Multi-style Generative Reading Comprehension

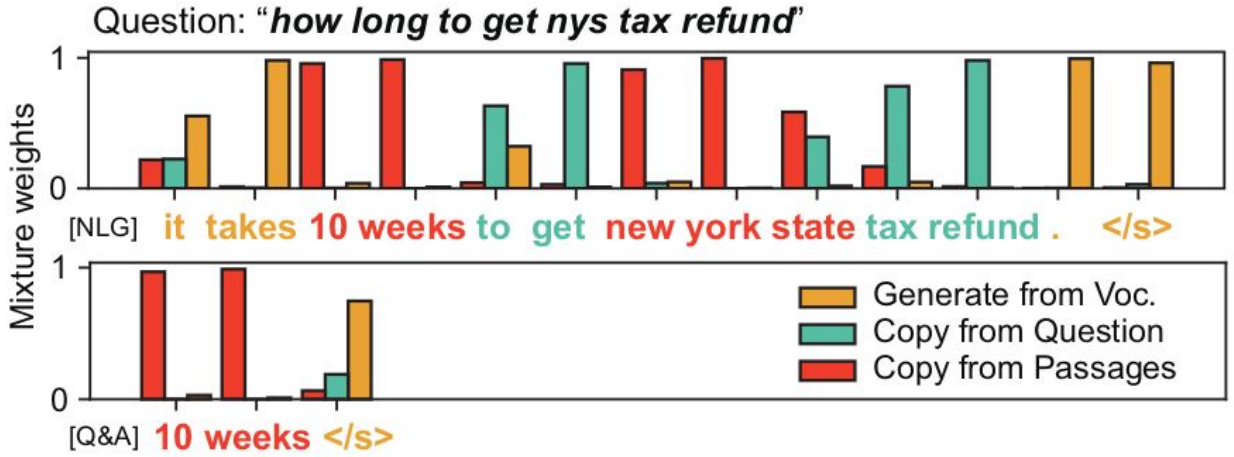
Nishida et al. 2019

Generative Reading Comprehension

From a **Q**uestion and a textual **C**ontext, *generate* an **A**nswer (MS MARCO)

They add a switch to determine the *style* of the answer: concise or well-formed

Use of Pointer-Generator Networks to copy or generate each word



Masque model

1) Question-passages reader

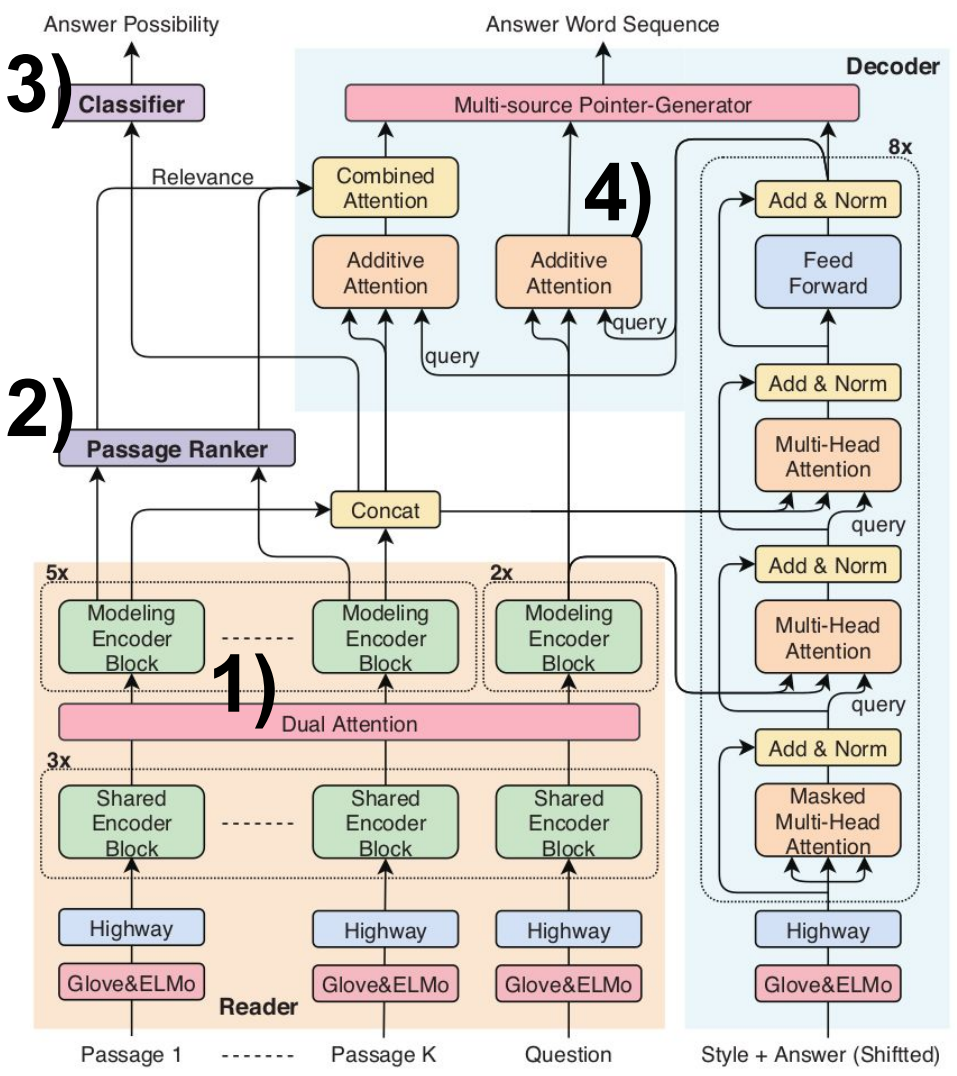
   → Compute M^q and M^{pk}

2) Passage ranker

3) Answer possibility classifier

4) Answer sentence decoder

Overall, *lots* of Transformers



Passage ranker

For each passage p_k use the representation **of the first word** M_1^{pk} to compute a **relevance score** B^{pk} $B^{pk} = \text{sigmoid}(w^{r\,T} M_1^{pk})$

Answer possibility classifier

Concatenate the representation of each passage and compute an **answerability probability** $P(a) = \text{sigmoid}(w^{c\,T} [M_1^{p1} ;...; M_1^{pk}])$

Answer sentence decoder

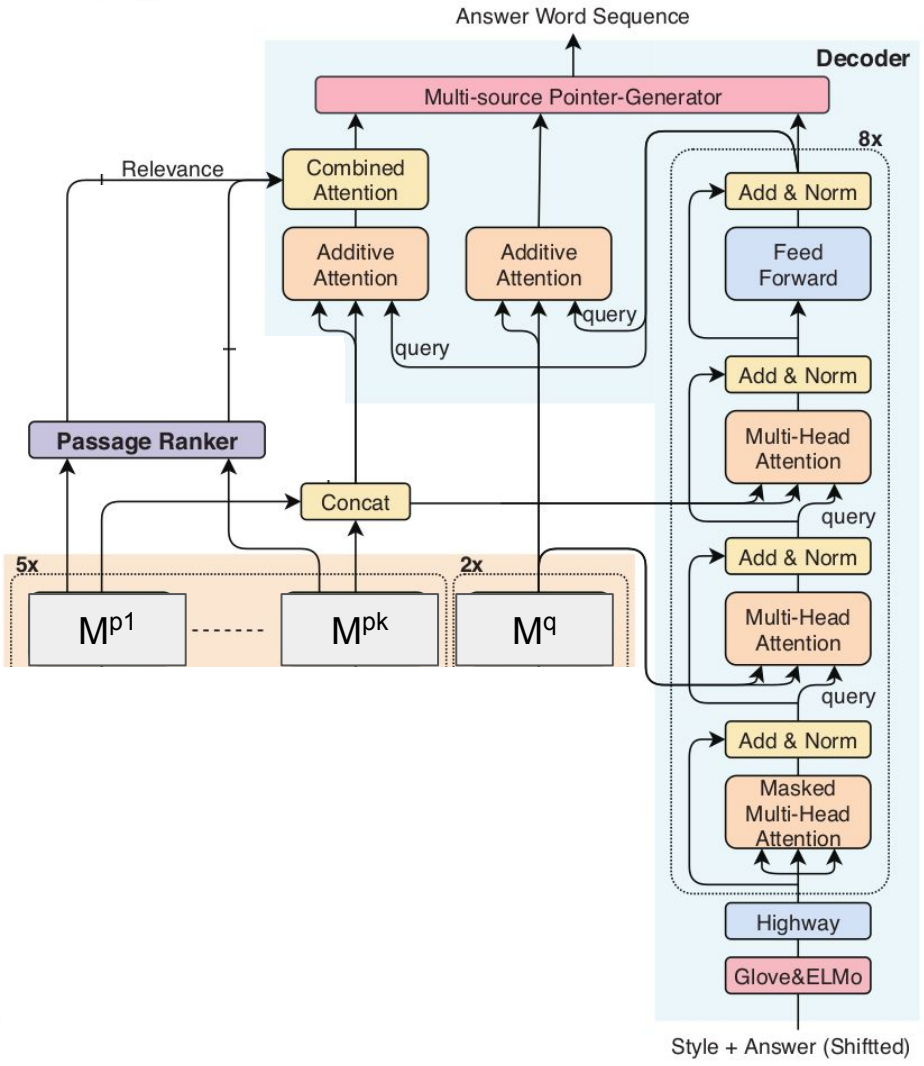
Iteratively generates each word via auto-regressive mechanism

An *artificial token* is used to control the style

Input: previous token

M^q and $[M_1^{p1};...;M_1^{pk}]$ are successively used

Words can be copied or generated from an *extended vocabulary*



Multi-task learning

$$L = L_{dec} + g_{rank}L_{rank} + g_{cls}L_{cls}$$

- L_{dec}: NLL of the whole target sentence
- L_{rank} and L_{cls}: binary cross entropy between the true and predicted *rank* and *answerability*

Results

SOTA on MS MARCO for both NLG and Q&A

Model	NLG		Q&A	
	Rouge-L	Bleu-1	Rouge-L	Bleu-1
BiDAF (2017)	16.91	9.30	23.96	10.64
Deep Cascade QA (2018)	35.14	37.35	52.01	**54.64**
S-Net (2018)[1]	45.04	40.62	44.96	46.36
VNET (2018)	48.37	46.75	51.63	54.37
Masque (NLG; single)	49.19	49.63	48.42	48.68
Masque (Q&A; single)	25.66	36.62	50.93	42.37
Masque (NLG; ensemble)	**49.61**	**50.13**	48.92	48.75
Masque (Q&A; ensemble)	28.53	39.87	**52.20**	43.77
Human Performance	63.21	53.03	53.87	48.50

Results

SOTA on MS MARCO for both NLG and Q&A

Trained on 8 P100 during 6 days...

Model	NLG		Q&A	
	Rouge-L	Bleu-1	Rouge-L	Bleu-1
BiDAF (2017)	16.91	9.30	23.96	10.64
Deep Cascade QA (2018)	35.14	37.35	52.01	**54.64**
S-Net (2018)[1]	45.04	40.62	44.96	46.36
VNET (2018)	48.37	46.75	51.63	54.37
Masque (NLG; single)	49.19	49.63	48.42	48.68
Masque (Q&A; single)	25.66	36.62	50.93	42.37
Masque (NLG; ensemble)	**49.61**	**50.13**	48.92	48.75
Masque (Q&A; ensemble)	28.53	39.87	**52.20**	43.77
Human Performance	63.21	53.03	53.87	48.50

Ablation

Model	train	Rouge-L	Bleu-1
Masque (NLG style; single)	ALL	**69.77**	**65.56**
w/o multi-style learning (§3.4.2)	WFA	68.20	63.95
↪ w/o Transformer (§3.1.2, §3.4.2)	WFA	67.13	62.96
w/o passage ranker (§3.2)	WFA	68.05	63.82
w/o possibility classifier (§3.3)	ANS	69.64	65.41
Masque w/ gold passage ranker	ALL	78.70	78.14

Removing each component separately only slightly hurts the performance

Interesting experiment to replace Transformer by LSTM