Question-Answering using BERT

Trevor Koenig, Pratim Ugale

University of Delaware

Abstract

We develop a system for Open Domain Question Answering using BERT

Model: BERT
Training: Fine-tuned on SQuAD 2.0
Evaluation: Manually evaluated against ChatGPT
Why: Lightweight, applicable to other domains

Motivation

- Large Language Models (LLMs) often face the following problems when answering factoid questions:
  - **Hallucination**
  - **Proprietary data**
  - **Expensive**

Problem Statement and Questions

Problem Statement

- Compare the Results of an open domain BERT-based QA model with the latent QA capabilities of ChatGPT

Research Questions

- Can BERT + Wikipedia produce answers of comparable quality to ChatGPT?
- Can BERT refrain from hallucinating?

Dataset

We will be fine-tuning a pretrained BERT model using the **Stanford Question Answering Dataset** (SQuAD 2.0)

- What is SQuAD 2.0?
  - JSON format includes Prompt, Context and Answer among other fields
  - Includes answerable and unanswerable questions
  - Designed for span-based QA: the model finds the exact span of text in a passage that answers the question

Methods

Training:

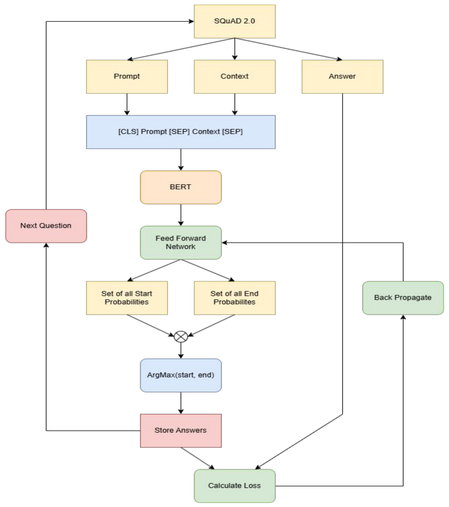
- Fine-tune BERT on SQuAD 2.0

Open Domain Question Answering Pipeline:

- Input: Natural language question
- Retriever:
  a. Wikipedia Search: Query via API
  b. Article Retrieval: Top N articles
  c. Chunking: Split text to fit BERT (≤512 tokens)
- Reader
  a. Formatting: [CLS] Question [SEP] Context [SEP]
  b. Run BERT on each chunk
  c. If answerable → extract span, else skip

ChatGPT Comparison:

- Ask same question in ChatGPT without context
- Evaluate manually for correctness and reliability



Training Pipeline (for Closed-Domain QA Model)

Modifications for improving 1st Method above:

1. Method 2: Normalization across sentences by feeding good candidates together
2. Method 3: Summed Softmax

Setup and Evaluation

- Train: A100 GPUs on Google Colab
- Test
  - Test Set: Manually generated - 50 Questions
  - We use top N = 3 articles across all results
  - Limit the open-domain test questions to factoid
  - Manually check if the answer highlighted by BERT, and generated by GPT is logically the same

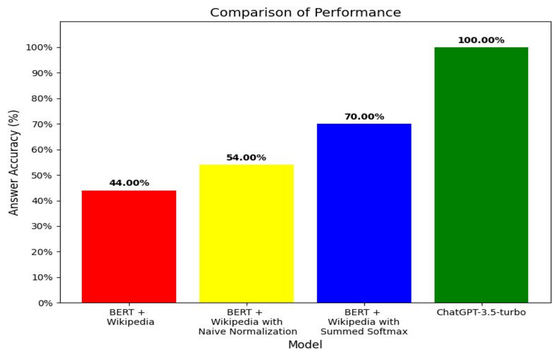
Results

Training Accuracy across Epochs = 3

Epoch	Training Loss	Validation Loss	Start Accuracy	End Accuracy
1	1.054300	1.057407	0.670092	0.678262
2	0.765200	1.122554	0.677588	0.688790
3	0.570500	1.244623	0.679609	0.688621

Selection Process:

- BERT + Wiki
- BERT+ Wiki with Naive Normalization across Best passages
- BERT + Wiki: Summed Softmax
- ChatGPT



Discussion

- Probabilities not initially normalized across chunks - leading to incorrect final answers although correct were considered
- Results dependent on Wikipedia's retrieval mechanism
- BERT + Wiki works well on current affairs, for example:

```
question = "Who was elected as the new Pope in the May 2025 Papal conclave?"

answer = ask_question_to_wiki(question)
```

```
Found the following relevant Wikipedia documents:
['2025 papal conclave papabili', 'Cardinal electors in the 2025 papal conclave', '2025 papal conclave']

Question: Who was elected as the new Pope in the May 2025 Papal conclave?
Best answer: Robert Prevost
```

- BERT method uses more tokens then GPT if they refer to the same document due to overlapping strides
- More specific questions are always helpful - BERT might not be knowing general implicit context. For example:

```
question = "Who invented the telephone?"

answer = ask_question_to_wiki(question)
```

```
Found the following relevant Wikipedia documents:
1: Telephone
2: The Telephone Cases
3: Antonio Meucci
Question: Who invented the telephone?
Best answer: Captain John Taylor
```

```
question = "Who invented the first electric telephone?"

answer = ask_question_to_wiki(question)
```

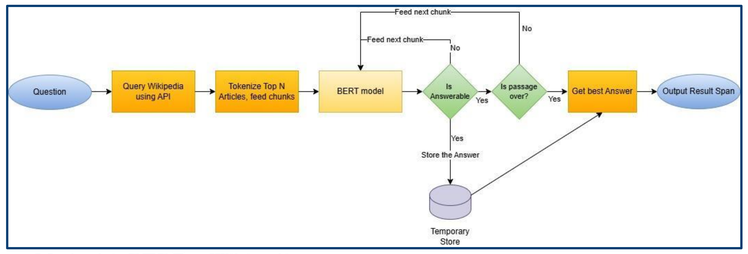
```
Found the following relevant Wikipedia documents:
1: History of the telephone
2: Invention of the telephone
3: Telephone
Question: Who invented the first electric telephone?
Best answer: Alexander Graham Bell
```

- When categories are too large (all animals, all countries) performs poorly. Performs better on smaller domains.

Conclusion and Future Work

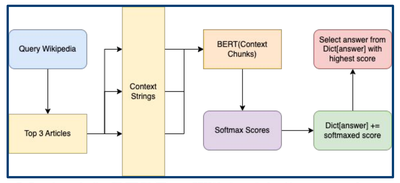
- BERT gave comparable results after some modifications **Best Method: Summed Softmax**
- With more ad-hoc improvements, the model could become even more accurate.

Open-Domain QA Pipelines



Method 1: BERT + Wikipedia

Method 2: BERT + Wiki with Naive Normalization



Method 3: BERT + Wiki with Summed Softmax

References

- Daniel Jurafsky and James H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition.
- MediaWiki API: https://wiki.creativecommons.org/api.php?action=help&modules=query%2Bsearch
- Wang, Zhiguo, et al. "Multi-passage BERT: A globally normalized BERT model for open-domain question answering." arXiv preprint arXiv:1908.08167 (2019).