Learning to Ask: Neural Question Generation for Reading Comprehension

Cornell University

Xinya Du, Junru Shao, Claire Cardie

QG dataset

What's QG and why QG?

A New Task: Automatic natural question **generation** for sentences from text passages in **reading comprehension**.

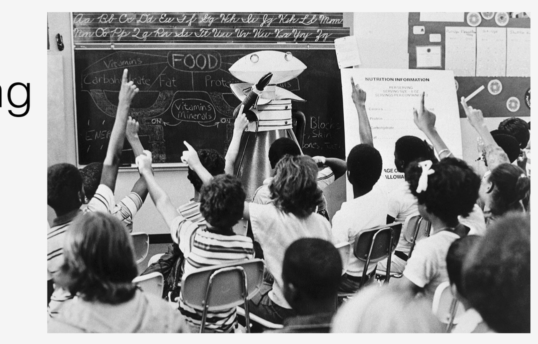
Example: From Wikipedia article *Oxygen*

Sentence:
Oxygen is used in cellular respiration and released by photosynthesis, which uses the energy of sunlight to produce oxygen from water.

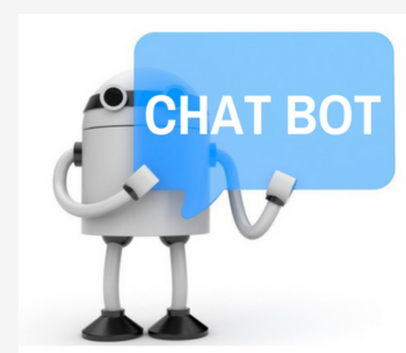
Questions:
– What life process produces oxygen in the presence of light?
photosynthesis
– Photosynthesis uses which energy to form oxygen from water?
sunlight
– From what does photosynthesis get oxygen?
water

Real Applications:

Education: Generating questions for testing understanding



Chat bot: asking questions to start a conversation or to request feedback.



Improving **question answering (QA)**

Sentence- and Paragraph-level seq2seq model

Task Objective:

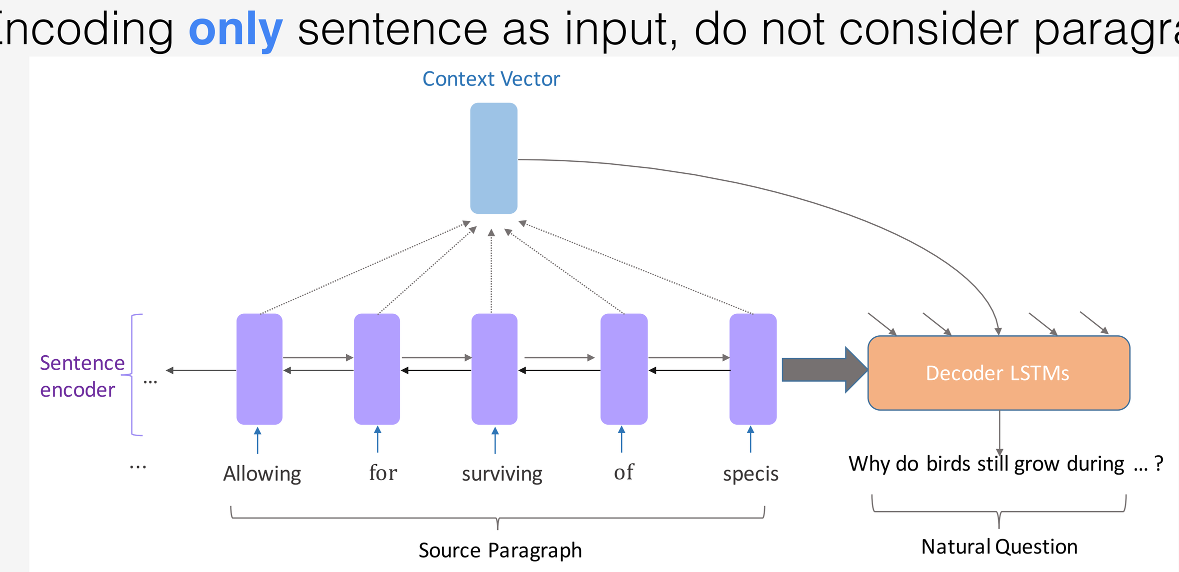
- **Input**: a natural language sentence $\mathbf{x}_s$ AND optionally a natural language paragraph $\mathbf{x}_p$
- **Objective**: To generate a question about the input sentence, such that: $\bar{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x})$
- We model the conditional next-word probability as:

Conditional log-likelihood of the predicted question y, given the input x.

Sentence-level model: $P(y_t|\mathbf{x}, y_{<t}) = \text{softmax}(\mathbf{W}_s \tanh(\mathbf{W}_t[\mathbf{h}_t; \mathbf{c}_t]))$

- Encoding **only** sentence as input, do not consider paragraph/context-level information.

$\mathbf{c}_t$ is the context vector, sum of the weighted avg. of encoder hidden units.

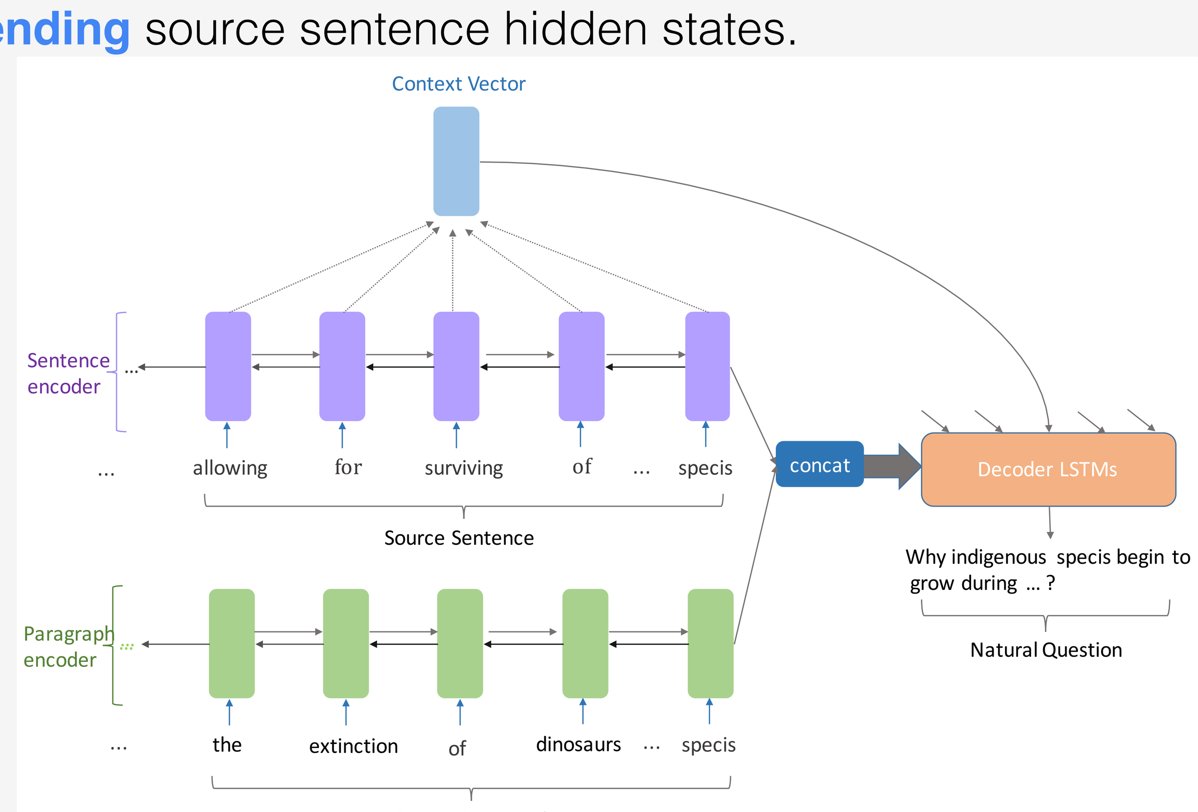


we use *bilinear* score to calculate the attention weights

$$a_{i,t} = \frac{\exp(\mathbf{h}_t^T \mathbf{W}_b \mathbf{b}_i)}{\sum_j \exp(\mathbf{h}_t^T \mathbf{W}_b \mathbf{b}_j)}$$

Paragraph-level model:

- Encoding **both** sentence and paragraph (that contains the sentence) as input, but only **attending** source sentence hidden states.



Also tried encoding *title/passage-level* information, but performance drops.

- Training: Minimize the negative log-likelihood with respect to θ: $\mathcal{L} = -\sum_{i=1}^{S} \log P\left(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}; \theta\right)$
- Inference: Beam search and **UNK replacement**

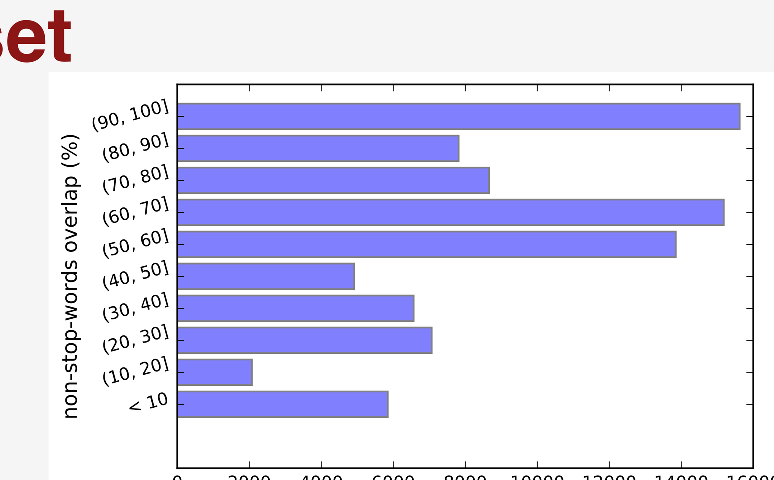
Very important!!! Even copy mechanism cannot *eliminate* UNKs.

For the UNK token at time step t, we replace it with the token in the *input sentence* with the **highest attention score**, the *index* of which is: $\arg\max_i a_{i,t}$

Dataset

# triple (Train)	70,484
# triple (Dev)	10,570
# triple (Test)	11,877
Sentence: avg. tokens	32.9
Question: avg. tokens	11.3
Avg. # questions per sentence	1.4

We pair *up* the questions with the *sentence*(s) which contain the answer span, and *paragraph* with contain the sentence.



Pruning constraint: the sentence-question pair have at least **one non-stop-word** in common.

SQuAD The Stanford Question Answering Dataset

We build the QG dataset based on the SQuAD corpus

Experiments

Automatic Evaluation:

Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	METEOR	ROUGE$_L$
IR$_{\text{Edit Distance}}$	18.28	5.48	2.26	1.06	7.73	20.77
DirectIn	31.71	21.18	15.11	11.20	14.95	22.47
H&S (rule-based)	38.50	22.80	15.52	11.18	15.95	30.98
MOSES+	15.61	3.64	1.00	0.30	10.47	17.82
Vanilla seq2seq	31.34	13.79	7.36	4.26	9.88	29.75
Our model (no pre-trained)	41.00	23.78	15.71	10.80	15.17	37.95
Our model (w/ pre-trained)	**43.09**	**25.96**	**17.50**	**12.28**	**16.62**	**39.75**
+ paragraph	42.54	25.33	16.98	11.86	16.28	39.37

Rule-based

Learning-based

- Our sentence-level model **beats** the strong rule-based system
- Directly copy (**DirectIn**) forms a very strong baseline.
- **Pre-trained** word embeddings help significantly.

Human Evaluation:

	Naturalness	Difficulty	Best %	Avg. rank
H&S (rule-based)	2.95	1.94	20.20	2.29
Ours	**3.36**	**3.03***	**38.38***	**1.94****
Human	3.91	2.63	66.42	1.46

Two-tailed t-test statistical significance: *($p < 0.005$), **($p < 0.001$)

- The neural model outperforms **significantly** rule-based methods by **human evaluation**.
- Larger margin compared with automatic eval., **better automatic metrics** to be designed.

Output Analysis

Paragraph 1 (truncated): ...... during the oligocene , for example , the rainforest spanned a relatively narrow band . it expanded again during the middle miocene , then retracted to a mostly inland formation at the last glacial maximum . however , the rainforest still managed to thrive during these glacial periods , allowing for the survival and evolution of a broad diversity of species .

Human: did the rainforest managed to thrive during the glacial periods ?

H&S (rule-based): what allowed for the survival and evolution of a broad diversity of species ?

Ours (sent.-level model): why do the birds still grow during glacial periods ?

Ours (para.-level model): why did the indigenous specis begin to grow during the glacial period ?

Green highlight shows the **input sentence**, which is used as input to both sent. and para.-level models

Our **sentence-level** model and **paragraph-level** both:

- *learns to select* an important aspect of the sentence
- Questions are more *natural sounding* and *vary more in terms of type*.

Para.-level model takes into account context info. **beyond sentence**.

Paragraph 2 (truncated): ...... kuznets ' curve predicts that income inequality will eventually decrease given time . as an example , income inequality did fall in the united states during its high school movement from 1910 to 1940 and thereafter . -lsb- citation needed -rsb- however , recent data shows that the level of income inequality began to rise after the 1970s . this does not necessarily disprove kuznets ' theory ....

Human: during what time period did income inequality decrease in the united states ?

H&S (rule-based): where did income inequality do fall during its high school movement from 1910 to 1940 and thereafter as an example ?

Ours (sent.-level model): when did income inequality fall in the us ?

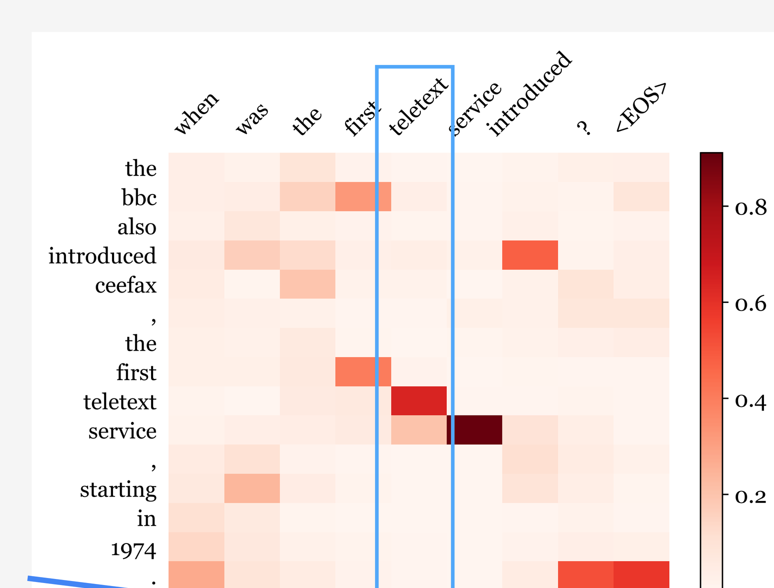
Ours (para.-level model): when did high school movement begin ?

Rule-based model copies nearly *word for word* the input sentence with minor syntactic change.

- redundant info.
- sometimes ungrammatical

Interpretability

Attention weight matrix shows the *soft alignment* between the sentence (left) and the generated question (top).



In this example, for the decoded token, the input sentence token with highest attention is ``teletext"

Media Coverage

- New Scientist "Inquisitive bot asks questions to test your understanding"
- Tech Republic "How researchers trained one AI system to start asking its own questions"

Conclusion

- We first proposed **the first fully data-driven** neural network approach for question generation in the reading comprehension setting. We investigated encoding **sentence-** and **paragraph-level** information for this task.
- Follow-up Work: Our **EMNLP17** paper on ***sentence selection for passage-level QG***, see you soon in Copenhagen :) !
- We release the **processed dataset** based on SQuAD.

Open question: How to better utilize QG for QA?