Question Generation Model for Educational Purposes

Joshua Fantillo, Amin Jamalkhah

Simon Fraser University

{jmf23,aja142}@sfu.ca

Contributions

- Designed a novel question generation pipeline using a fine-tuned T5 model for educational purposes.
- Leveraged the SQuAD dataset for robust context-question pair preparation and fine-tuning.
- Implemented preprocessing scripts to process raw text and PowerPoint slides into structured datasets.
- Fine-tuned T5 with custom datasets and optimized model performance through prompt engineering.
- Evaluated models with BLEU, ROUGE, semantic similarity, and grammar metrics, achieving significant improvements.
- Developed user-friendly tools for generating questions from user-provided text or PowerPoint inputs.

Proposed Method

- **SQuAD:**
  Input: Context, Question → Answer

Context:
The SQuAD dataset is a collection of question-answer pairs. The questions in the dataset are based on a set of Wikipedia articles. Each question is crafted such that it requires the model to extract a span from the passage.

Question:
What is the SQuAD dataset?

Answer:
The SQuAD dataset is a collection of question-answer pairs.

- **Proposed: Question Generation**
  Context → Question

Context:
Albert Einstein was a theoretical physicist who developed the theory of relativity. He is widely known for his mass-energy equivalence formula.

Question:
Who developed the theory of relativity?

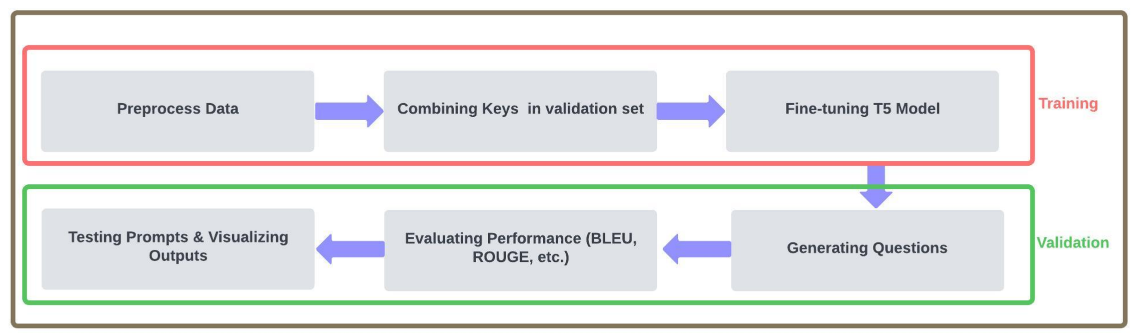
Dataset

- **Dataset Overview:**
  - SQuAD: Stanford Question Answering Dataset.
  - 100,000 question-answer pairs across diverse articles.
  - Suitable for training models in question generation and comprehension.
- **Data Preparation and Grouping:**
  - Grouped the dataset by context as the unique key.
  - Aggregated questions associated with each context into a single entry.
  - This key-pairing approach simplifies model training and ensures contextual alignment.

Example Dataset Format

```
{ "context": "Some sample context" [
"This is Question 1?",
"This is Question 2?",
"This is Question 3?"
] }
```

Process



- **Step 1: Data Pre-Processing**
  - Prepare raw datasets for model input.
  - Remove invalid or incomplete entries.
  - Group contexts and their related questions.
  - Format data to fit the fine-tuning requirements.
- **Step 2: Fine-Tuning the Model**
  - Train the T5 model with SQuAD dataset variants.
  - Implement prompt-based variations for diverse outputs.
- **Step 3: Model Evaluation**
  - Evaluate the model using BLEU, ROUGE, and semantic similarity.
  - Analyze grammar consistency in generated questions.
- **Step 4: Question Generation**
  - Generate new questions based on user-provided context (text/PPT).
  - Provide diverse output with various prompt types.

Examples

Fine Tuned Model Sample Validation and Generated Question

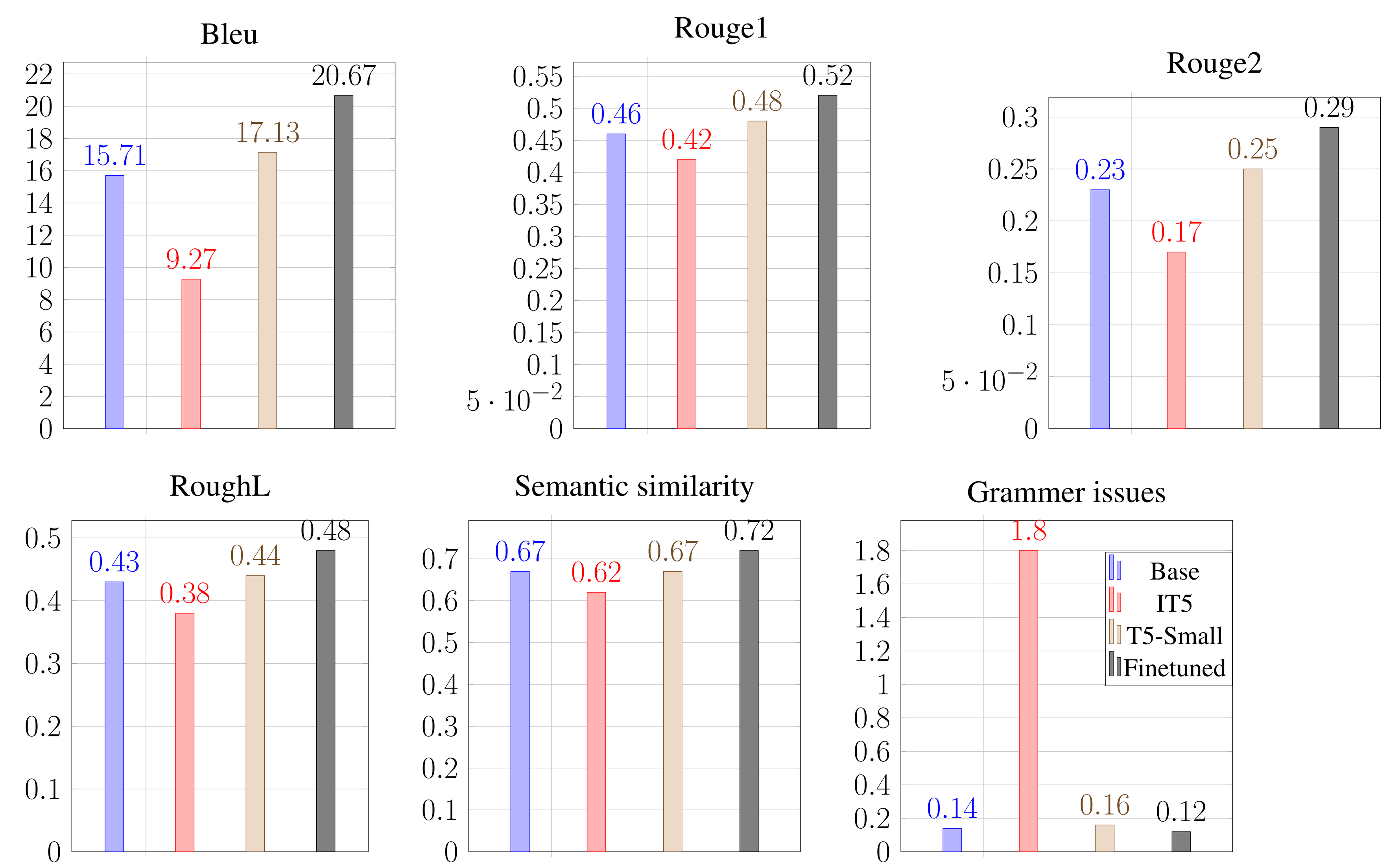
```
{
"context": "Religious and spiritual teachers, such
as gurus, mullahs, rabbis, pastors/youth pastors and
lamas, may teach religious texts such as the Quran,
Torah or Bible.",

"ground truth questions": [
"What is another general name for a religious
teacher?",
"What type of text is the Quran?",
"Name a text that might be used by a religious
teacher to teach.",
"Who would generally teach from a work like the
Quran, Torah or Bible?"
],

"generated question": "What are some religious
and spiritual teacher names?"
}
```

Prompt Variants Sample

```
prompt_variants = [
"Generate a detailed question based on the context",
"Generate a short question based on the context",
"Generate a creative question based on the context",
"Generate a factual question based on the context:",
]
```

Results

- **Performance Comparison:**
  - The fine-tuned model outperformed the base model across almost all metrics.
  - Achieved a **32% increase in BLEU score**, indicating better alignment with ground-truth questions.
  - ROUGE scores improved significantly: **ROUGE-1 (+11%)**, **ROUGE-2 (+26%)**, and **ROUGE-L (+12%)**.
  - Semantic similarity increased by **7%**, demonstrating enhanced contextual relevance in the generated questions.
  - Grammar issues were reduced by **9%**, though there is still room for further improvement.
- **Performance Visualization:**
  - Compared models: Base model, IT5 model, T5-small-SQuAD2-question-generation, and the fine-tuned model.
  - Metrics analyzed: BLEU, ROUGE (1/2/L), Semantic similarity, and Grammar quality.



References

- 1.
  Asahi Ushio and Fernando Alva-Manchego and Jose Camacho-Collados Generative Language Models for Paragraph-Level Question Generation
- 2.
  Sahan Bulathwela, Hamze Muse and Emine Yilmaz Scalable Educational Question Generation with Pre-trained Language Models
- 3.
  Bidyut Das, Mukta Majumder, Santanu Phadikar Arif Ahmed Sekh Automatic question generation and answer assessment: a survey