Lab Assignment – 4
Conditional Text Generation with Context

Dataset

Name: Stanford Question Answering Dataset (SQuAD v1.1)

Size: 87,599 training examples, 21,900 validation examples, 10,570 test examples

Preprocessing Steps:

- Loaded dataset using Hugging Face datasets library
- Split training data into 80% train / 20% validation
- Formatted input as "answer_question: [context]" and target as questions
- Tokenized using T5Tokenizer with max length 512 for context, 128 for questions
- Saved as CSV files for training pipeline with proper input-target formatting

Model

Pre-trained Model: T5-small (Text-to-Text Transfer Transformer)

Configuration:

- 60M parameters encoder-decoder architecture
- Sequence-to-sequence conditional generation model
- Fine-tuned for question generation from context passages
- Optimized for Apple MPS device (M1/M2 MacBook Pro)
- Legacy tokenizer behavior disabled for improved performance

Training Details

Batch Size: 8 per device (train and evaluation)

Epochs: 3 epochs with epoch-based evaluation and saving

Optimizer: AdamW with 3e-4 learning rate

Training Strategy:

- Evaluation after each epoch instead of step-based
- Model checkpoints saved after each epoch
- Best model selection based on validation loss
- DataCollatorForSeq2Seq for dynamic padding and MPS compatibility

Generation

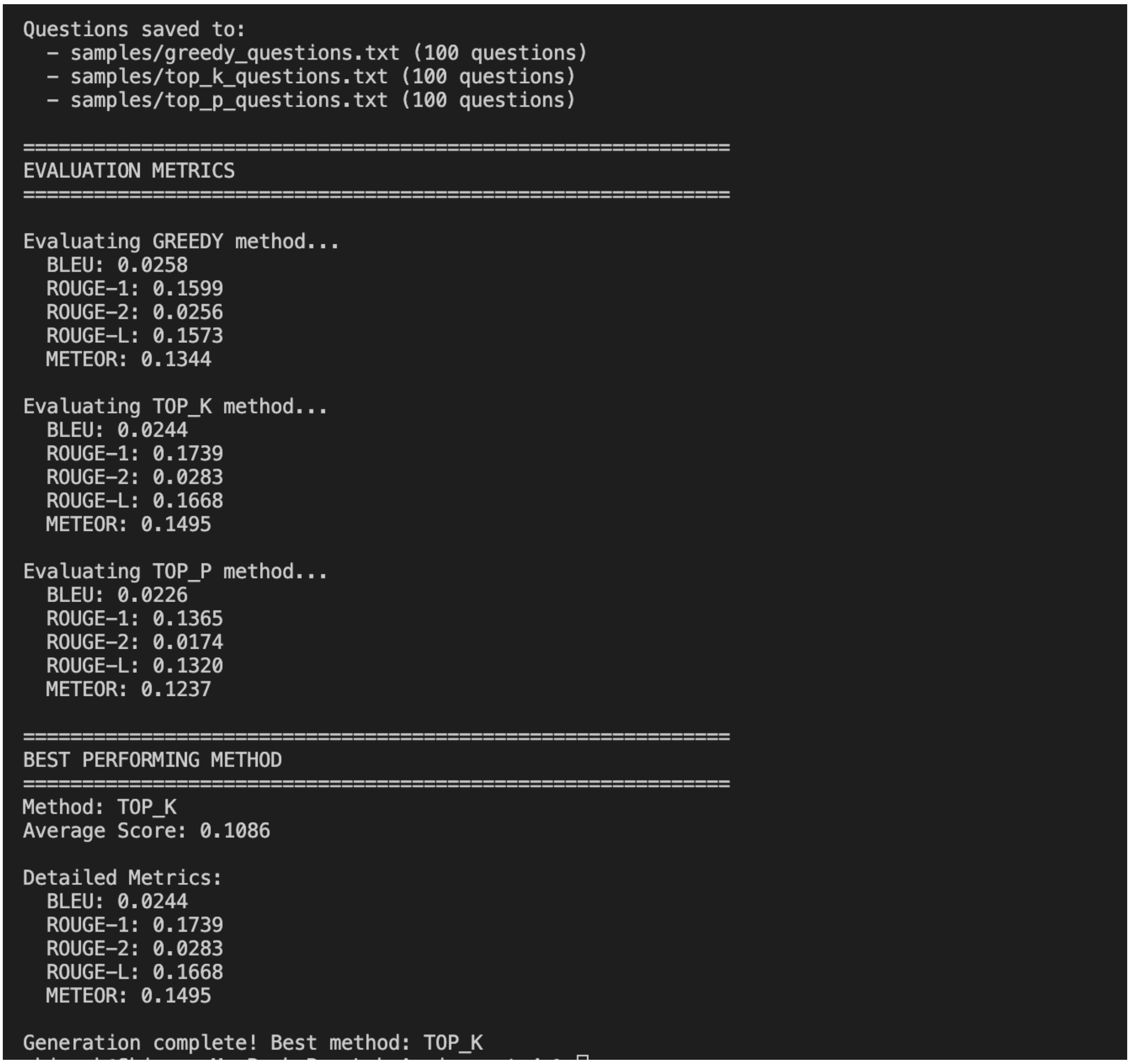
Prompt Format: "answer_question: [context]" → questions

Decoding Strategies:

1. **Beam Search (Greedy):** num_beams=10, deterministic diverse outputs
2. **Top-k Sampling:** k=50, temperature=0.7, stochastic sampling from top tokens
3. **Top-p Sampling:** p=0.9, temperature=0.8, nucleus sampling approach

- Generated 100 questions per method (300 total questions)
- Used 10 diverse test contexts covering multiple domains

Evaluation



```
Questions saved to:
  - samples/greedy_questions.txt (100 questions)
  - samples/top_k_questions.txt (100 questions)
  - samples/top_p_questions.txt (100 questions)

============================================================
EVALUATION METRICS
============================================================

Evaluating GREEDY method...
  BLEU: 0.0258
  ROUGE-1: 0.1599
  ROUGE-2: 0.0256
  ROUGE-L: 0.1573
  METEOR: 0.1344

Evaluating TOP_K method...
  BLEU: 0.0244
  ROUGE-1: 0.1739
  ROUGE-2: 0.0283
  ROUGE-L: 0.1668
  METEOR: 0.1495

Evaluating TOP_P method...
  BLEU: 0.0226
  ROUGE-1: 0.1365
  ROUGE-2: 0.0174
  ROUGE-L: 0.1320
  METEOR: 0.1237

============================================================
BEST PERFORMING METHOD
============================================================
Method: TOP_K
Average Score: 0.1086

Detailed Metrics:
  BLEU: 0.0244
  ROUGE-1: 0.1739
  ROUGE-2: 0.0283
  ROUGE-L: 0.1668
  METEOR: 0.1495

Generation complete! Best method: TOP_K
```

Observations

Strengths:

- Successfully implemented three distinct decoding strategies for question generation
- TOP_K sampling achieved best overall performance with balanced diversity and quality
- Stable training convergence with epoch-based evaluation on MPS device
- Generated contextually relevant questions across diverse domains

Challenges:

- Greedy decoding required beam search modification for multiple sequence generation
- Low absolute metric scores indicate room for improvement in question quality
- Memory optimization needed for MPS device compatibility during generation