Identify Queries With Similar Intent

A key element for question and answer website efficiency such as Quora relies on properly categorizing questions such that questions with identical intent are not duplicated. This allows users to both find and answer questions easily without having to search through all the different duplicate question pages. For example, the questions “How do you start a bakery?” and “How can one start a bakery business?” should not have separate pages because the intent of both questions are identical. This problem makes it imperative for us to recognize entailment relations between the pairs of sentences using neural network architectures. This strikes me as an interesting problem because the application is not just limited to finding sentences with the same intent but this ML solution can be potentially extended for search suggestions, grammar correction, chatbots etc.

Dataset:

Dataset released by Quora that consists of over 400,000 lines of potential duplicate pairs will be used. The dataset is supplemented with negative examples to balance the number of positive and negative examples. Figure below shows the format of the raw dataset. That is for each pair of questions a sample ID, an individual question ID, the questions and their corresponding labels are provided (duplicate = 1, not duplicate = 0). https://www.kaggle.com/c/quora-question-pairs/data

id	qid1	qid2	question1	question2	label
20	41	42	Why do rockets look white?	Why are rockets and boosters painted white?	1
21	43	44	What's causing someone to be jealous?	What can I do to avoid being jealous of someone?	0
22	45	46	What are the questions should not ask on Quora?	Which question should I ask on Quora?	0

Approach:

This problem can be solved using “traditional” machine learning as well as Deep learning approach. We can propose various RNN and evaluate various models based on complexity and accuracy. The problem can be approached as a supervised learning problem. The model will simply take 2 sentences as an input and respond with yes/no indicating if the intent of question same or not with addition to the percentage indicating confidence level.

Deliverables:

The final deliverable will be an api driven web application hosted on a cloud that can be used as a demo to showcase the learning.

Resources:

16GB or more RAM, 8 core or more CPU or a GPU for enhanced performance if needed.