IITB-Monash Research Academy

An Indian-Australian research partnership

Indian Institute of Technology Bombay

MONASH University

Automating reading comprehension by generating question and answer pairs

Vishwajeet Kumar [1] Kireeti Boorla [2] Ganesh Ramakrishnan [2] Yuan-Fang Li [3]

[1]IITB-Monash Research Academy, India

[2]IIT Bombay, India

[3]Monash University, Australia

Automatic question and answer generation

A system to automatically generate questions and answers from text.

Some text

Sachin Tendulkar received the Arjuna Award in 1994 for his outstanding sporting achievement, the Rajiv Gandhi Khel Ratna award in 1997...

Questions

1. When did Sachin Tendulkar received the Arjuna Award?
   Ans: 1994
2. which award did sachin tendular received in 1994 for his outstanding sporting achievement?
   Ans: Arjuna Award
3. when did Sachin tendulkar received the Rajiv Gandhi Khel Ratna Award?
   Ans: 1997

Motivation

Sachin Ramesh Tendulkar is a former Indian cricketer and captain, widely regarded as one of the greatest batsmen of all time. He took up cricket at the age of eleven, made his Test debut on 15 November 1989 against Pakistan in Karachi at the age of sixteen, and went on to represent Mumbai domestically and India internationally for close to twenty-four years.....

Motivation

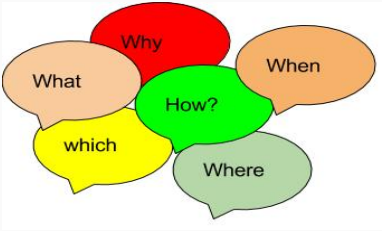
Sachin Ramesh Tendulkar is a former Indian cricketer and captain, widely regarded as one of the greatest batsmen of all time. He took up cricket at the age of eleven, made his Test debut on 15 November 1989 against Pakistan in Karachi at the age of sixteen, and went on to represent Mumbai domestically and India internationally for close to twenty-four years.....

How would someone tell that you have read this text?

Motivation

Sachin Ramesh Tendulkar is a former Indian cricketer and captain, widely regarded as one of the greatest batsmen of all time. He took up cricket at the age of eleven, made his Test debut on 15 November 1989 against Pakistan in Karachi at the age of sixteen, and went on to represent Mumbai domestically and India internationally for close to twenty-four years.....

How would someone tell that you have read this text?



Why is this problem Challenging?

- Question Must be Relevant to the Text

Why is this problem Challenging?

- Question Must be Relevant to the Text
- Answer Must be Unambiguous

Why is this problem Challenging?

- Question Must be Relevant to the Text
- Answer Must be Unambiguous
- Question must be challenging and well formed

Existing Work

Template Based [Mazidi and Nielsen, 2014, Mostow and Chen, 2009]

- Use crowd sourced templates such as What is X ?

Syntax Based [Heilman, 2011]

- Rules for declarative-to-interrogative sentence transformation
- Only syntax is considered not semantics.
- Rely heavily on NLP tools.

Vanilla Seq2Seq for Question Generation [Du et al., 2017]

- First approach towards question generation from text using neural network.
- Uses vanilla Seq2Seq model for question generation.

Some other related work

Generate question given a fact/triple from KB/Ontology.

Some other related work

Generate question given a fact/triple from KB/Ontology.
Example: <**Fires Creek**, contained by, nantahala national forest> $\Rightarrow$ Which forest is **Fires Creek** in?

Some other related work

Generate question given a fact/triple from KB/Ontology.

Example: <**Fires Creek**, contained by, nantahala national forest> ⇒ Which forest is **Fires Creek** in?

Template based [Seyler et al., 2015]

- Assumption: Facts are present in Domain dependent knowledge base.
- Generates question using templates based on facts.

Factoid question generation using RNN [Serban et al., 2016]

- Propose generating factoid question generation from freebase triples(subject,relation,object).
- Embeds fact using KG embedding techniques such as TransE.

Limitations of previous approaches

- Mostly rule based or template based.

Limitations of previous approaches

- Mostly rule based or template based.
- Do not generate answer corresponding to the question.

Limitations of previous approaches

- Mostly rule based or template based.
- Do not generate answer corresponding to the question.
- Overly simple set of linguistic features.

Our contribution

- Pointer network based method for automatic answer selection.

Our contribution

- Pointer network based method for automatic answer selection.
- Sequence to sequence model with attention and augmented with rich set of linguistic features and answer encoding

Our contribution

- Pointer network based method for automatic answer selection.
- Sequence to sequence model with attention and augmented with rich set of linguistic features and answer encoding

Our contribution

- Pointer network based method for automatic answer selection.
- Sequence to sequence model with attention and augmented with rich set of linguistic features and answer encoding

Sentence: It was adopted into an Oscar-winning film in 1962 by director Robert Mulligan, with a screenplay by Horton Foote.
Feature Tagged Sentence: It|PRP|O|nsubjpass was|VBD|O|auxpass adapted|VBN|O|ROOT into|IN|O|case
.... Horton|N|N|O|Compound Foote|N|N|O|Compound

With Features:	Who was the director of the film ?
With Features and Answer as "Horton Foote":	Who wrote the movie ?
With Features and Answer as "Robert Mulligan":	Who was the director of the Oscar-winning movie ?

Our contribution

- Pointer network based method for automatic answer selection.
- Sequence to sequence model with attention and augmented with rich set of linguistic features and answer encoding

Sentence: It was adopted into an Oscar-winning film in 1962 by director Robert Mulligan, with a screenplay by Horton Foote.
Feature Tagged Sentence: It|PRP|O|nsubjpass was|VBD|O|auxpass adapted|VBN|O|ROOT into|IN|O|case
.... Horton|N|N|O|Compound Foote|N|N|O|Compound

With Features:	Who was the director of the film ?
With Features and Answer as "Horton Foote":	Who wrote the movie ?
With Features and Answer as "Robert Mulligan":	Who was the director of the Oscar-winning movie ?

Automatic question and answer generation using seq2seq model with pointer network

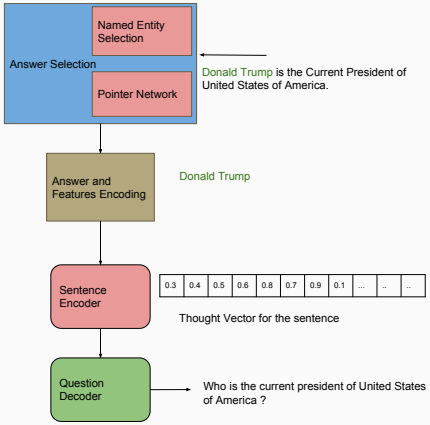
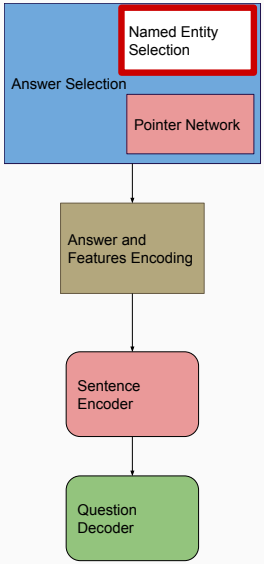


Figure 1: High level architecture of our question generation model

Named Entity Selection

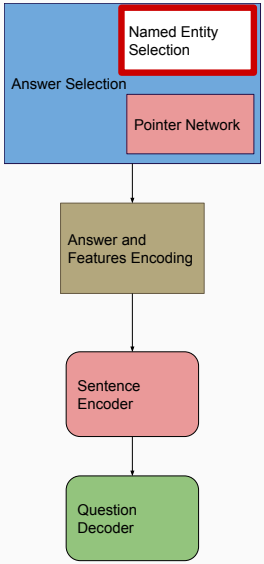


- Sentence $S = (w_1, w_2, ..., w_n)$ is encoded using a 2-layer LSTM network into hidden states $H = (h_1^s, h_2^s, ..., h_n^s)$.

---

where h_n^s is final state
h_{mean}^s is the mean of all activations
h_{mean}^{ne} is mean of activations in NE span $(h_i^s, ..., h_j^s)$
[a] Most relevant answer to ask question about

Named Entity Selection

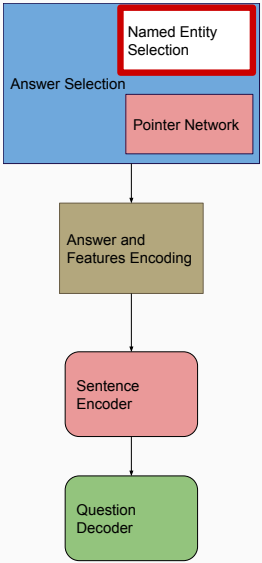


- Sentence $S = (w_1, w_2, ..., w_n)$ is encoded using a 2-layer LSTM network into hidden states $H = (h_1^s, h_2^s, ..., h_n^s)$.
- For each NE, $NE = (n_i, ..., n_j)$, create representation (**R**) $=< h_{mean}^{ne} >$,

---

where h_n^s is final state
h_{mean}^s is the mean of all activations
h_{mean}^{ne} is mean of activations in NE span $(h_i^s, ..., h_j^s)$
[a] Most relevant answer to ask question about

Named Entity Selection

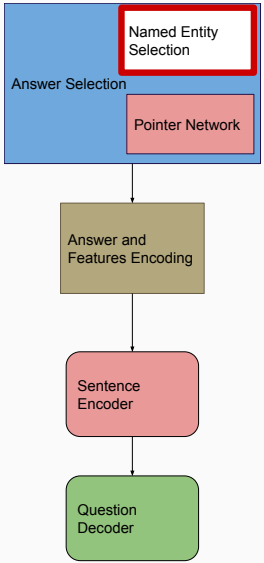


- Sentence $S = (w_1, w_2, ..., w_n)$ is encoded using a 2-layer LSTM network into hidden states $H = (h_1^s, h_2^s, ..., h_n^s)$.
- For each NE, $NE = (n_i, ..., n_j)$, create representation (**R**) $=< h_{mean}^{ne} >$,
- **R** is fed to MLP along with $< h_n^s; h_{mean}^s; >$ to get probability of named entity being pivotal answer[a].

---

where h_n^s is final state
h_{mean}^s is the mean of all activations
h_{mean}^{ne} is mean of activations in NE span $(h_i^s, ..., h_j^s)$
[a] Most relevant answer to ask question about

Named Entity Selection

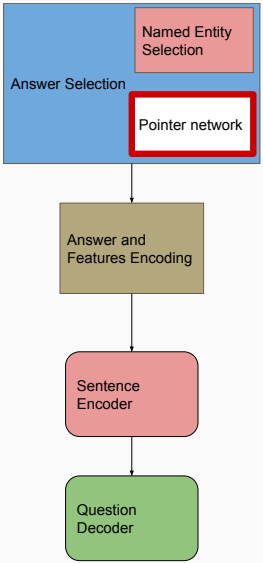


- Sentence $S = (w_1, w_2, ..., w_n)$ is encoded using a 2-layer LSTM network into hidden states $H = (h_1^s, h_2^s, ..., h_n^s)$.
- For each NE, $NE = (n_i, ..., n_j)$, create representation $(\mathbf{R})$ $=< h_{mean}^{ne} >$,
- $\mathbf{R}$ is fed to MLP along with $< h_n^s; h_{mean}^s; >$ to get probability of named entity being pivotal answer[a].
- $P(NE_i|S) = softmax(\mathbf{R}_i.W + B)$

---

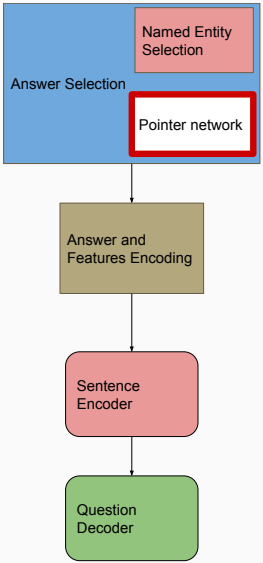
where h_n^s is final state
h_{mean}^s is the mean of all activations
h_{mean}^{ne} is mean of activations in NE span $(h_i^s, ..., h_j^s)$
[a] Most relevant answer to ask question about

Answer selection using Pointer networks



- Given encoder hidden states $H = (h_1, h_2, \ldots, h_n)$, the probability of generating $O = (o_1, o_2, \ldots, o_m)$ is : $P(O|S) = \prod P(o_i|o_1, o_2, o_3, \ldots, o_{i-1}; H)$

Answer selection using Pointer networks



- Given encoder hidden states $H = (h_1, h_2, \ldots, h_n)$, the probability of generating $O = (o_1, o_2, \ldots, o_m)$ is : $P(O|S) = \prod P(o_i|o_1, o_2, o_3, \ldots, o_{i-1}; H)$
- Probability distribution is modeled as:

$$u^i = v^T tanh(W^e \hat{H} + W^d D_i) \quad (1)$$

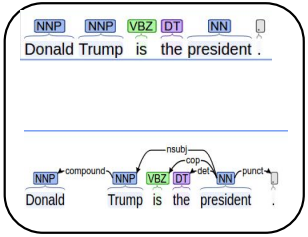
$$P(O|S) = softmax(u^i) \quad (2)$$

Sentence:

Donald Trump is the President.

Question:

Who is Donald Trump ?



POS Tag and Dependency Label

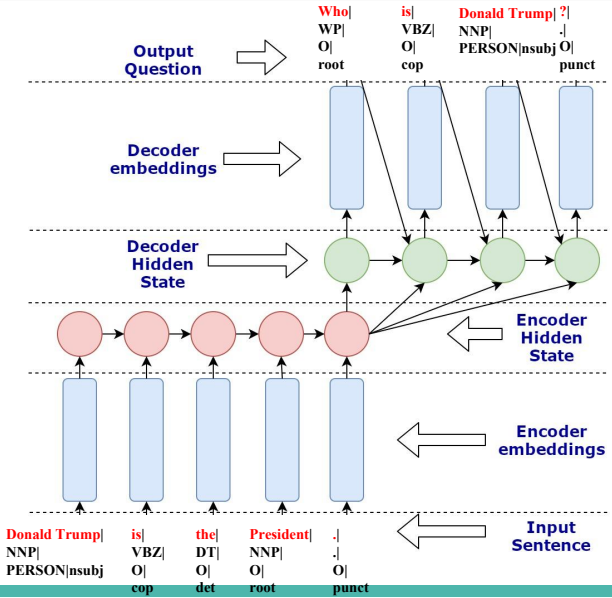
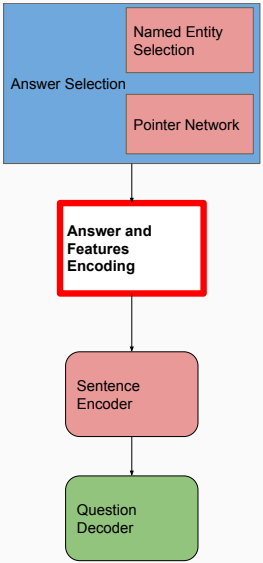


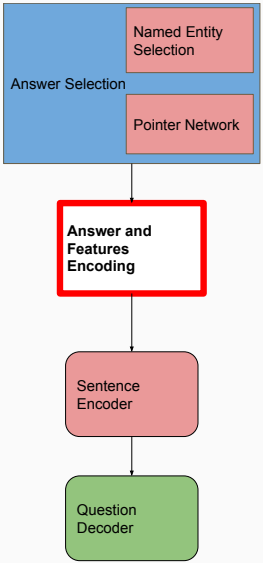
Figure 2: Question generation

Features and Answer Encoding



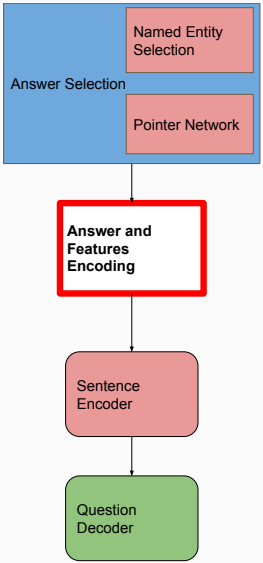
- POS tag, Named Entity tag, Dependency label as linguistic features.

Features and Answer Encoding



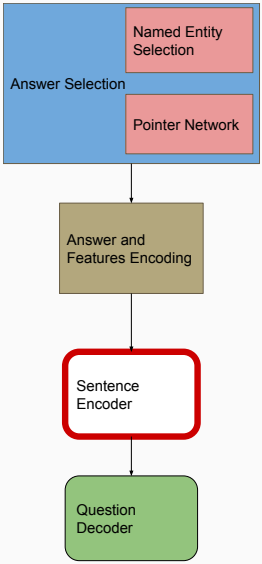
- POS tag, Named Entity tag, Dependency label as linguistic features.
- Rich set of linguistic features help model learn better generalize transformation rules.

Features and Answer Encoding



- POS tag, Named Entity tag, Dependency label as linguistic features.
- Rich set of linguistic features help model learn better generalize transformation rules.
- Dependency label is the edge label connecting each word with the parent in the dependency tree.

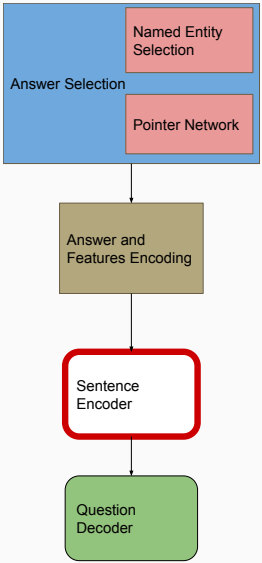
Sentence Encoder



- BiLSTM to capture both left context and the right context.

$\hat{h_t}$ is the thought vector $W, V,$ *and* $U \in R^{n \times m}$ are trainable parameters, $w_t \in R^{p \times q \times r}$ is feature encoded word embedding at time step t.

Sentence Encoder

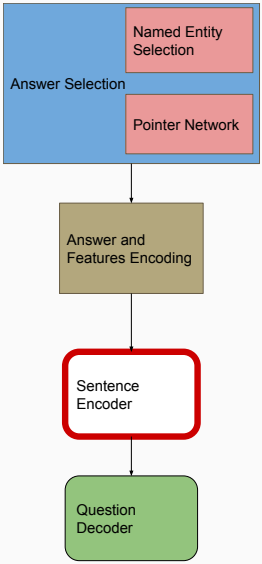


- BiLSTM to capture both left context and the right context.
-

$$\overrightarrow{\hat{h}_t} = f(\overrightarrow{W} w_t + \overrightarrow{V}\overrightarrow{\hat{h}_{t-1}} + \overrightarrow{b}), \overleftarrow{\hat{h}_t} = f(\overleftarrow{W} w_t + \overleftarrow{V}\overleftarrow{\hat{h}_{t+1}} + \overleftarrow{b}) \quad (3)$$

$\hat{h}_t$ is the thought vector $W, V,$ *and* $U \in R^{n \times m}$ are trainable parameters, $w_t \in R^{p \times q \times r}$ is feature encoded word embedding at time step t.

Sentence Encoder



- BiLSTM to capture both left context and the right context.
-

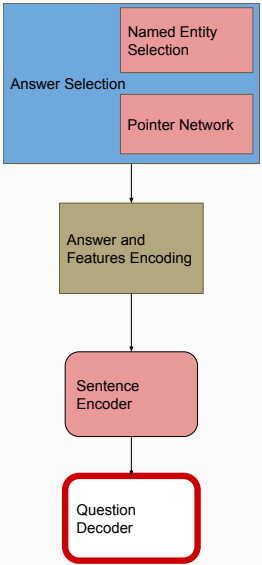
$$\overrightarrow{\hat{h}_t} = f(\overrightarrow{W} w_t + \overrightarrow{V}\overrightarrow{\hat{h}_{t-1}} + \overrightarrow{b}), \overleftarrow{\hat{h}_t} = f(\overleftarrow{W} w_t + \overleftarrow{V}\overleftarrow{\hat{h}_{t+1}} + \overleftarrow{b}) \quad (3)$$

-

$$\hat{h}_t = g(Uh_t + c) = g(U[\overrightarrow{\hat{h}_t}, \overleftarrow{\hat{h}_t}] + c) \quad (4)$$

$\hat{h}_t$ is the thought vector $W, V,$ *and* $U \in R^{n \times m}$ are trainable parameters, $w_t \in R^{p \times q \times r}$ is feature encoded word embedding at time step t.

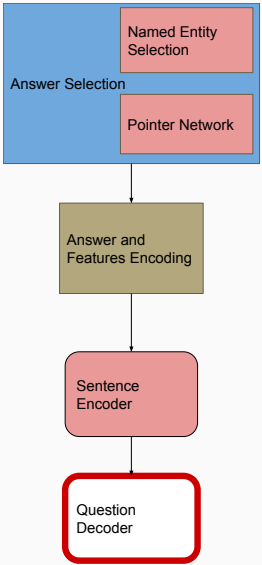
Question Decoder



- 2-layer LSTM network.

where W_s and W_r are weight vectors and *tanh* is the activation function.

Question Decoder

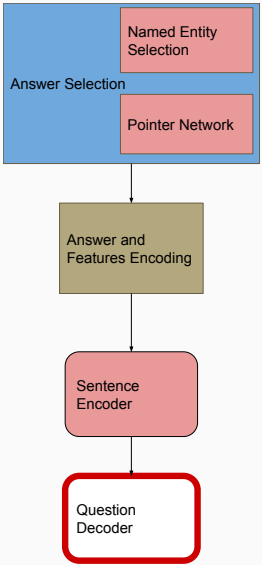


- 2-layer LSTM network.
- Decoder:

$$P(Q|S;\theta) = softmax(W_s(tanh(W_r[h_t, c_t] + b)) \quad (5)$$

where W_s and W_r are weight vectors and *tanh* is the activation function.

Question Decoder



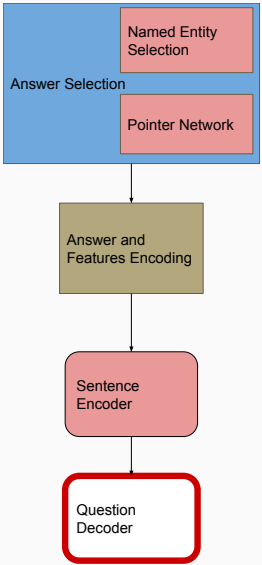
- 2-layer LSTM network.
- Decoder:

$$P(Q|S;\theta) = softmax(W_s(tanh(W_r[h_t, c_t] + b)) \quad (5)$$

- Beam search with beam_size 3 to decode question.

where W_s and W_r are weight vectors and *tanh* is the activation function.

Question Decoder



- 2-layer LSTM network.
- Decoder:

$$P(Q|S;\theta) = softmax(W_s(tanh(W_r[h_t, c_t] + b)) \quad (5)$$

- Beam search with beam_size 3 to decode question.
- Suitably modified decoder integrated with an attention mechanism to handle rare word problem.

---

where W_s and W_r are weight vectors and *tanh* is the activation function.

Attention Mechanism

Attention distribution:

W_{eh}, W_{sh} and $batt$ are learnable model parameters.
W_v and b_v are trainable parameter.

Attention Mechanism

Attention distribution:

$$e_i^t = v^t tanh(W_{eh} h_i + W_{sh} s_t + b_{att}) \tag{6}$$

W_{eh}, W_{sh} and $batt$ are learnable model parameters.
W_v and b_v are trainable parameter.

Attention Mechanism

Attention distribution:

$$e_i^t = v^t tanh(W_{eh} h_i + W_{sh} s_t + b_{att}) \tag{6}$$

$$a^t = softmax(e^t) \tag{7}$$

W_{eh}, W_{sh} and $batt$ are learnable model parameters.
W_v and b_v are trainable parameter.

Attention Mechanism

Attention distribution:

$$e_i^t = v^t tanh(W_{eh} h_i + W_{sh} s_t + b_{att}) \tag{6}$$

$$a^t = softmax(e^t) \tag{7}$$

$$c_t^* = \Sigma_i a_i^t h_i \tag{8}$$

W_{eh}, W_{sh} and $batt$ are learnable model parameters.
W_v and b_v are trainable parameter.

Attention Mechanism

Attention distribution:

$$e_i^t = v^t tanh(W_{eh} h_i + W_{sh} s_t + b_{att}) \tag{6}$$

$$a^t = softmax(e^t) \tag{7}$$

$$c_t^* = \Sigma_i a_i^t h_i \tag{8}$$

Probability distribution over vocabulary is:

$$P_{vocab} = sofmax(W_v[s_t, c_t^*] + b_v) \tag{9}$$

W_{eh}, W_{sh} and $batt$ are learnable model parameters.
W_v and b_v are trainable parameter.

Attention Mechanism

Attention distribution:

$$e_i^t = v^t tanh(W_{eh} h_i + W_{sh} s_t + b_{att}) \tag{6}$$

$$a^t = softmax(e^t) \tag{7}$$

$$c_t^* = \Sigma_i a_i^t h_i \tag{8}$$

Probability distribution over vocabulary is:

$$P_{vocab} = sofmax(W_v[s_t, c_t^*] + b_v) \tag{9}$$

Overall loss is calculated as:

$$LOSS = \frac{1}{T} \Sigma_{t=0}^{T} - \log P_{vocab}(word_t) \tag{10}$$

W_{eh}, W_{sh} and $batt$ are learnable model parameters.
W_v and b_v are trainable parameter.

Human evaluation results

System	p1(%)	p2(%)	p3(%)
QG [Du et al., 2017]	51.6	48	52.3
QG+F	59.6	57	64.6
QG+F+NE	57	52.6	67
QG+GAE	44	35.3	50.6
QG+F+AES	51	47.3	55.3
QG+F+AEB	61	60.6	**71.3**
QG+F+GAE	**63**	**61**	67

Table 1: Human evaluation results on $\mathcal{S}^{te}$. Parameters are, **p1**: percentage of syntactically correct questions, **p2**: percentage of semantically correct questions, **p3**: percentage of relevant questions.

F	Features
NE	Named entity selection
AES	Sequence pointer network
AEB	Boundary pointer network
GAE	Ground truth answer encoding

blue $\Rightarrow$ different alternatives for encoding the pivotal answer.
green $\Rightarrow$ set of linguistic features that can be optionally added to any model.

Automatic evaluation results

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
QG [Du et al., 2017]	39.97	22.39	14.39	9.64	14.34	37.04
QG+F	41.89	24.37	15.92	10.74	15.854	37.762
QG+F+NE	41.54	23.77	15.32	10.24	15.906	36.465
QG+GAE	43.35	24.06	14.85	9.40	15.65	37.84
QG+F+AES	**43.54**	**25.69**	**17.07**	**11.83**	**16.71**	**38.22**
QG+F+AEB	**42.98**	**25.65**	**17.19**	**12.07**	**16.72**	**38.50**
QG+F+GAE	**46.32**	**28.81**	**19.67**	**13.85**	**18.51**	**41.75**

blue ⇒ different alternatives for encoding the pivotal answer.
green ⇒ set of linguistic features that can be optionally added to any model.

Some sample questions generated

Sentence 1: The museum was founded by the nurse and explorer Kate Marsden and Reverend J.C. Thompson FGS.
Answer Predicted: Nurse and explorer Kate Marsden
Question Generated: who founded the museum ?

Sentence 2: American idol premiered in June 2002 and became the surprise summer hit show of 2002.
Answer Predicted: June 2002
Question Generated: When did American idol begin ?

Sentence 3: Shuman then constructed a full-scale steam engine powered by low-pressure water, enabling him to patent the entire solar engine system by 1912.
Answer Predicted: 1912
Question Generated: When was the solar engine system invented ?

Sentence 4: Journalist Vinod Verma arrested for alleged extortion bid on Chhattisgarh minister .
Answer Predicted: alleged extortion bid on chhattisgarh minister
Question Generated: what did journalist vinod arrested for ?

Sentence 5: Donald Trump is the current president of the United States .
Answer Predicted: Donald Trump
Question Generated: who is the current president of the United States ?

Sentence 6: Manhattan was on track to have an estimated 90,000 hotel rooms at the end of 2014, a 10% increase from 2013.
Answer Predicted: 90000
Question Generated: How many hotel rooms did Manhattan have ?

Conclusion

- We introduced a novel two-stage process to generate question-answer pairs from text.
- We proposed an automatic answer selection technique using pointer network.
- We incorporate attention mechanism to the decoder to handle rare word problem.

Questions?

References I

Du, X., Shao, J., and Cardie, C. (2017).
Learning to ask: Neural question generation for reading comprehension.
In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1342–1352.

Heilman, M. (2011).
Automatic factual question generation from text.
PhD thesis, Carnegie Mellon University.

Mazidi, K. and Nielsen, R. D. (2014).
Linguistic considerations in automatic question generation.
In *ACL (2)*, pages 321–326.

References II

Mostow, J. and Chen, W. (2009).
Generating instruction automatically for the reading strategy of self-questioning.
In *AIED*, pages 465–472.

Serban, I. V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., and Bengio, Y. (2016).
Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 588–598.

Seyler, D., Berberich, K., and Weikum, G. (2015).
Question Generation from Knowledge Graphs.
PhD thesis, Universität des Saarlandes Saarbrücken.