Identifying Where to Focus in Reading Comprehension for Neural Question Generation

Cornell University

Xinya Du and Claire Cardie

Passage-level QG

Question: How to do automatic natural **question generation (QG)** at **document-level**?

Example: From Wikipedia article *Fresno, California*

Fresno is the largest U.S. city not directly linked to an Interstate highway. When the Interstate Highway System was created in the 1950s, the decision was made to build what is now Interstate 5 on the west side of the Central Valley, and thus bypass many of the population centers in the region, instead of upgrading what is now State Route 99. Due to rapidly rising population and traffic in cities along SR 99, as well as the desirability of Federal funding, much discussion has been made to upgrade it to interstate standards and eventually incorporate it into the interstate system, most likely as Interstate 9. Major improvements to signage, lane width, median separation, vertical clearance, and other concerns are currently underway.

Q: Which is the largest city not connected to an interstate highway?
A: - *Fresno*

Q: Which State Route has been in discussion to upgrade to interstate standards?
A: *SR 99*

Q: What are the factors that are contributing to the desire to have SR 99 improved to be of interstate standards?
A: *rapidly rising population and traffic in cities along SR 99, as well as the desirability of Federal funding*

How?:
A first step, important (***question-worthy***) sentences ***selection***.

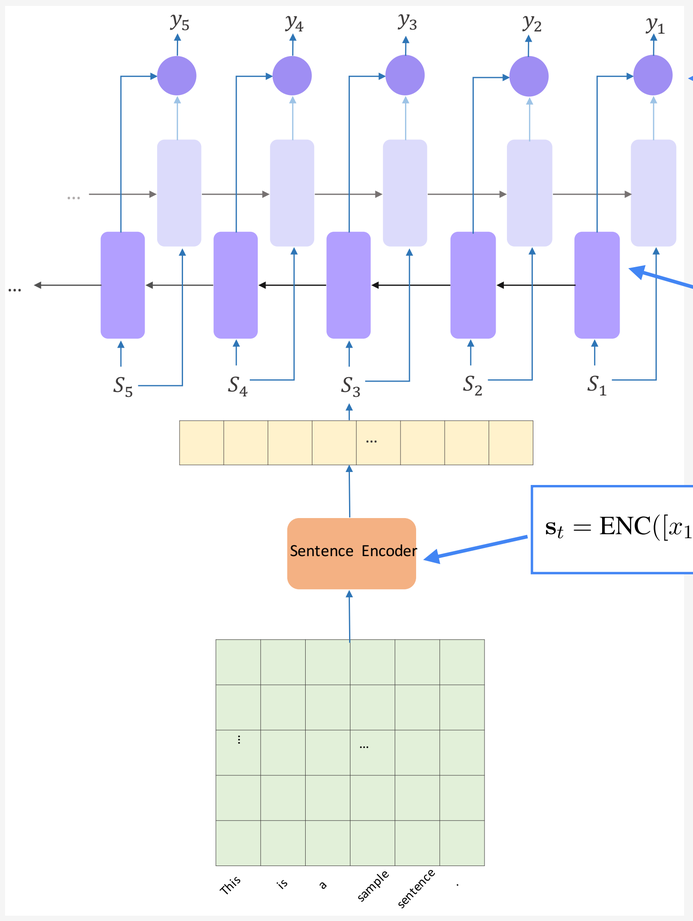
Why?:
- Educational app.
- Question Answering (QA)
- Generating FAQs
- Conversational agent

Hierarchical sequence tagging model

Task Objective:

- **Input**: a paragraph D consisting of a sequence of sentences $\{s_1, ..., s_m\}$
- **Objective**: To select a subset of k question-worthy sentences (k < m), such that:

$$\overline{\mathbf{y}} = \arg\max_{\mathbf{y}} \log P_1(\mathbf{y}|D)$$



$$P_1(y_t|D;\theta) = \text{softmax}\left(\text{MLP}\left(\tanh\left([\overrightarrow{\mathbf{h}_t};\overleftarrow{\mathbf{h}_t}]\right)\right)\right)$$

where **MLP** is multi-layer neural network and **tanh** is the activation function.

$$\overrightarrow{\mathbf{h}_t} = \overrightarrow{\text{LSTM}}\left(\mathbf{s}_t, \overrightarrow{\mathbf{h}_{t-1}}\right)$$
$$\overleftarrow{\mathbf{h}_t} = \overleftarrow{\text{LSTM}}\left(\mathbf{s}_t, \overleftarrow{\mathbf{h}_{t+1}}\right)$$

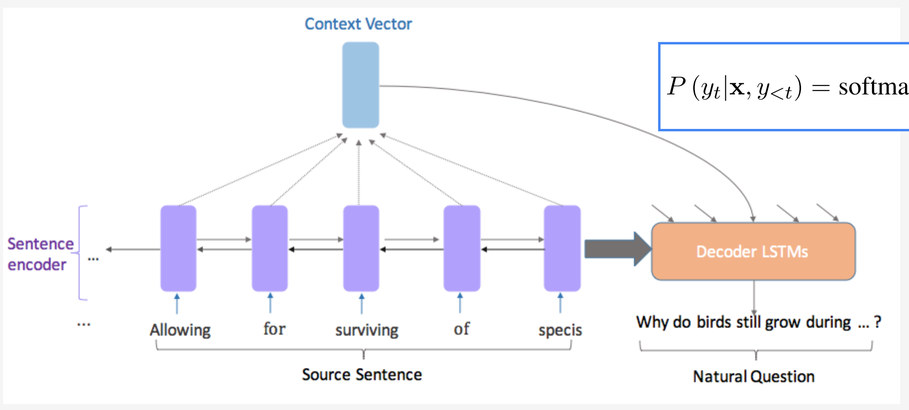
Then we use a **bidirectional LSTM** to encode the paragraph.

$$\mathbf{s}_t = \text{ENC}([x_1, ..., x_n])$$

The first layer encodes the sentences of the paragraph, with two encoding alternatives:
- **CNN + max pooling** (Kim, 2014; dos Santos and Zadrozny, 2014)
- **Sum operation**

Sentence-level question generation

- Encoding **only** sentence as input, do not consider paragraph/context-level, **attending** to source sentence hidden states.



$$P(y_t|\mathbf{x}, y_{<t}) = \text{softmax}\left(\mathbf{W}_s \tanh\left(\mathbf{W}_t[\mathbf{h}_t; \mathbf{c}_t]\right)\right)$$

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

- 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:

$$\arg\max_i a_{i,t}$$

Experiments

Automatic Evaluation:

Model	Precision	Recall	F-measure	Acc.	Paragraph-level Acc.
RANDOM	63.45	50.29	56.11	50.27	11.69
Majority Baseline	63.21	100.00	77.46	63.21	32.30
CNN (Kim, 2014)	68.35	**90.13**	77.74	67.38	24.73
LREG(w/ BOW)	68.52	86.55	76.49	66.37	31.36
LREG(w/ para.-level) (Cheng and Lapata, 2016)	70.49	89.08	78.70	69.52	33.95
Ours_{SUM} (no pre-trained)	73.02	89.23	80.32	72.36	36.46
Ours_{SUM} (w/ pre-trained)	73.85	87.65	80.16	**72.58**	36.30
Ours_{CNN} (no pre-trained)	73.15	89.29	**80.42***	72.52	35.93
Ours_{CNN} (w/ pre-trained)	**74.35**	86.11	79.80	72.44	**36.87**

Rule-based / Learning-based

- Our hierarchical sequence-tagging model **beats** the strong linear system.
- *Majority* forms a very strong baseline.
- **Pre-trained** word embeddings **do not** help significantly.

Full QG Evaluation:

Metric	Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	METEOR
Conservative	LREG(C&L) + NQG	38.30	23.15	15.64	10.97	15.09
	Ours + NQG	40.08	24.26	16.39	11.50	15.67
Liberal	LREG(C&L) + NQG	51.55	40.17	34.35	30.59	24.17
	Ours + NQG	52.89	41.16	35.15	31.25	24.76

The full QG system with our sentence selection component achieves SOTA performance

For **detailed explanation** for the evaluation metric, plz refer to our paper.

Output Analysis

Paragraph 1 in 1939 , coinciding with the start of world war ii , rene dubos reported the discovery of the first naturally derived antibiotic , tyrothricin , a compound of 20 % gramicidin and 80 % tyrocidine , from b. brevis . it was one of the first commercially manufactured antibiotics universally and was very effective in treating wounds and ulcers during world war ii . gramicidin , however , could not be used systemically because of toxicity . tyrocidine also proved too toxic for systemic usage . research results obtained during that period were not shared between the axis and the allied powers during the war .
Our questions:
Q1: what was the name of the compound that was discovered in 1939 ?
Q2: what was one of the first commercially manufactured antibiotics ?

Gold questions:
Q1: (1) what was the first antibiotic developed from nature ? (2) when was tyrothricin created ? (3) what also happened in 1939 besides tyrothricin ?
Q2: what was tyrothricin used for during the war ?

Paragraph 2 after the english civil war the royal citadel was built in 1666 on the east end of plymouth hoe , to defend the port from naval attacks , suppress plymothian parliamentary leanings and to train the armed forces . Guided tours are available in the summer months . further west is smeaton 's tower , which was built in 1759 as a lighthouse on rocks 14 miles -lrb- 23 km -rrb- off shore , but dismantled and the top two thirds rebuilt on the hoe in 1877 . It is open to the public and has views over the plymouth sound and the city from the lantern room . Plymouth has 20 war memorials of which nine are on the hoe including : plymouth naval memorial , to remember those killed in world wars i and ii , and the armada memorial , to commemorate the defeat of the spanish armada .
Our questions:
Q1: when was the royal basilica fort built ?
Q2: when was the smeaton 's tower built ?
Q3: how many war memorials did plymouth have ?

Gold questions:
Q1: in what year was the royal citadel constructed ?
Q2: (1) when was smeaton 's tower first constructed ? (2) in kilometers , how far off the coast was smeaton 's tower originally built ?
Q3: what memorial commemorates the naval victory over the spanish armada ?

Paragraph 3 the city receives 49.9 inches -lrb- 1,270 mm -rrb- of precipitation annually , which is fairly spread throughout the year . average winter snowfall between 1981 and 2010 has been 25.8 inches -lrb- 66 cm -rrb- , but this varies considerably from year to year . hurricanes and tropical storms are rare in the new york area , but are not unheard of and always have the potential to strike the area . hurricane sandy brought a destructive storm surge to new york city on the evening of october 29 , 2012 , flooding numerous streets , tunnels , and subway lines in lower manhattan and other areas of the city and cutting off electricity in many parts of the city and its suburbs . the storm and its profound impacts have prompted the discussion of constructing seawalls and other coastal barriers around the shorelines of the city and the metropolitan area to minimize the risk of destructive consequences from another such event in the future .
Our questions:
Q1: how much precipitation does the city receive annually ?
Q2: what is the average winter snowfall between 1981 and 2010 ?
Q3: what caused the storm surge to new york city ?

Gold questions:
Q1: (1) in millimeters , how much precipitation does new york receive a year ? (2) how many inches of precipitation does nyc get in a year ?
Q2: (1) in centimeters , what is the average winter snowfall ? (2) the mean snowfall between 1981 and 2010 in nyc has been how many inches ?
Q3: (1) when did hurricane sandy strike new york ? (2) which natural disaster occurred on october 29 , 2012 in nyc ?

- Red highlight shows the **selected sentence**, which is used as input for question generation.
- Wave-lined sentences shows the **ground truth** question-worthy sentences.
- Appropriate questions might still be generated for **wrongly selected sentences**, generated question sometime ask **different aspects** with ground truth question.
- **Quality** of some generated questions to be improved.

Conclusion

- We introduce the **new task** of question-worthy sentences selection.
- We introduce a neural **sentence-level sequence tagging** approach for this task, with sum or CNN for encoding the sentences.
- Our system outperforms the baselines (e.g. feature-rich linear systems) significantly.

Open questions:
- **Connecting QG and QA!**
- **Better dataset for the task.**