CS7650 Term Project Proposal — Automatic Labeling of Google QUEST Dataset

James Hahn jhahn37@gatech.edu
Bharat Rahuldhev Patil rahuldhev@gatech.edu
Cyrille Combettes cyrille@gatech.edu

1 Motivation

Question-answering (Q&A) is the task of an intelligent system to, given a question, provide an appropriate query response. Recently, with the help of deep learning, this field of machine learning has expanded through varying dataset domains, including audio Q&A [1], visual Q&A [2], and text Q&A [3], and have led to exciting applications, e.g., textual customer support [4] and embodied Q&A visual navigation [5].

However, a major shortfall of current question-answering systems is their failure to deal with open-ended questions. In other words, factual, single-answer questions are their forte. In order to trend toward artifical general intelligence, systems should be taught to distinguish between the two types of questions and reply accordingly, allow them to reasoning about opinion-based or dynamic questions. In this sense, autonomous systems should be able to match human performance on subjective questions requiring additional thought and context.

The first step of this process is detecting the various characteristics of questions. Questions may require one-sentence answers or multi-sentence answers. Additionally, they may require additional, external knowledge graphs, as well as information about the expected answer (e.g., free-form, multiple choice, text, or count), or whether the question has a commonly-accepted answer (e.g., "How many presidents has the U.S. had throughout history?"). On the other side of the question-answering task, a dataset's labeled answers might require additional meta information, such as how satisfied the user was with the response, its fluency, or its relevance.

Unfortunately, no datasets exist to tackle this above challenge. They do not contain the level of detail needed for an insightful training algorithm capable of reasoning about the meta details of the question or answer. As such, no related work exists in this domain, and it is a novel challenge. Thankfully, online challenges exist, which we discuss below. As such, the greater goal of the proposed task is to predict the ground-truth values of these annotations for a dataset, which will hopefully assist in automatically labeling common, large-scale datasets, such as the Stanford Question Answering Dataset (SQuAD) [3]. The closest work to this task are other papers attempting to automatically label data [6, 7, 8, 9], however none of the papers are associated with question-answer metadata labeling.

2 Goal

Before we explore further details of the project, it is important to note this project is motivated by the Google QUEST Q&A Labeling task on Kaggle [10]. A general task overview can be seen in Figure 1. Google's internal CrowdSource team labeled 6079 training samples and 476 testing samples across 30 question and answer attributes. Each training sample consists of a textual question, corresponding textual answer, and 30 attributes in the range $[0, 1]$, assessing varying qualities of the question and answer (e.g., "question_fact_seeking", "question_multi_intent", "answer_relevance", and "answer_well_written"). Given a test sample containing a question and its corresponding answer, the task is to label the question and answer across these 30 attributes.

While the immediate task is accurately labeling question/answer attributes, we plan to carry out an additional task of semi-supervised learning by labeling SQuAD with these 30 attributes and attempting the question-answering task to verify whether these additional attributes boost a model's performance.

These two tasks go in tandem with each other, hopefully showcasing success on the ability for sys-

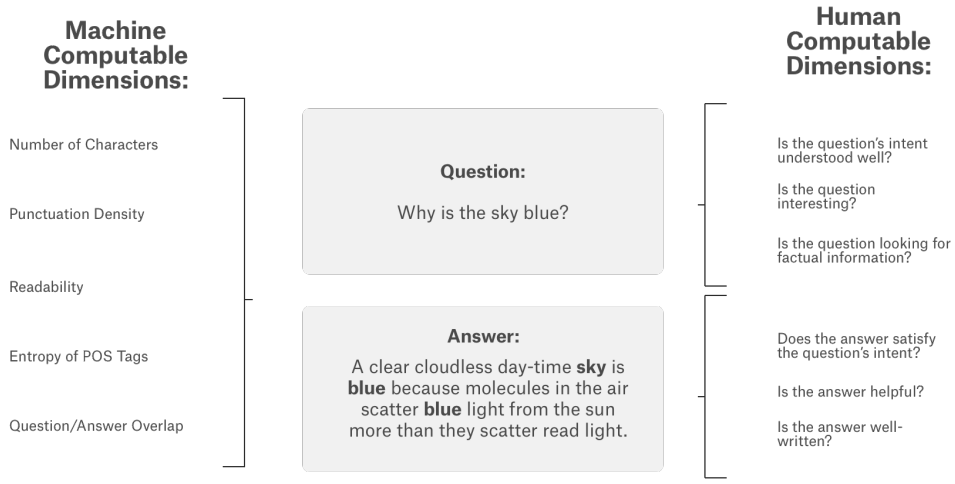

Figure 1: General task overview in the Google QUEST Q&A Labeling challenge on Kaggle [10].

tems to provide additional metadata for question-answer pairs, as well as providing verification of its usefulness by increasing performance on the current textual Q&A state-of-the-art on a pre-existing unlabeled (without metadata) dataset, such as SQuAD. In summary, this project has an initial 30-way prediction task with the additional proposition of utilizing the trained model to label SQuAD and evaluate on the question-answering task.

3 Plan

Fortunately, the dataset is already provided. We do not have to collect anything, as Google provides both the training and testing sets through Kaggle as csv files. The only downside is the test set does not have labels. As such, to evaluate our method, despite the Kaggle competition being closed, we will make a late submission to evaluate our approaches. It is important to note this is not a binary classification problem, but rather a 30-way regression task to predict values in [0, 1] for each attribute in the dataset. Additionally, each question-answer pair was labeled by several annotators, as to gather stable ratings, and several samples may contain the same question, but that is because they all contain different answers to the same question.

To evaluate on all 476 test samples, we will use the column-wise mean of the Spearman's correlation coefficient. Kaggle calculates the coefficient for each column and uses the mean of the calculated values as an indicator of the model's performance. In the case that we are not able to submit our models to Kaggle for evaluation, we will randomly select $\sim$500 test samples from our training data of 6079 samples. We plan to use a 5579/500/476 train/val/test split for our evaluations and hyperparameter tuning assuming we do not have to resort to that scenario.

For modeling, we plan to utilize naive Bayes and logistic regression, which will help for cases where we prefer inference in our predictions. These models will provide us with a baseline to start with, before switching to a deep learning approach, namely LSTMs. If time permits, we will experiment with transformers (e.g., BERT [11]), which have shown promising results on existing datasets such as SQuAD [12]. Due to our relatively small dataset size, training and testing will only require use of personal computing machines. In case we need more computing power, we will resort to Google Colab's Tesla K80 GPU.

By the time of the midway report, we plan to complete data analysis, data preprocessing, and minimal model training (both inference and deep learning models) in that order. After the midway report, we will carry out model analysis, explore implementing BERT, expand our approach to labeling SQuAD, and analyzing our findings. In terms of splitting up the work, we each will be responsible for a single model, but will carry out analysis and preprocessing of datasets as a team.

Thankfully, we do not require a backup plan as the training data, testing data, and evaluation metric are provided, our goal is defined, and our plan is well-thought out with responsibilities properly delegated to each individual of the group. For the final report, each member is responsible for writing about their model, but we will all collaborate on data analysis and findings/conclusion.

References

[1] Haytham M. Fayek and Justin Johnson. Temporal reasoning via audio question answering, 2019.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[3] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[4] Zendesk. Text by Zendesk, 2020.

[5] Cătălina Cangea, Eugene Belilovsky, Pietro Liò, and Aaron Courville. VideoNavQA: Bridging the gap between visual and embodied question answering, 2019.

[6] P.-Y Hsueh, Prem Melville, and Vikas Sindhwani. Data quality from crowdsourcing: A study of annotation selection criteria. *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35, 01 2009.

[7] Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. Modeling social annotation data with content relevance using a topic model. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 835–843. Curran Associates, Inc., 2009.

[8] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 101, 01 2012.

[9] Boris Katz, Gary Borchardt, and Sue Felshin. Natural language annotations for question answering. pages 303–306, 01 2006.

[10] Google. Google QUEST Q&A labeling, 2020.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2018.

[12] Zhangning Hu. Question answering on SQuAD with BERT. 2019.