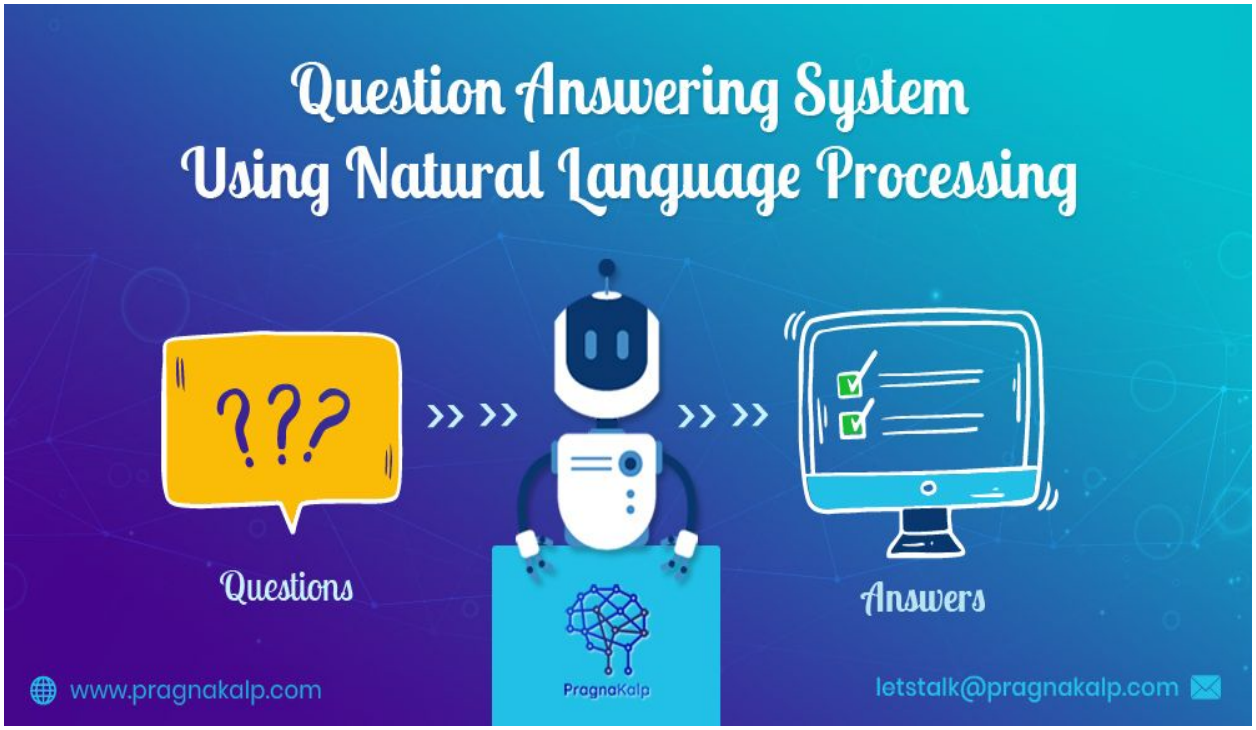


Google QUEST Q&A Labelling

29/09/2020

Manav Nitin Kapadnis
Department of Electrical Engineering
IIT Kharagpur

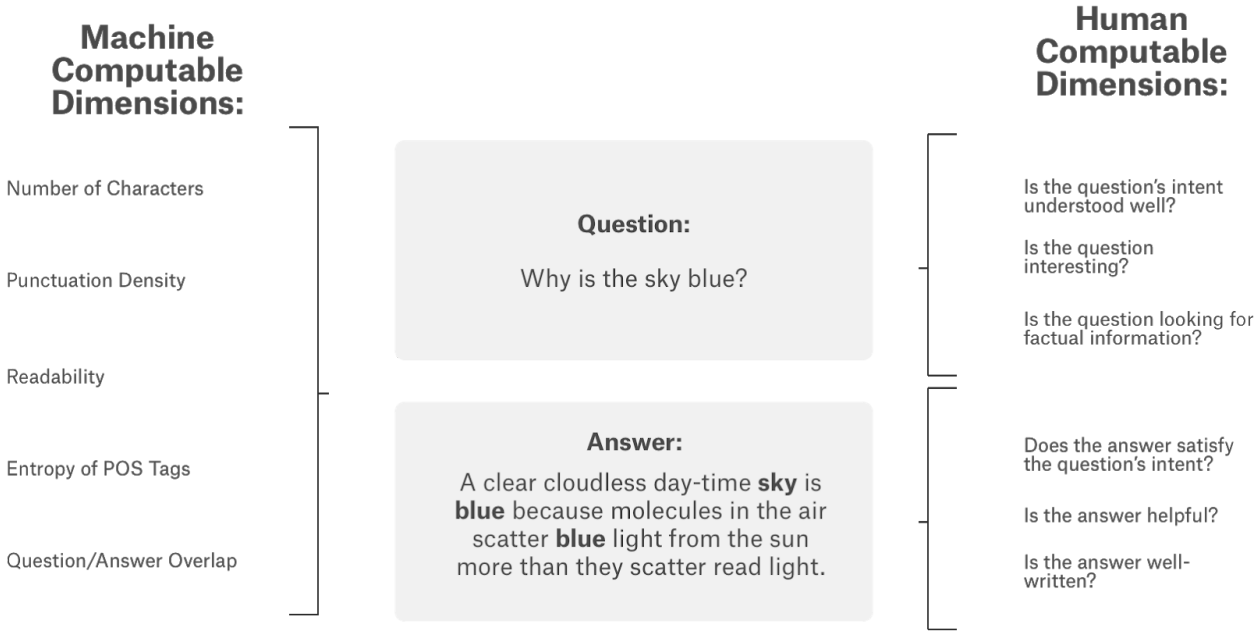
Overview/Domain Background

The main aim behind this project is to understand and learn a major part of Natural Language Processing Applications which Question and Answering using a given phrase of sentence.

Problem Statement

Computers are really good at answering questions with single, verifiable answers. But, humans are often still better at answering questions about opinions, recommendations, or personal experiences.

Humans are better at addressing subjective questions that require a deeper, multidimensional understanding of context - something computers aren't trained to do well...yet.. Questions can take many forms - some have multi-sentence elaborations, others may be simple curiosity or a fully developed problem. They can have multiple intents, or seek advice and opinions. Some may be helpful and others interesting. Some are simple right or wrong.



Unfortunately, it's hard to build better subjective question-answering algorithms because of a lack of data and predictive models. That's why the CrowdSource team at Google

Research, a group dedicated to advancing NLP and other types of ML science via crowdsourcing, has collected data on a number of these quality scoring aspects.

In this dataset, you're challenged to use this new dataset to build predictive algorithms for different subjective aspects of question-answering. The question-answer pairs were gathered from nearly 70 different websites, in a "common-sense" fashion.

Datasets And Inputs

The dataset used for this project is the dataset given in Google Quest QnA labelling competition hosted on kaggle 8 months ago.I'll attach the dataset files when in the same github repository when I submit my Capstone Project.

The data for this competition includes questions and answers from various StackExchange properties. Your task is to predict target values of 30 labels for each question-answer pair.

The list of 30 target labels are the same as the column names in the `sample_submission.csv` file. Target labels with the prefix `question_` relate to the `question_title` and/or `question_body` features in the data. Target labels with the prefix `answer_` relate to the `answer` feature.

Each row contains a single question and a single answer to that question, along with additional features. The training data contains rows with some duplicated questions (but with different answers). The test data does not contain any duplicated questions.

This is not a binary prediction challenge. Target labels are aggregated from multiple raters, and can have continuous values in the range `[0,1]`. Therefore, predictions must also be in that range.

Since this is a synchronous re-run competition, you only have access to the Public test set. For planning purposes, the re-run test set is no larger than 10,000 rows, and less than 8 Mb uncompressed.

Additional information about the labels and collection method will be provided by the competition sponsor in the forum.

File descriptions

- train.csv - the training data (target labels are the last 30 columns)
- test.csv - the test set (you must predict 30 labels for each test set row)
- sample_submission.csv - a sample submission file in the correct format; column names are the 30 target labels

Benchmark model

There is no such benchmark model however the highest team on leaderboard got 0.43100 .Therefore I will try to get my score as near as possible.In order to evaluate my code and my output,I will make a submission to the competition,whose score later I'll will take a screenshot and add it to the github repository of submission.

Evaluation Metric

Submissions are evaluated on the mean column-wise Spearman's correlation coefficient. The Spearman's rank correlation is computed for each target column, and the mean of these values is calculated for the submission score.

Submission File

For each `qa_id` in the test set, you must predict a probability for each target variable. The predictions should be in the range `[0,1]`. The file should contain a header and have the following format:

```
qa_id,question_asker_intent_understanding,...,answer_well_written
6,0.0,...,0.5
8,0.5,...,0.1
18,1.0,...,0.0
etc.
```

Solution Statement/Project Design

I am going to try four different models for QnA on this dataset,which will be BERT,DistilBERT,RoBerta and BART(if possible).This type of project is also new to me.I have taken this as a challenge.

Firstly I will clean the data using the tokenizer of Bert which does most of text preprocessing for me in most of the cases . Then it will be followed by some EDA ,which will

be done in another ipython notebook.The EDA will mostly be done in Plotly or Seaborn library.

For the Embeddings, I will try to use Glove Embeddings of different sizes and check whichever is suitable for giving higher accuracy.I will also try the pre trained model embeddings that are already there.

Thankyou For Reading....