MacNet: Transferring Knowledge from Machine Comprehension to Sequence-to-sequence Models

Boyuan Pan, Yazheng Yang, Hao Li, Zhou Zhao, Yueting Zhuang, Deng Cai, Xiaofei He.
Zhejiang University

Introduction

We propose MacNet: a novel encoder-decoder supplementary architecture to the widely used attention-based sequence-to-sequence models, which transfers knowledge learned from machine comprehension to the sequence-to-sequence tasks to deepen the understanding of the text.

Machine Comprehension (MC):

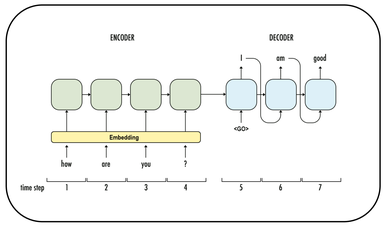
➢It requires to provide an answer (usually a subspan in the passage) given a passage and a question:

Context: The Alpine Rhine is part of the Rhine, a famous European river. The Alpine Rhine begins in the most western part of the Swiss canton of Graubünden, and later forms the border between Switzerland to the West and **Liechtenstein** and later Austria to the East. On the other hand, the Danube separates Romania and Bulgaria.

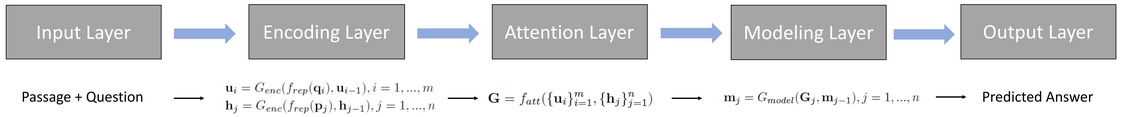
Question: What is the other country the Rhine separates Switzerland to?

Answer: **Liechtenstein**

Sequence-to-sequence Model:



Machine Comprehension



MacNet Architecture

Encoder: Original encoder + transferred encoder.

$$\bar{\mathbf{h}}_s = F_{enc}(\mathrm{Emb}(x_s), \bar{\mathbf{h}}_{s-1})$$

$$\tilde{\mathbf{e}}_s = G_{enc}(\mathrm{Emb}(x_s), \tilde{\mathbf{e}}_{s-1})$$

$$\bar{\mathbf{h}}_s = F_{int}([\bar{\mathbf{h}}_s; \tilde{\mathbf{e}}_s], \bar{\mathbf{h}}_{s-1})$$

Decoder & Attention Mechanism: Additionally send the attention vector into the modeling layer of the pre-trained MC model.

$$\alpha_{ts} = \frac{\exp(\mathrm{score}(\bar{\mathbf{h}}_s, \bar{\mathbf{h}}_t))}{\sum_{s'=1}^{T_x} \exp(\mathrm{score}(\bar{\mathbf{h}}_{s'}, \bar{\mathbf{h}}_t))}$$

$$\mathbf{c}_t = \sum_s \alpha_{ts} \bar{\mathbf{h}}_s$$

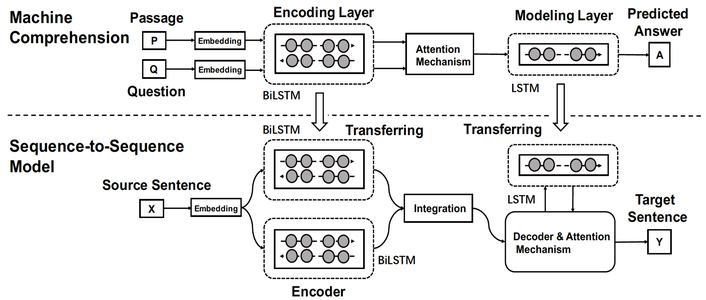
$$\mathbf{a}_t = g_a(\mathbf{c}_t, \bar{\mathbf{h}}_t) = \tanh(\mathbf{W}_a[\mathbf{c}_t; \bar{\mathbf{h}}_t] + \mathbf{b}_a)$$

$$P(y_t|y_{<t}, x) \propto \mathrm{softmax}(\mathbf{W}_p \mathbf{a}_t + \mathbf{b}_p)$$

Knowledge Transfer

$$\mathbf{r}_t = G_{model}(\mathbf{a}_t, \mathbf{r}_{t-1})$$

$$P(y_t|y_{<t}, x) \propto \mathrm{softmax}(\mathbf{W}_p \mathbf{a}_t + \mathbf{W}_q \mathbf{r}_t + \mathbf{b}_p)$$



Training: Add a modulating factor to the cross entropy loss.

$$\Theta^* = \arg\max_{\Theta} \sum_{(x,y)\in D} P(y|x;\Theta)$$

$$= \arg\max_{\Theta} \sum_{(x,y)\in D} \sum_{t=1}^{T_y} \log P(y_t|y_{<t}, x; \Theta)$$

Simplifying $P(y_t|y_{<t}, x; \Theta)$ as p_t

$$\Theta^* = \arg\max_{\Theta} \sum_{(x,y)\in D} \sum_{t=1}^{T_y} (1-p_t)^\gamma \log(p_t)$$

Experiments

➢ BLEU scores on official NMT test sets (WMT English-German for newtest2014 and newtest2015).

NMT Systems	WMT14		WMT15	
	En→De	De→En	En→De	De→En
Baseline	22.1	26.0	24.5	27.5
Baseline + Encoding Layer	23.2	27.0	25.3	28.3
Baseline + Modeling Layer	22.4	26.4	24.8	27.8
Baseline + Encoding Layer + Modeling Layer	23.4	27.3	25.6	28.5
Baseline + Random Initalized Framework	21.6	25.6	24.2	27.0
Baseline + MacNet	**24.2**	**28.1**	**26.3**	**29.4**

➢ ROUGE F1 evaluation results on the CNN/Daily Mail test set and the English Gigaword test set.

Summarization Models	CNN/Daily Mail			Gigaword		
	RG-1	RG-2	RG-L	RG-1	RG-2	RG-L
words-lvt5k[Nallapati et al., 2016]	35.46†	13.30†	32.65†	35.30†	16.64†	32.62†
SummaRuNNer[Nallapati et al., 2017]	39.60†	16.20†	35.30†	–	–	–
ConvS2S[Gehring et al., 2017]	–	–	–	35.88†	17.48†	33.29†
SEASS[Zhou et al., 2017]	–	–	–	36.15†	17.54†	33.63†
RL with intra-attn[Paulus et al., 2017]	**41.16**†	15.75†	**39.08**†	–	–	–
Pointer-Generator[See et al., 2017]	39.69	17.26	36.38	36.44	17.26	33.92
Pointer-Generator + Encoding Layer	40.38	17.75	37.24	37.30	17.83	34.41
Pointer-Generator + Modeing Layer	39.92	17.58	36.65	36.85	17.45	34.12
Pointer-Generator + MacNet	40.87	**18.02**	37.54	**37.97**	**18.16**	**34.93**

➢ Performance with different pre-trained machine comprehension models for our NMT model on De-En of WMT 14.

MC Attention	EM	BLEU
Context to Query Attention	63.3	25.1
Query to Context Attention	56.9	25.3
BiDAF	67.1	27.5
BiDAF + Self-Attention	68.2	27.4
BiDAF + Memory Network	68.5	27.6

➢ An example of summary on English Gigaword.

Article: Israeli warplanes raided Hezbollah targets in south Lebanon after guerrillas killed two militiamen and wounded seven other troops on Wednesday, police said.
Reference: Israeli warplanes raid south Lebanon.
PG + MacNet: Israeli warplanes attack Hezbollah targets in south Lebanon.
PG: Hezbollah targets Hezbollah targets in south Lebanon.

Article: The dollar racked up some clear gains on Wednesday on the London forex market as operators waited for the outcome of talks between the White House and Congress on raising the national debt ceiling and on cutting the American budget deficit.
Reference: Dollar gains as market eyes US debt and budget talks.
PG + MacNet: : Dollar racked up some clear gains.
PG: London forex market racked gains.

Conclusion

➢We improve the sequence-to-sequence model via transferring knowledge of several neural network layers from another supervised task.

➢We conduct extensive experiments on two typical seq2seq tasks to show that our method achieves significant improvement on the baseline model.