Automatic Quiz Generation from Textbooks

Ethan A. Chi, Jillian Tang, Emily Wen

{ethanchi, jiltang, ewen22}@stanford.edu

Background

- Educational information is most commonly found in the form of textbooks; however, textbooks are often densely written and difficult to understand.
- Our goal is to generate reading comprehension questions from textbook passages that a) are nontrivial, b) are answerable upon reading the text, and c) test the reader on understanding of a central question of the passage.

Datasets

- We evaluate our quality of our models on the reverse AllenAI Textbook Question Answering dataset [1].
- Since our dataset size is small, we train our machine learning-based models on the reverse Stanford Question Answering Dataset [2], which has over 100,000+ questions. We then perform zero-shot evaluation on the textbook dataset.

Example

- **Text:** *About half the energy used in the U.S. is used in homes and for transportation.*
- **Target Question:** *What are the main ways energy is used in the U.S.?*
- **Text:** *A chemical bond is a force of attraction between atoms.*
- **Target Question:** *What is a chemical bond?*

Baselines

- **LastSent:** choose last sentence in the source text
- **LongestSent:** choose longest sentence in the source text

Neural Models

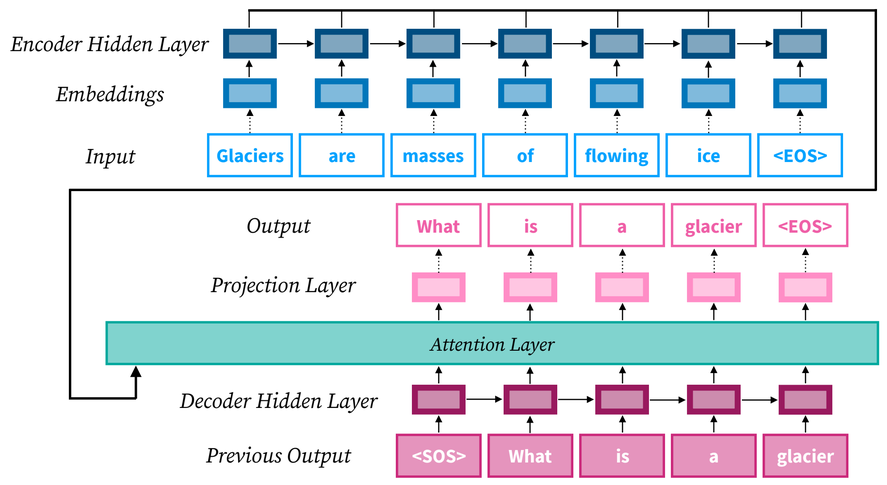
- We use pretrained 100-dim GloVe embeddings to embed input tokens.
- For sequential input, we use **LSTMs** (long short-term memory networks).
- At each timestep, a single input token is fed through a cell to generate another hidden state. We decode using **beam search** (beam size=5).

Model 1: Sequence to Sequence (Seq2Seq)

- We jointly train **encoder** and **decoder** models, two unidirectional LSTMs. The final hidden state of the encoder is used to initialize the decoder.
- The decoder applies **global attention**. To decode output tokens, we concatenate the output of each LSTM cell with a context vector calculated from the encoder's hidden states weighted by the *attention distribution* (a weighting over the source words calculated by a dot-product similarity metric, which tells the decoder which input tokens to focus on.)

Model 2: Seq2Seq with Copy Attention

- Our first seq2seq model is unable to handle unseen words in the input, especially if the training and evaluation sets are from different domains.
- To avoid this issue, we implement a copy-attention network as described in See et al [3]. At each time step, the network calculates a probability p_{gen}, from which we sample whether to generate or copy a word drawn from the attention distribution. This allows for copying unseen words.



Results

- Evaluation metrics:
  - Bilingual evaluation understudy (BLEU): calculates fraction of n-grams that appear in the target
  - Recall-Oriented Understudy for Gisting Evaluation (ROUGE): ROUGE-1 (unigrams), ROUGE-2 (bigrams), ROUGE-L (LCS)

Model		BLEU	ROUGE-1	ROUGE-2	ROUGE-L
LastSent	SQuAD	5.83	0.28	0.11	0.19
	Textbook	2.01	**0.18**	0.03	0.13
LongestSent	SQuAD	7.83	0.28	0.11	0.22
	Textbook	3.49	0.17	0.04	0.13
Seq2Seq	SQuAD	6.21	0.36	0.09	0.32
	Textbook	2.67	0.16	0.04	0.14
Seq2Seq w/ Copy Attention	SQuAD	**8.47**	**0.39**	**0.12**	**0.34**
	Textbook	**4.22**	0.16	**0.04**	**0.14**

Analysis

- Zero-shot transfer to textbook domain was generally unsuccessful due to different style of questions
- Copula-based questions: our model often asks about the subject complement and not the subject:
  - **Text**: *Muscles are the main organs of the muscular system.*
  - **Target Question:** *What are muscles?*
  - **Predicted:** *what are the main organs of the muscular system ?*
- Difficulty with words not in the GloVe dataset (attention-based <UNK> replacement did not perform well):
  - **Text:** *Earths rotation influences their direction. This is called the Coriolis effect.*
  - **Target Question:** *what is the Coriolis effect?*
  - **Predicted:** *what is the effect effect ?*

Citations

[1] Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., & Hajishirzi, H. (2017). Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4999-5007).

[2] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

[3] See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.