Capstone Project - Machine Learning Engineer Nanodegree

Renee Jia
July.5th, 2020

QUEST Q&A Labeling
—— Improving automated understanding of complex question answer content

Domain background

The distinguishment of computers and human-beings is that the computers are really good at answering the questions with a fixed answer. However, humans are better at those questions which are more comprehensive and require more emotional experiences.

We know that machine learning could take great advantage of the super power using the computers to learn a lot of the fixed historical data and solve problems. However, due to the factor that the computers' lack of knowledge in a deeper, multidimensional understanding of context, computers may need to improve in this part.

This question aroused a lot researchers in the field to dive deep in the area of 'common sense', trying to figure out by simulating the human ability to make presumptions about the type and essence of ordinary situations they encounter every day.These assumptions include judgments about the physical properties, purpose, intentions and behavior of people and objects, as well as possible outcomes of their actions and interactions.[1]

Problem statement

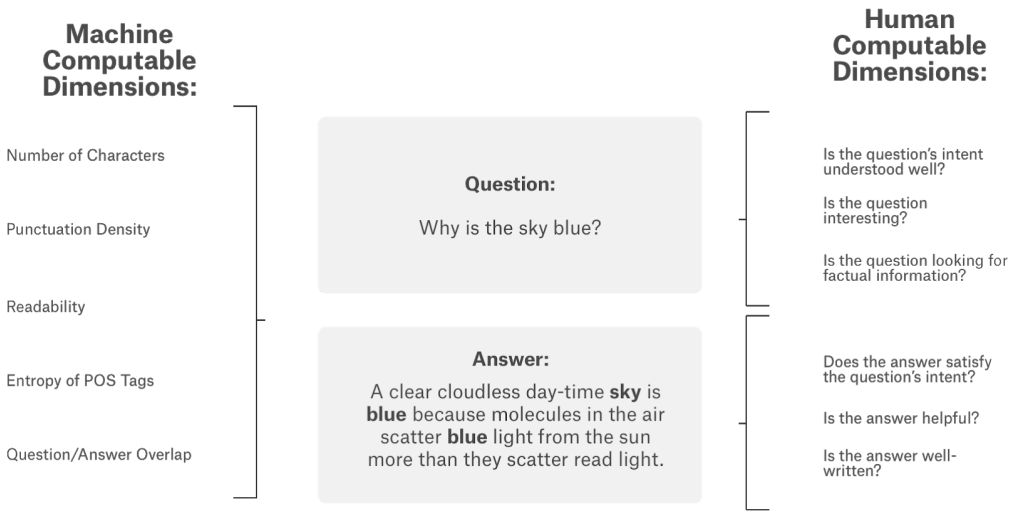
The questions can take many forms - some have multi-sentence elaborations, others may be simple curiosity or a fully developed problem. They can have multiple intents, or seek advice and opinions. Some may be helpful and others interesting. Some are simple right or wrong.

The overall problem could be categorized as a supervised learning prediction problem: to use the dataset to build the prediction for different subjective aspects of question-answering.

The question-answer pairs were gathered from nearly 70 different websites, in a "common-sense" fashion. The raters received minimal guidance and training, and relied largely on their subjective interpretation of the prompts. As such, each prompt was crafted in the most intuitive fashion so that raters could simply use their common-sense to complete the task.

The chart below gives a general overview of the problem, and what the emphasis of this problem is trying to achieve in terms of bridging the gap between the human's comprehension of a subjective question and where could we take the advantage of the computer's computational power:

[1] Common sense reasoning wiki: https://en.wikipedia.org/wiki/Commonsense_reasoning



Dataset and inputs[2]

The CrowdSource team[3] 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.

The data comes from the google-quest-challenge data on Kaggle, which 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.

The picture below shows the data format:

[2] Data input: https://www.kaggle.com/c/google-quest-challenge/data
[3] Google crowdsource team: https://crowdsource.google.com/

```
train_data.head()
```

	qa_id	question_title	question_body	question_user_name	question_user_page
0	0	What am I losing when using extension tubes in...	After playing around with macro photography on...	ysap	https://photo.stackexchange.com/users/1024
1	1	What is the distinction between a city and a s...	I am trying to understand what kinds of places...	russellpierce	https://rpg.stackexchange.com/users/8774
2	2	Maximum protusion length for through-hole comp...	I'm working on a PCB that has through-hole com...	Joe Baker	https://electronics.stackexchange.com/users/1015
3	3	Can an affidavit be used in Beit Din?	An affidavit, from what i understand, is basic...	Scimonster	https://judaism.stackexchange.com/users/5151
4	5	How do you make a binary image in Photoshop?	I am trying to make a binary image. I want mor...	leigero	https://graphicdesign.stackexchange.com/users/...

Solution statement

Since 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. This challenge could be considered as a prediction problem of predicting the probability of the 30 labels for each question-answer pair and calculate the probability for each of the labels.

Benchmark Model

Since this is defined as a prediction model, the benchmark model would be using the MultiOutputRegression model for setting up a decent baseline. By starting from a linear regression model, this could provide a solid baseline for further model explorations.

Evaluation metrics

The model would be 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 overall performance of the model.

Specifically, the Spearman’s correlation coefficient is calculated as:

$$r_s = \rho_{\mathrm{rg}_X, \mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X} \sigma_{\mathrm{rg}_Y}},$$

Outline of project design

1. Retrieving the data
2. Data Overview
    a. Overview of the tables to get a sense of the datasets
    b. Conduct statistical analysis of the dataset and draw plots to get a sense of the files
3. Data Exploration
    a. Draw distribution plots of host, categories, target variables
    b. Draw distributions of the Question Title/ Body and distribution of the answers
4. Data preparation
    a. Data cleaning: Check with missing data and using techniques to manipulate data (eg: deal with null values, etc.)
    b. Feature Engineering
        i. Text based features
        ii. TF - IDF features exploration
5. Setup baseline model as benchmark
6. Data splitting
    a. Split the training dataset into testing & validation datasets
7. Model building
    a. Build different models for comparison
8. Define testing metrics
9. Model validation and testing
    a. Using the validation dataset to evaluate the performance of different models and compare
    b. Using the test dataset to generate the submission file and upload to kaggle competition
10. Write up
    a. General summary of the model comparisons
    b. Business insights generation