LEARNING TO ASK: NEURAL QUESTION GENERATION FOR READING COMPREHENSION

ADVISOR: JIA-LING KOH
SPEAKER: YIN-HSIANG LIAO
2018/08/21, FROM ACL 2017

Outline

- Introduction
- Definition
- Model
- Experiment
- Conclusion

Introduction

- Question generation (QG)
  - Purpose: creating natural questions from a given sentence or paragraph.
  - Real application: education, chat bot.

Introduction

- Example:
- (from Wiki, 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

Introduction

Past

Deep linguistic knowledge → Deep neural model

Rule-based **Data-driven**

Syntactic role

Semantic role

Outline

- Introduction
- Definition
- Model
- Experiment
- Conclusion

Definition

- Task definition:
  - Find $\bar{y}$

$$\overline{\mathbf{y}} = \arg\max_{\mathbf{y}} P\left(\mathbf{y}|\mathbf{x}\right)$$

- x: input sentence, y: natural question
- P(y|x) : conditional log-likelihood of the predicted y.

Outline

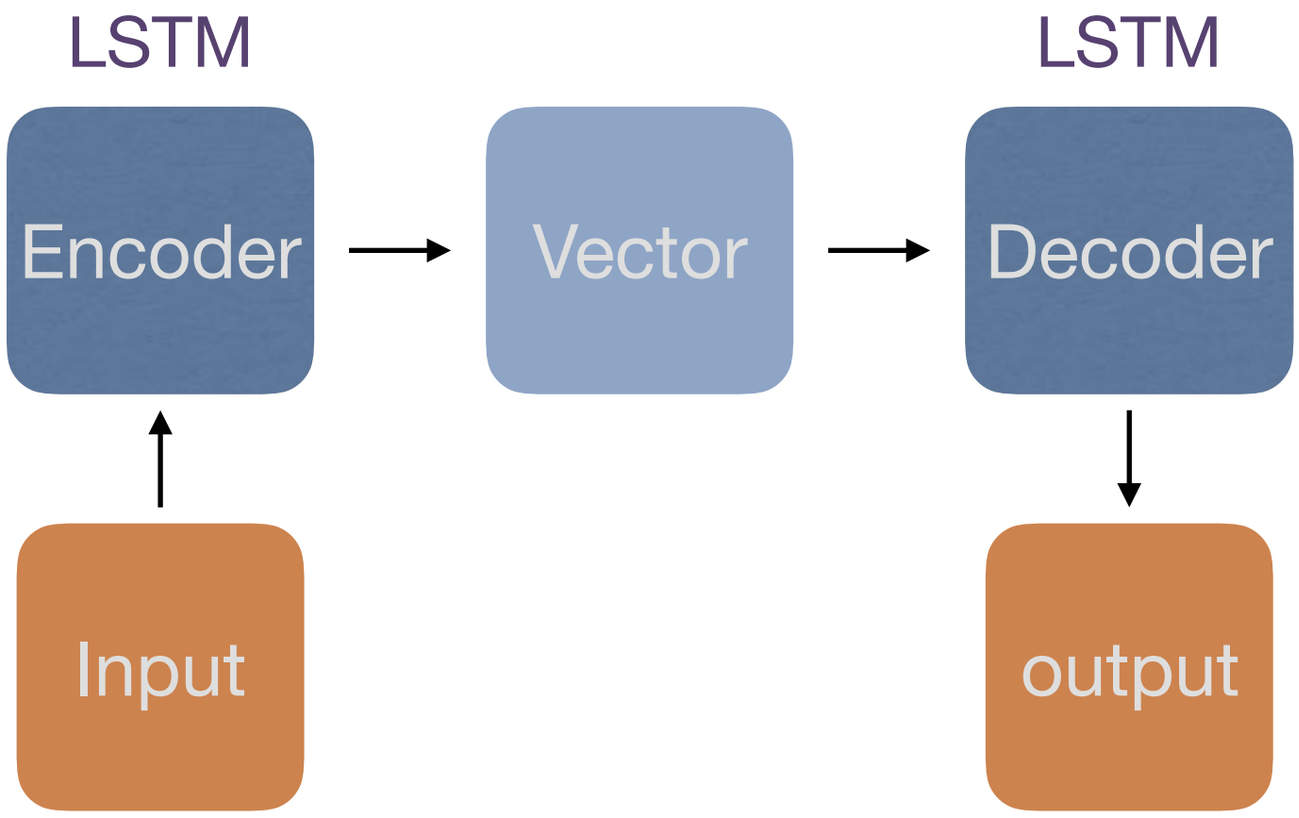
- Introduction
- Definition
- Model
- Experiment
- Conclusion

Model

- Inspired by the way in which people ask question.
- People usually pay ***attention*** to certain parts of the incoming sentence.
- RNN ***encoder-decoder*** architecture with global attention mechanism.

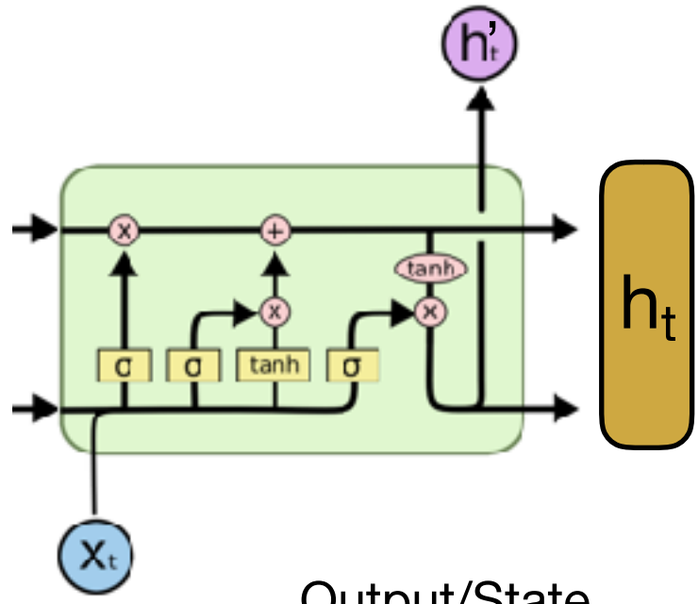
Model

- Encoder-decoder mechanism:



Model

- Decoder:



Output/State

$$\overline{\mathbf{y}} = \arg\max_{\mathbf{y}} P\left(\mathbf{y}|\mathbf{x}\right)$$

$$P\left(\mathbf{y}|\mathbf{x}\right) = \prod_{t=1}^{|y|} P\left(y_t|\mathbf{x}, y_{<t}\right)$$

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

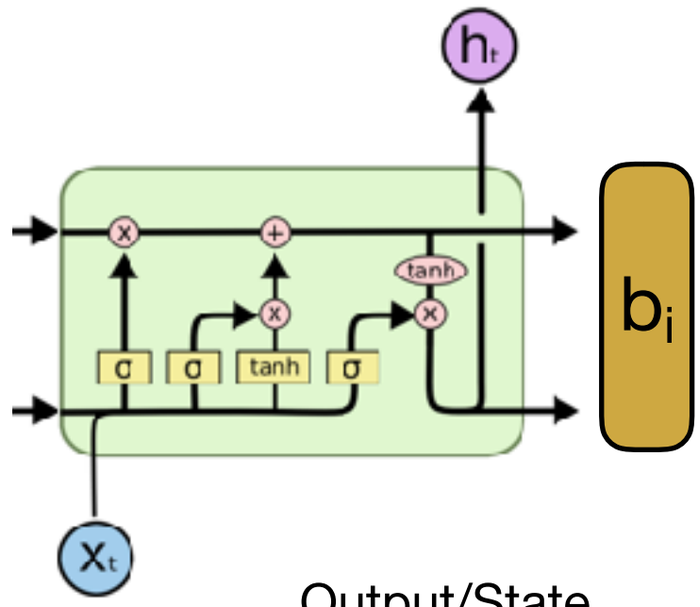
$$\mathbf{h}_t = \text{LSTM}_1\left(y_{t-1}, \mathbf{h}_{t-1}\right)$$

Model

- Encoders
- Attention-based sentence encoder
  - Given a sentence x , we encode the sentence.
    - bi-LSTM and attention weights
- Paragraph encoder
  - Given a sentence x, we encode the paragraph containing x.
  - bi-LSTM without attention on outputs of states.

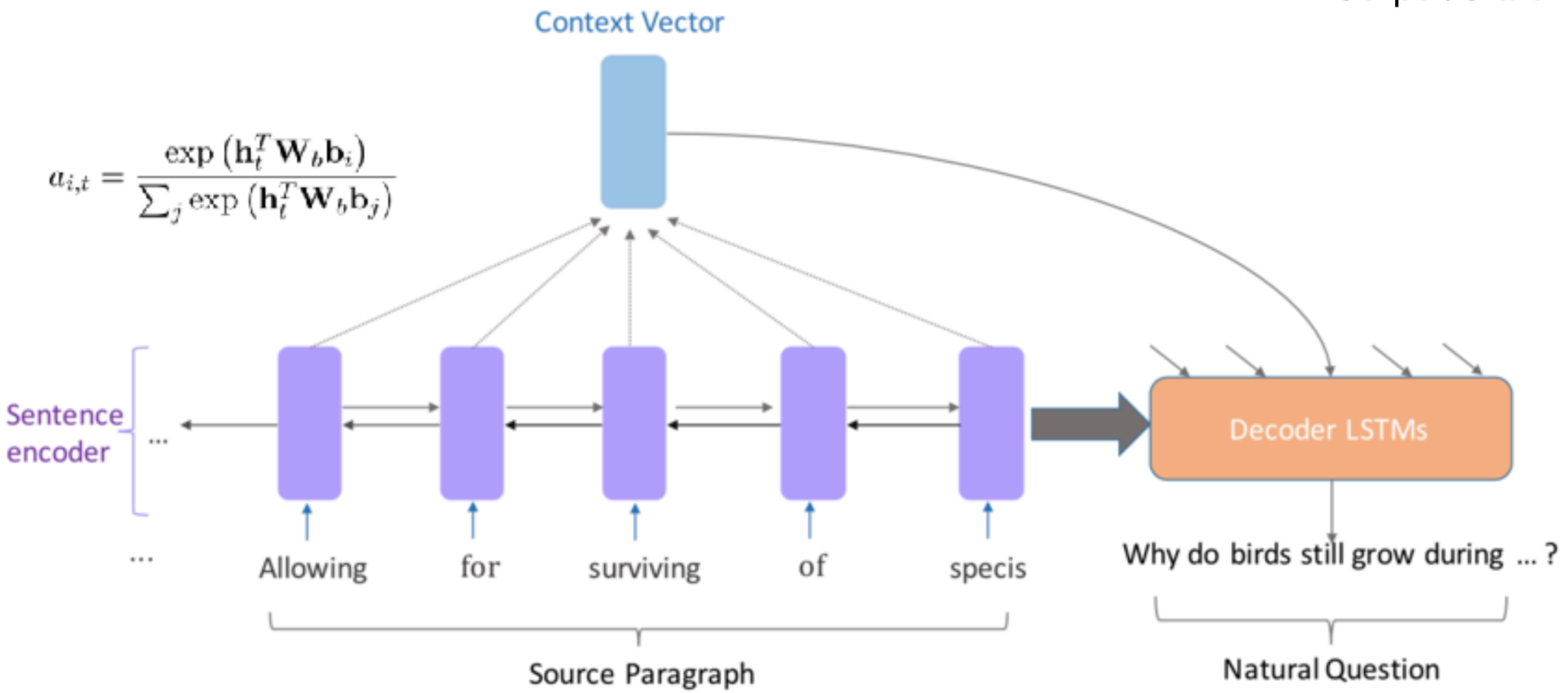
Model

- Sentence-level model:



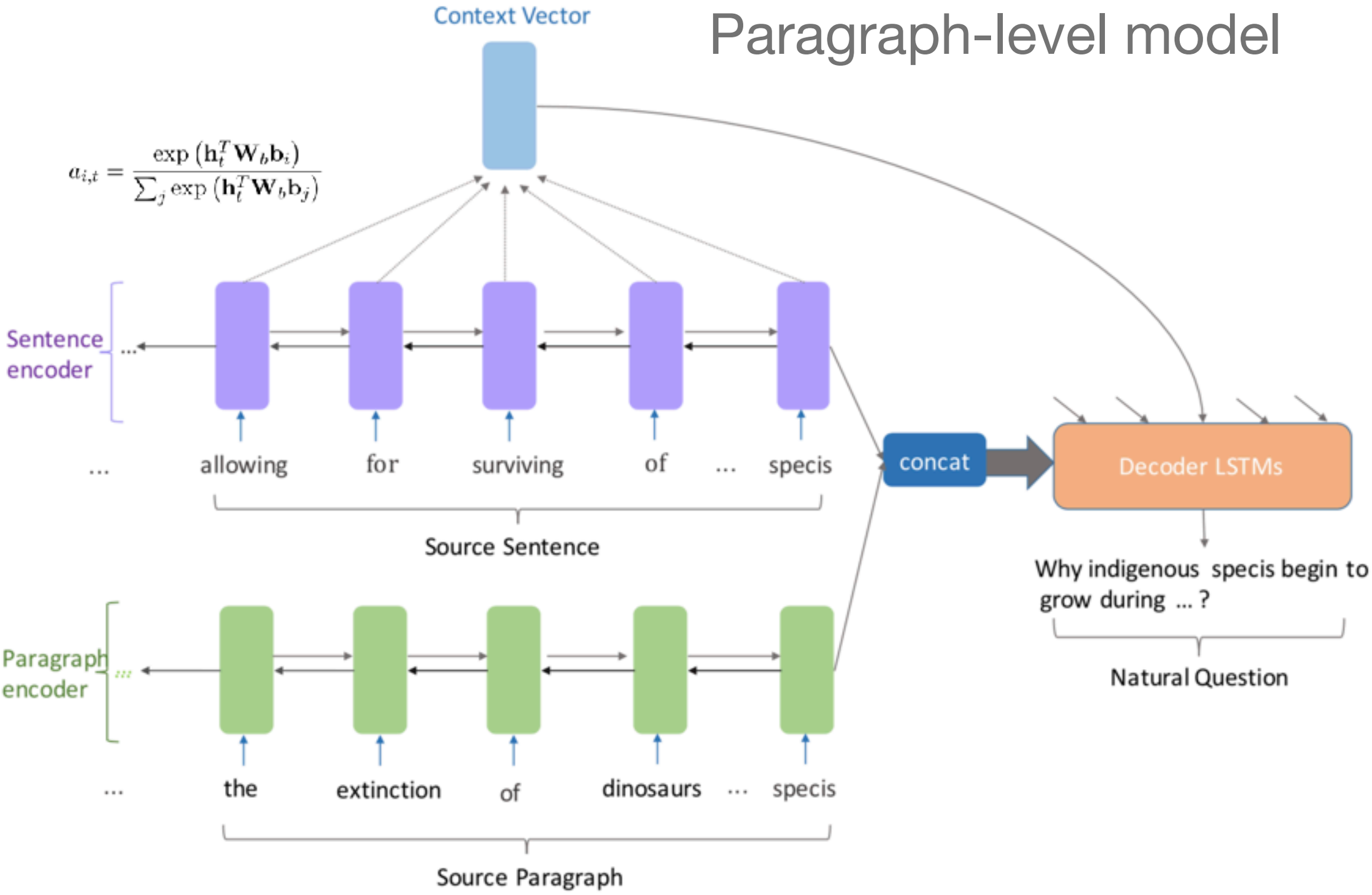
Output/State

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



Model

Paragraph-level model



Model

- Training:
  - Minimize the negative log-likelihood w.r.t θ.
  - That is,

$$\mathcal{L} = -\sum_{i=1}^{S} \log P\left(\mathbf{y}^{(i)}|\mathbf{x}^{(i)};\theta\right)$$

$$= -\sum_{i=1}^{S}\sum_{j=1}^{|\mathbf{y}^{(i)}|} \log P\left(y_j^{(i)}|\mathbf{x}^{(i)}, y_{<j}^{(i)};\theta\right)$$

  - UNK handling: replace it with the token with highest attention.

Outline

- Introduction
- Definition
- Model
- Experiment
- Conclusion

Experiment

- Dataset:
  - Stanford Question Answering Dataset (SQuAD)
  - SQuAD overcomes small size and synthetic issues.
- Preprocessing:
  - CoreNLP (tokenization & splitting)

Experiment

- A training technique:
  - Initial learning rate of 1.0
  - Starting halve the learning rate at epoch 8

Experiment

- Evaluation:
  - Automatic:
    - **BLEU:** avg. n-gram precision, short snt. penalty
    - **ROUGE:** recall for gold sentence, L for LCS.
    - **METEOR:** recall-oriented, considering synonyms, stemming & paraphrases.
  - Human: naturalness, difficulty. (range from 1 to 5)

Experiment

- Results: auto-

Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	METEOR	ROUGE_L
IR_{BM25}	5.18	0.91	0.28	0.12	4.57	9.16
$\text{IR}_{\text{Edit Distance}}$	18.28	5.48	2.26	1.06	7.73	20.77
MOSES+	15.61	3.64	1.00	0.30	10.47	17.82
DirectIn	31.71	21.18	15.11	11.20	14.95	22.47
H&S	38.50	22.80	15.52	11.18	15.95	30.98
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

Glove

Experiment

- Results: human

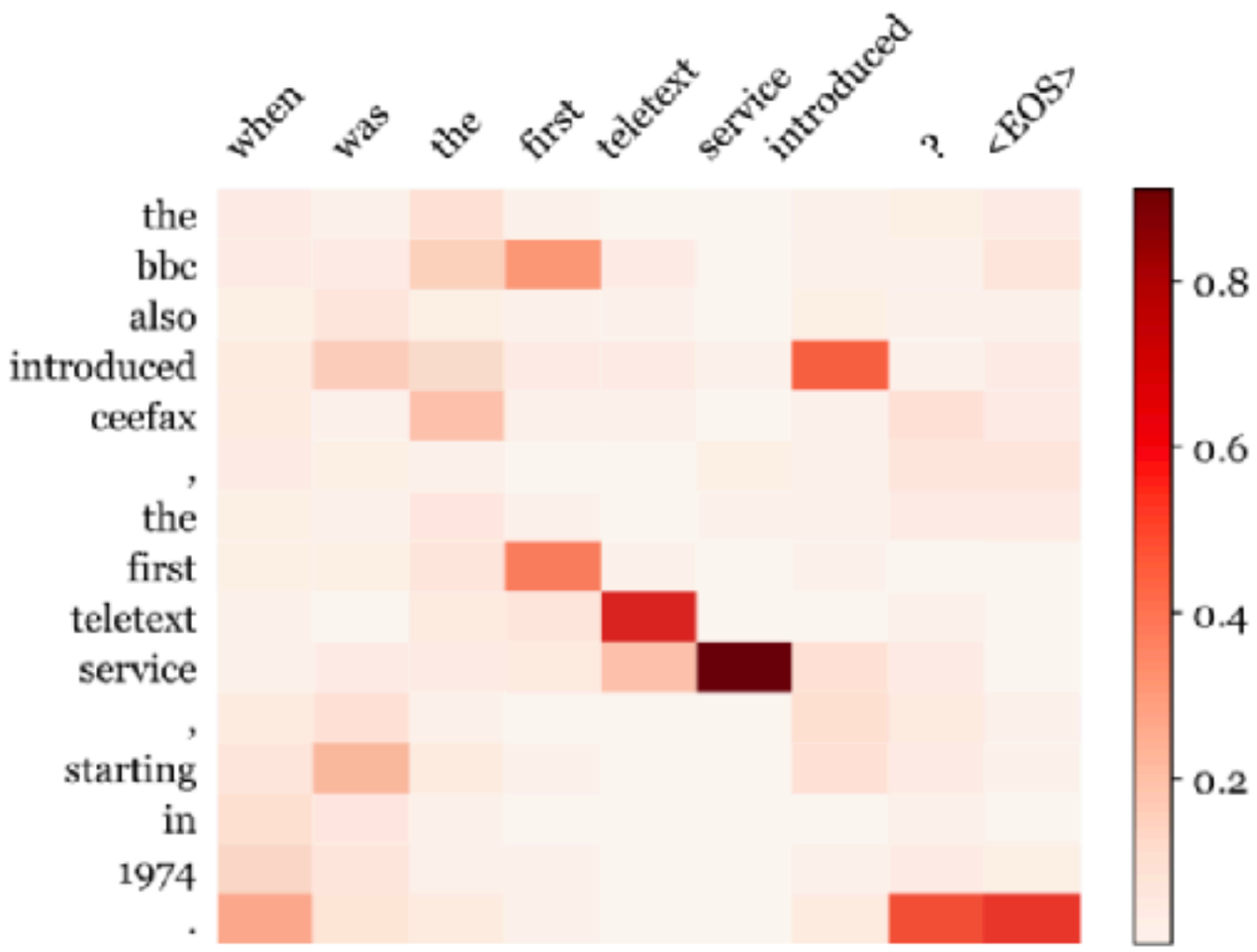
	Naturalness	Difficulty	Best %	Avg. rank
H&S	2.95	1.94	20.20	2.29
Ours	**3.36**	**3.03***	**38.38***	**1.94****
Human	3.91	2.63	66.42	1.46

Random sampling 100 sentence-quetion pairs

Experiment

Category	(%)	H&S			Ours			Ours + paragraph		
		BLEU-3	BLEU-4	METEOR	BLEU-3	BLEU-4	METEOR	BLEU-3	BLEU-4	METEOR
w/ sentence	70.23 (243)	20.64	15.81	16.76	**24.45**	**17.63**	17.82	24.01	16.39	**19.19**
w/ paragraph	19.65 (68)	6.34	< 0.01	10.74	3.76	< 0.01	11.59	**7.23**	**4.13**	**12.13**
All*	100 (346)	19.97	14.95	16.68	23.63	**16.85**	17.62	**24.68**	16.33	**19.61**

Experiment


when
was
the
first
teletext
service
introduced
?
<EOS>
the
bbc
also
introduced
ceefax
,
the
first
teletext
service
,
starting
in
1974
.
0.8
0.6
0.4
0.2

Conclusion

- A fully data-driven neural networks approach to automatic QG for reading comprehension.
- The best model they presented had achieved the state-of-the-art performance.

Fin.