Name: Cody Yu, Pon Swarnalaya Ravichandran, Ei Tanaka
Class: DATS 6312 - 10 Natural Language Processing
Date: November 5, 2023
Topic: https://www.kaggle.com/datasets/stanfordu/stanford-question-answering-dataset

Final Project Proposal - Group 7

Project Overview

- A meticulously curated set of question-answer pairs that have been skillfully extracted from a wide range of Wikipedia articles is called the Stanford Question Answering Dataset (SQuAD). SQuAD is unique because it defines correct answers with remarkable flexibility, covering various token sequences in the provided text. Crowdsourcing initiatives have allowed for the laborious assembly of human-crafted questions and answers, resulting in inclusivity.
- The dataset is a flexible yet complex resource for question-answering research, with over 100,000 question-answer pairs from over 500 articles on average. Custom models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) address the intricacy of SQuAD because they are superior at understanding context and semantics at a deeper level. The SQuAD dataset can be used to optimize these models' performance when answering questions.
- Several packages, including Transformer, PyTorch, Spacy, and NLTK, support text processing and feature extraction. For question-answering tasks, two main paradigms—Knowledge-Based QA and Information Retrieval (IR) based QA—are examined, emphasizing precise and effective factoid question answering. Metrics such as Mean Reciprocal Rank (MRR), Exact Match, and F1 Score are used to evaluate these systems to determine the relevance and accuracy of the model's responses. This all-encompassing method seeks to improve machine learning models for answering questions while advancing natural language understanding.

About Dataset

- The Stanford Question Answering Dataset (SQuAD) is a carefully selected collection of question-answer pairs that have been carefully removed from an extensive body of Wikipedia articles.
- The SQuAD framework allows for extensive latitude in defining correct answers, which includes a variety of token sequences in the given text. The human-crafted, painstakingly put-together answers and questions via crowdsourcing initiatives are responsible for this broad inclusivity. Because of this special quality, SQuAD stands out from some other question-answering datasets.
- SQaAD dataset typically contains 100,000+ question-answer pairs on 500+ articles.

Method

- Customized Models: BERT(Bidirectional Encoder Representations from Transformers) GPT(Generative Pre-trained Transformer)

Why Customized Models?

Compared with classical Models such as BOW, TF-IDF, or LSTM, the SQuaAD dataset is too complex to understand the context and semantics at a deeper level. BERT(Devlin J, 1970) is designed to understand the context of a word in a sentence. It is essential for this answering questions project, where the meaning of a word could change based on its surrounding words. Another benefit of these customized models is the Fine-Tuning Capability. With these pre-trained models, they can be fine-tuned with the SQuAD dataset, and Fine-tuning also allows the model to adjust its parameters specifically for the task of question answering.

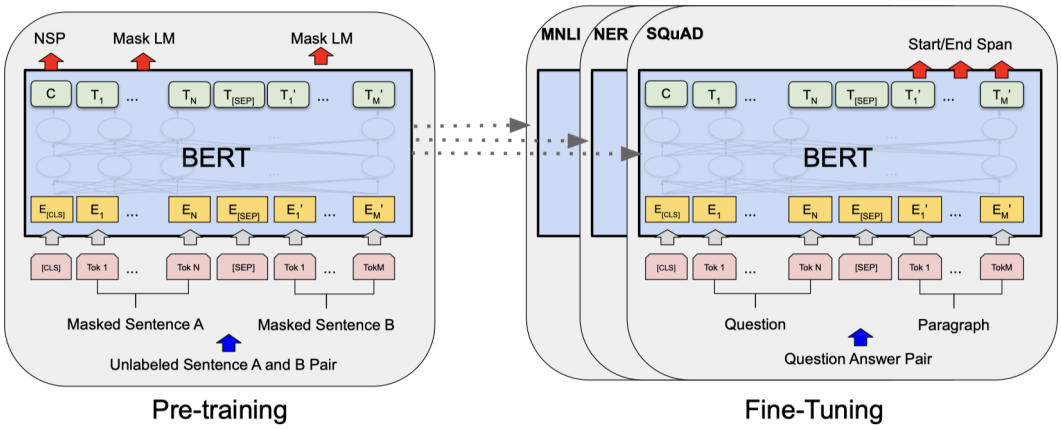

Figure 1: Overall pre-training and fine-tuning procedures for BERT.

Packages

- Transformer, PyTorch or TensorFlow, Spacy, NLTK

Why use these packages?

Transformer: It offers pre-trained models and provides functionalities for fine-tuning these models.

PyTorch: Primary Deep Learning Framework.

Spacy: Text Preprocessing and feature extraction.

NLTK: tokenization, stemming, and other NLP tasks

NLP Tasks

As outlined by Jurafsky and Martin (2023), there are two major paradigms for addressing question-answering (QA):

1. Information-Retrieval (IR) Based QA (Open Domain QA): This approach involves retrieving relevant passages or documents in response to a user's question. It then employs neural reading comprehension algorithms to interpret these passages and extract direct answers. A key challenge in this paradigm is handling the vocabulary

mismatch problem, where the precision of answers depends on the overlap of terminology between the question and the source material. To overcome this, modern approaches often use dense embeddings like BERT, which effectively understand contextual meanings and synonyms.

2. Knowledge-Based QA: This paradigm leverages structured databases or knowledge bases to derive answers. Here, the focus is on building semantic representations of queries, such as mapping them to logical forms or identifying relationships. These representations are then used to query structured sources of facts. A critical component in this approach is entity linking, which involves associating mentions in the text with real-world entities in a knowledge base, like Wikipedia. This process often involves stages of mention detection and disambiguation to identify and link entities accurately.

Our goal is to develop a system that can efficiently and accurately provide answers to factoid questions, harnessing the strengths of both IR-based and knowledge-based approaches or (just either one due to time restriction).

Performance of the model/metrics

In evaluating factoid question-answering systems, Jurafsky and Martin (2023) suggest using the following metrics:

- Mean Reciprocal Rank (MRR): This measures how well a system ranks the correct answer within its top responses. For each question, the reciprocal rank of the first correct answer is calculated. The MRR is then the average of these values across all test questions.

Reading Comprehension Metrics:

- Exact Match: The percentage of system answers that match the gold-standard answers.
- F1 Score: An average measure of the word/token overlap between the predicted and gold-standard answers, allowing for partial credit for near-correct solutions.

These metrics assess the accuracy and the relevance of the answers provided by the system.

Time Schedule

- Week 1 (Oct 29 ~ Nov 4): Final Project Proposal, Literature Review
- Week 2 (Nov 5 ~ 11): Data Loading, EDA, Data Preprocessing, (Final Project Report)
- Week 3 (Nov 12 ~ 18): Model Training, (Final Project Report), Hyperparameter Tuning
- Week 4 (Nov 19 ~ 25) Fall Break
- Week 5 (Nov 26 - Dec 3): Model Evaluation, Building the GUI with Stram (Final Project Report)
- Week 6 (Dec 4 ~ 11): Preparation for Presentation and Final Submission

GitHub Link: https://github.com/eitanaka/NLP_Final_Project_Group7

References

- Alberti, C., K. Lee, and M. Collins. 2019. A BERT baseline for the natural questions. https://arxiv.org/abs/1901.08634
- Chen, D., A. Fisch, J. Weston, and A. Bordes. 2017. Reading Wikipedia to answer open-domain questions. ACL. https://aclanthology.org/P17-1171.pdf
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. N. (1970, January 1). *Bert: Pre-training of deep bidirectional Transformers for language understanding*. Google Research. https://research.google/pubs/pub47751/
- Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing (3rd ed.). Chapter 14, Stanford University

Dataset

- Stanford Question Answering Dataset, Kaggle https://www.kaggle.com/datasets/stanfordu/stanford-question-answering-dataset