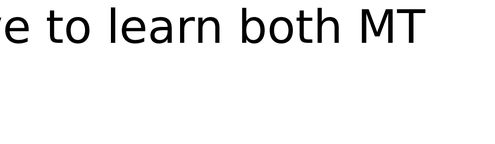


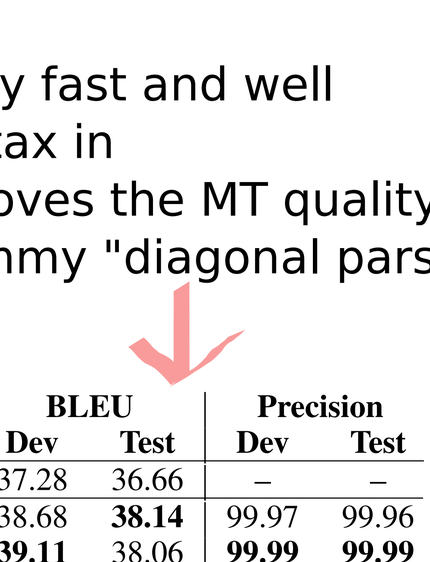
- interpret self-attention weights as probabilities of being a father in depedency tree
- modify the training objective to learn both MT and parsing of source

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

	BLEU Dev	BLEU Test	UAS Dev	UAS Test	BLEU Dev	BLEU Test	Precision Dev	Precision Test
TransformerBase	37.28	36.66	–	–	37.28	36.66	–	–
Parse from layer 0	36.95	36.60	81.39	82.85	38.68	**38.14**	99.97	99.96
Parse from layer 1	**38.51**	**38.01**	90.17	90.78	**39.11**	38.06	**99.99**	**99.99**
Parse from layer 2	38.50	37.87	91.31	91.18	37.85	37.85	99.98	99.98
Parse from layer 3	38.37	37.67	91.43	91.43	37.93	37.70	99.97	99.98
Parse from layer 4	37.86	37.60	**91.65**	**91.56**	37.68	37.47	99.98	99.96
Parse from layer 5	37.63	37.67	91.44	91.46	37.53	37.54	99.96	99.95

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.