Project Proposal: Enterprise Domain Question & Answer

Motivation

- Task: Question Answering (QA) in one enterprise domain
- Focus: Usable QA tool for enterprise internal technical inquiry
- Reasons for Choice: Many large companies rely on frontline IT/HR/Finance human staff to handle internal employee inquiries, which are quite human resources intensive and have high answering latency. Of these internal inquiries, 70% of them are quite basic and repetitive. Technical Q&A is one big portion of these enterprise internal inquiries critical for company operation efficiency. Answering these basic inquiries with AI will shorten employee waiting time, reduce IT/HR/Finance human resources consumption and improve overall enterprise operation efficiency. We choose the IT domain to work out an AI tool to answer internal employee technical questions. If this works well, it could be easily deployed to HR/Finance domain by just following a similar approach and replacing the datasets as HR/Finance datasets. Developing internal usage AI tool will not only improve operation efficiency, but also benefit cybersecurity because question & answer contents within the enterprise are quite sensitive and need to be protected. Utilizing a separate internal AI tool rather than putting those questions to public open tools like Chat-GPT will protect the privacy of enterprises and eliminate diverse risks. The potential market for these highly customized enterprise internal AI tools could be billions of dollars.

Approach

- Potential Solution:
  - Step 1: Process one public technical QA dataset to make it suitable for in-context language understanding model training.
  - Step 2: Fine-tune a model based on a mid-size pre-trained language model to find the most relevant technical notes for each question. The relevance will be estimated by comparing the similarities of the inquiry and the tech notes title in the large technical notes pool. This will bring shortlisted technical notes for each question. The potential pre-trained model could be all-mpnet-base-v2. We could also try other models like all-distilroberta-v1, etc.
  - Step 3: Fine-tune a pre-trained model to find the short answer for each question according to the long contents of those most relevant technical notes. One potential pre-trained model is bert-large-cased. We could also try other similar-size models good for in-context understanding.
  - Step 4: With fine-tuned models, we could test whether the accuracy of answering, and the latency can fulfill a real usage scenario in the enterprise IT domain.

Dataset

- IBM TECHQA dataset, containing
  - a training set: 450 answerable questions and 150 non-answerable questions.
  - a development set: 160 answerable and 150 non-answerable questions.
  - a test set: 490 questions with similar answerable vs. non-answerable proportions as the development set
  - full collection of the unique 801, 998 Technotes that were available on the web as of April 4, 2019.

  The dataset is designed for machine reading comprehension, rather than for open-domain question answering, which is a quite similar dataset to an enterprise IT department technical support knowledge dataset.