Numerical Question Generation in Science

Harish Umasankar
Ruhul Ameen

Guide: Pawan Kumar
Panelists: Shantanav Chakraborty, Indranil Chakrabarty

Objective

- To develop a model that can automatically generate numerical questions in science based on a given theory.
- To explore different approaches of modelling relationships between context and questions to improve diversity of question generated
- To assess the quality and diversity of questions generated

Why We Chose Diffusion Models For This task?

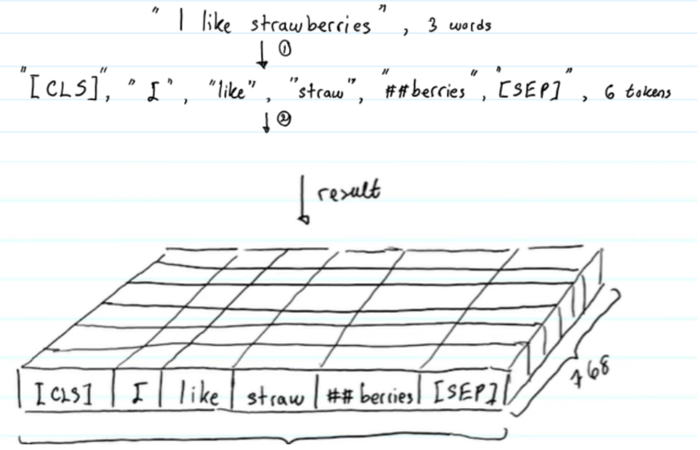
- Generate diverse questions
  From one topic, the model must generate different types of questions.
  For eg: For context as description of Ohms Law, the various questions that can be generated are
  1. Calculating Voltage when Resistance and Current are given
  2. Calculating Current when Voltage and Resistance are given
  3. Calculating Resistance when Voltage and Current are given
- Conditional nature
  Conditional modelling nature of the DiffuSeq is key to our objective of generating questions based on the context.
- Flexibility and adaptability
  We expect the model to generate questions from different topics of Science - Electricity, Magnetism, Acids and Bases etc.

Challenges In Using Diffusion Models on Text

- Image is Continuous Data
  - Image data is continuous in nature because it represents a continuous range of pixel values.
- Text is Discrete Data
  - Text data is discrete in nature because it consists of discrete symbols, such as words or characters.

Text Embedding: BERT

- Tokenization
- Adding Special Tokens
  - CLS
  - SEP
  - PAD
- Encoding

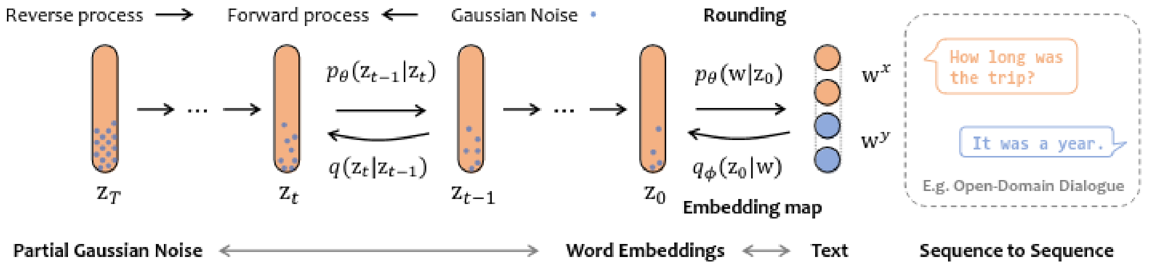


Model Architecture: Text Embedding

- Map the discrete input text data to a continuous space using an embedding function
  - A pair of sequences - w^x, w^y are concatenated = $w^{x \oplus y}$
  - EMB($w^{x \oplus y}$) = [EMB(w_1^x), EMB(w_2^x), ..., EMB(w_m^x), EMB(w_1^y), EMB(w_2^y) ... EMB(w_n^x)]
  - m is length of sequence w^x and n is length of sequence of w^y

Model Architecture: Forward Process with Partial Noising

- For each forward step, $q(z_t|z_{t-1})$, we gradually inject noise into z_{t-1} to get z_t
- Conventional Models: Corrupt whole z_t
- Partial Noising: Corrupt only y_t



Model Architecture: Reverse Process with Conditional Denoising

- Goal: Recover z_0 by denoising z_t
- Diffusion model will learn the denoising process $p_\theta(z_{t-1}|z_t) = N(z_{t-1}; \mu_\theta(z_t), \sigma_\theta(z_t))$
- The input will condition the denoising process.
- We train the model by maximizing the likelihood of observed data given the parameters
- Classifier Free Guidance

$$\nabla_x \log p_\gamma(x \mid y) = (1-\gamma)\nabla_x \log p(x) + \gamma \nabla_x \log p(x \mid y)$$

  - x - input
  - y - class label/text sequence
  - γ - guidance scale

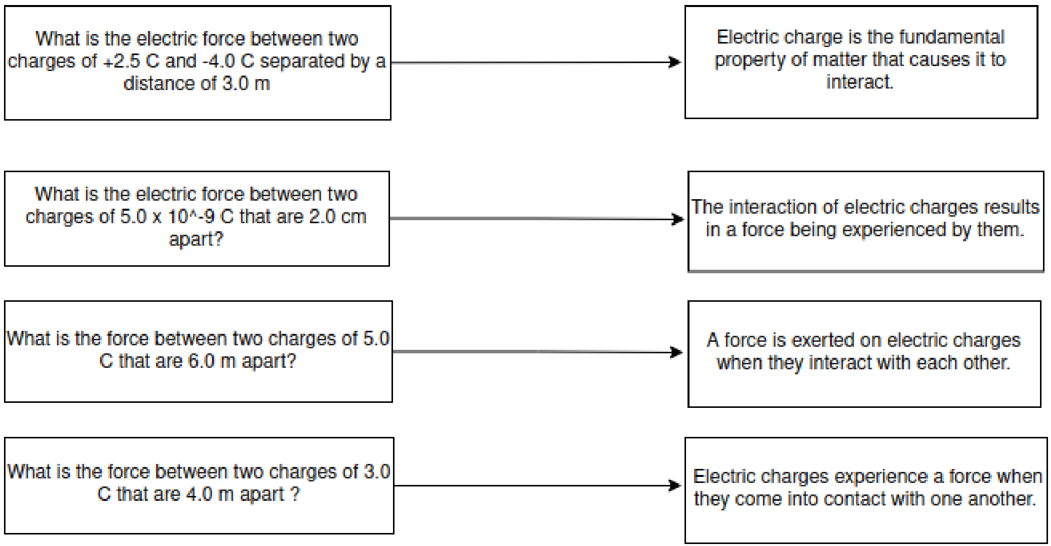
Dataset

- OpenQA dataset - contains 117k question-context pairs in English natural language. Used to train the model on English language.
- Numerical Question datasets - for our specific purpose contains numerical question-context pairs in certain topics of science. These datasets are created by us. It is done in two ways.
  - Forward mapped dataset
  - Forward + Reverse mapped dataset

Dataset

Forward mapped dataset

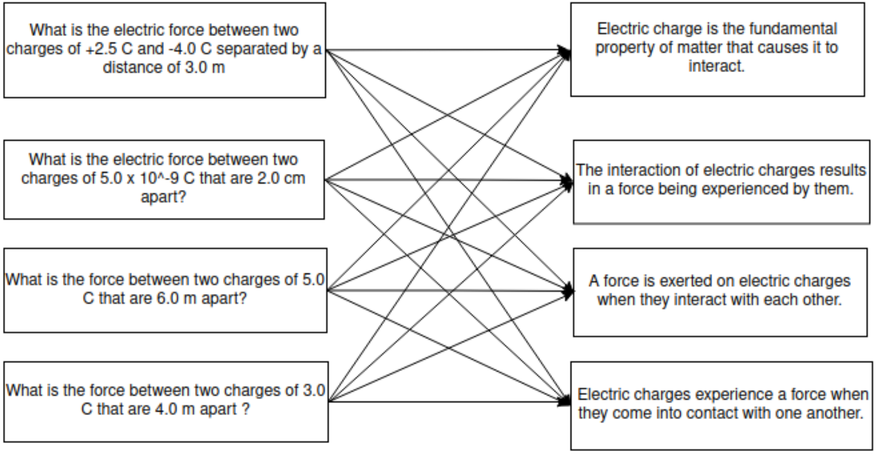
- Contains 20 unique contexts for each science topic.
- 20 different numerical questions for each context.
- Totally, 200 question-context pairs for each science topics.



DATASET

FORWARD + REVERSE MAPPED DATASET

- Contains 40 contexts for each science topic. In sets of 4 contexts, where each context is paraphrased versions of each other
- 20 different numerical questions for each context set.
- Totally, 800 question-context pairs for each science topics.



Loss

- Goal: Maximize the log-likelihood of the data given the noise-corrupted samples
- But computing the log-likelihood involves integrating over all possible paths of the diffusion process, which is computationally expensive
- Solution: Use a variational lower bound (VLB) on the log-likelihood, which can be optimized
- By optimizing the VLB instead of the actual log-likelihood, we can train the model more efficiently and effectively.

Result

- Electricity

Context	the relationship between current and voltage in a conductor is described by ohm's law, which states that the current is proportional to the voltage
Question	what current for voltage 2. 5 V through a wire with a resistance of 7 ohms?

Context	the relationship between current and voltage in a conductor is described by ohm's law, which states that the current is proportional to the voltage
Question	required voltage to produce a current of 12. 5 amps through a wire with a resistance of 7 ohms?

RESULT

- Acceleration

Context	acceleration is rate of change of velocity with respect to time
Question	what is acceleration is train to from 10 m/s to 20/s in 20 seconds?

Context	the acceleration due to gravity varies depending on the planet or celestial body being considered
Question	if a ball is at height 1 km, what is time to freefall due to gravity in s ?

Metrics for Evaluation

- **Analyze Quality:** BLEU Score
  - Measures the similarity between the machine-generated text and the reference question generated
  - Higher BLEU Score ensures more quality
- **Analyze Diversity:** Self BLEU Score
  - Computes the BLEU score between an n-gram of the generated text and all other n-grams in the text except for itself
  - Lower Self BLEU Score ensures more diversity

Score

Model	BLEU	Self BLEU
DiffuSeq - Question Generation	0.1731	0.2732
DiffuSeq - Numerical Science Question Generation	0.1775	0.029

Challenges Faced

- Initially, we tried with the pre-trained model RoBERTa but it couldn't capture the numerical structure of the questions.
  - Solution: Moved to DiffuSeq model
- DiffuSeq did not give good results when trained on our science question dataset
  - Solution: Used Transfer Learning Approach
- Training DiffuSeq Model gave memory error frequently
  - Solution: Reduced training parameters like batch-size and n-proc-nodes

References

- Neeraj Kollepara and Snehith Kumar Chatakonda and Pawan Kumar, "SCIMAT: Science and Mathematics Dataset".
- Gong, Shansan and Li, Mukai and Feng, Jiangtao and Wu, Zhiyong and Kong, Lingpeng, "DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models".
- Xiang Lisa Li and John Thickstun and Ishaan Gulrajani and Percy Liang and Tatsunori Hashimoto, "Diffusion-LM Improves Controllable Text Generation".