Heinz 95-845: Project Proposal

Ao Luo AOL@ANDREW.CMU.EDU
Porter Hall A107
Carnegie Mellon University
Pittsburgh, PA, United States

Dingze Wang DINGZEW@ANDREW.CMU.EDU
Heinz College A206
Carnegie Mellon University
Pittsburgh, PA, United States

1. Project Details

1.1 Objectives

- Construction and description of an analytic framework that motivates the use of machine learning for your task

  Quora is a website that contains millions of different questions. For users, sometimes it's hard to use the search engine to find the existing questions that almost identical to what they want to ask. Let's say we have two questions:
  1. "What would a Trump presidency mean for current international masters students on an F1 visa?"
  2. "How will a Trump presidency affect the students presently in US or planning to study in US?"
  The structure of both sentences are different but actually they are identical question. How to use machine learning technology to tag same question? This is one of the biggest problems that Quora is trying to solve, and also for some other big company to solve such as Alibaba.

- Presentation of machine learning techniques appropriate for the task

  Currently Quora use Random Forest to tag the identical questions. However, this model has it's limitation on prediction. We are going to using different machine models such as SVM and Neural Networks with LSTM to solve the problem and compare their performance. We will combine several machine learning models together if the pipeline can achieve better result and analyze the performance of each of them. Also, we'll test the behaviour of the model in different dataset, even in different language.

- Description of the data

  The goal of this project is to predict which of the provided pairs of questions contain two questions with the same meaning. The ground truth is the set of labels that have been supplied by human experts. The ground truth labels are inherently subjective, as the true meaning of sentences can never be known with certainty. Human labeling is also a 'noisy' process, and reasonable people will disagree. So, thanks for so many experts works on the dataset so we can study on these labels.

Data fields

id : the id of a training set question pair
qid1, qid2 - unique ids of each question (only available in train.csv)
question1, question2 - the full text of each question
isDuplicate - the target variable, set to 1 if question1 and question2 have essentially the same meaning, and 0 otherwise.

- Description of the likely analysis outcomes and their impact.

  For the outcome, we are expected to develop a model that can distinguish if two question are actually asking the same thing. By implementing this machine learning model, the employees on websites such as Quora can greatly save their time on manage the duplicate information which can greatly save the time and disk space. Users can also find the related information very quickly.

2. Proposal Details (10 points)

2.1 What is your proposed analysis? What are the likely outcomes?

Compare the performance and complexity of different models on same dataset, and compare the performance of these model on datasets within different field. The analysis could help us find the model with best generality and the reason behind.

2.2 Why is your proposed analysis important?

It helps us find the model with good generality for tackling the Duplicate Question Detection task, and the reason behind.

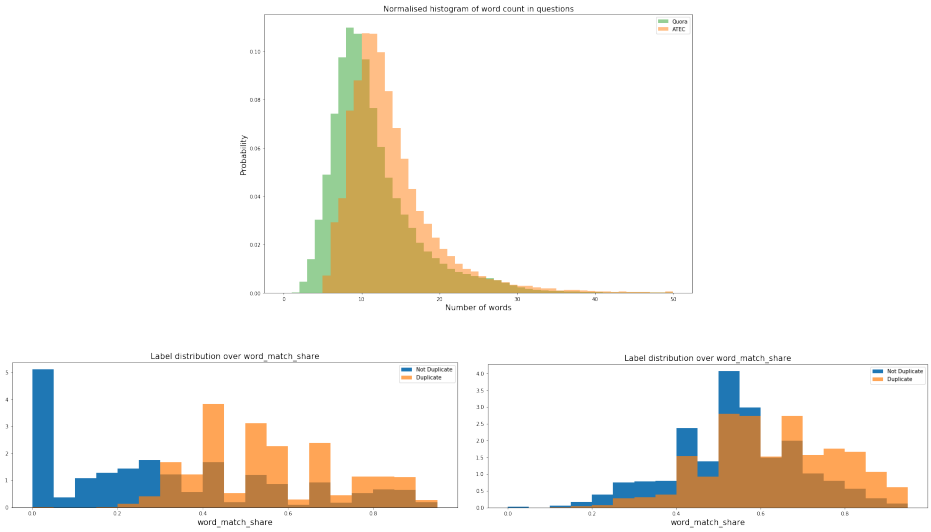
2.3 How will your analysis contribute to existing work?

On Duplicate Question Detection task, there are some papersSaedi et al. (2017) focused on comparing the state-of-art model on different size of dataset, or on certain dataset. Few researches are done comparing different models on datasets within different field, and analyze the generality. Our project could fill this gap.

2.4 Describe the data. If applicable, please also define Y outcome(s), U treatment, V covariates, and W population.

Figure 2.4 shows the length distribution of the question pairs for two data sets. Figure 2.4 and 2.4 shows the label distribution with number different shared word for training set. It could be noticed that the two data sets are quit different. Also, since simple number of word shared length could not differentiate two sentences, more deep features should be explored.

- **Quora question pairs**. The total number of question pairs is 2750086. Each pair contains two short questions. We split the data equally to train and test set. For train set, question pairs with positive label account for 21.68%.



- **Financial services question pairs**. The total number of question pairs is 39346. Each pair contains two short questions. We leave 80% of the data as test set. For train set, question pairs with positive label account for 36.92%.

2.5 What evaluation measures are appropriate for the analysis? Which measures will you use?

- **Log loss**. Log-loss takes into account the uncertainty of the prediction based on how much it varies from the actual label. In this binary classification task,

$$Log - loss = -(y \log p + (1 - y) \log(1 - p))$$

- **ROC-AUC**. The probabilistic interpretation of ROC-AUC score is that if you randomly choose a positive case and a negative case, the probability that the positive case outranks the negative case according to the classifier is given by the AUC. Mathematically, it is calculated by area under curve of sensitivity (TPR) vs. FPR(1-specificity).

- **F1-score**. In the real situations, labels imbalance might exist. F1-score makes more sense in these cases.

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

2.6 What study design, pre-processing, and machine learning methods do you intend to use? Justify that the analysis is of appropriate size for a course project.

- **Study design**. Comparison: 1) performance of same model on different datasets 2) performance and complexity of different model on same datasets.

- **Pre-processing**. To improve the question quality before training, we need do text normalization, which includes removing non-ascii, digits, punctuation and stopwords. In addition, low-case transformation, stemming and lemmatization are applied to English text, while common word correction is applied to Chinese text.
- **machine learning methods**. Neuron networks models are the main frameworks. Also, boosting tree models(LGB and XGB) are used to handle the tabular features and increase the model diversity.

2.7 What are possible limitations of the study?

Since the limitation of time and resource, only a portion of the models are compared, a simple model ensemble techniques(10-fold-cross-validation) is applied. Also, data sometimes can be dirty. Sometimes features can be missing or very hard to extracted. Can we just simply use interpolation or chose some different approach. This is another challenge

2.8 Who will use your analytic pipeline? In one or two sentences, describe an example of its use.

If a company want train a question pair similarity classification model based on their own datasets, our analytic pipeline could help them compare the generality of these models, and choose the model that best satisfy their needs.

References

C. Saedi, J. Rodrigues, J. Silva, A. Branco, and V. Maraev. Learning profiles in duplicate question detection. In *2017 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 544–550, Aug 2017. doi: 10.1109/IRI.2017.39.