Machine Learning Engineer Nanodegree

Capstone Proposal

Elvyna Tunggawan
March 20, 2018

Domain Background

Emergence of web 2.0 has increased variety, veracity, velocity and volume of textual data. Demand of text analytics researches also increases, such as sentiment analysis, topic modelling, text summarization, and text duplication detector. One of the challenges on text analytics, especially on text duplication detector, is identifying semantically equivalent sentences. Different phrases could have similar meaning, while a word also could have different meanings. Moreover, there are thousands of language in the world with various grammars and vocabularies, which add the complexity.

Number of past researches were conducted to identify text similarity. Achananuparp et al.[1] evaluated performance of different sentence similarity measures and found that linguistic measures perform better than word overlap and TF-IDF measures. However, it couldn't produce satisfactory result on text which requires more specific textual entailment. Bogdanova et al.[2] represented documents from Ask Ubuntu Community Questions and Answers site using word embeddings and fed them into convolutional neural network (CNN). CNN with pre-trained in-domain word embeddings resulted in higher accuracy than using English Wikipedia's word2vec. Based on experiments on different domains, neural network could achieve high accuracy regardless of the domain and size of training data.

Problem Statement

As a question-and-answer site, Quora enables people to submit questions, give high quality answers, and learn from each other. Having more than 100 million users, submitted questions could have similar words, and multiple questions might have similar intent. Quora emphasizes on providing canonical questions in order to ease user experience on finding the best answers of similar questions and prevent users from answering similar questions. This problem also arises on other question-and-answer sites, such as Stack Overflow. To provide canonical questions, identification of duplicate questions could be done using a classification model. Each question pair could be represented as numerical features and be used as input on the model.

---

[1] Achananuparp et al. (2008) The Evaluation of Sentence Similarity Measures. In DaWaK '08 Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery, 305-316.

[2] Bogdanova et al. (2015) Detecting Semantically Equivalent Questions in Online User Forums. In Proceedings of the 19th Conference on Computational Language Learning, 123-131.

Datasets and Inputs

The dataset used on this work is retrieved from Quora Question Pairs Kaggle Competition[3], which contains full text of the question pair. The dataset contains 404,290 question pairs, which will be split into training, validation, and test set. Each question pair is labeled manually by human experts, which might not be 100% accurate. However, the labels are considered to represent reasonable consensus and will be used as the ground truth. The dataset structure is described on the table below.

Field Name	Description	Data Type
id	id of a question pair on training set	int64
qid1	unique ids of each question	int64
qid2	unique ids of each question	int64
question1	full text of each question	string
question2	full text of each question	string
is_duplicate	target variable (boolean); 1 if *question1* and *question2* have essentially same meaning, otherwise 0	int64

Features on *question1* and *question2* will be extracted and used as input variables to the model; while *is_duplicate* resembles the target variable.

Solution Statement

As a solution this problem, I plan to implement Siamese neural network model, since it is popular among tasks that involve finding similarities or relationships between objects[4]. Prior to model training, data preprocessing and feature engineering will be done to extract important features of the dataset. I consider using word2vec to extract semantic representation of each word in the question pairs.

Benchmark Model

Quora developed random forest model as initial solution to this classification task, then built another Long Short Term Memory (LSTM)-based model[5]. Their best model achieved 87% accuracy and 88% F1 score, which will be an aspirational target. Another research done by

[3] Quora Question Pairs Competition. https://www.kaggle.com/c/quora-question-pairs/data. Accessed on February 16, 2018.
[4] What are Siamese neural networks, what applications are they good for and why. https://www.quora.com/What-are-Siamese-neural-networks-what-applications-are-they-good-for-and-why. Accessed on March 11, 2018.
[5] Semantic Question Matching with Deep Learning. https://engineering.quora.com/Semantic-Question-Matching-with-Deep-Learning. Accessed on March 10, 2018.

Shankar and Shenoy achieved 85% accuracy from Support Vector Machine (SVM) model[6]. Homma et al. also achieved 85% accuracy with 84% F1 score from Siamese Gated Recurrent Unit (GRU) with 2-layer similarity network trained on an augmented dataset[7]. Thus, 85% accuracy should be a more attainable target.

Evaluation Metrics

Model performance will be measured based on its accuracy, which is defined as number of correctly predicted class (duplicate or not) divided by number of question pairs on the test set. In addition, F1 score will be calculated as measure of model performance on identifying the duplicate pairs.

Project Design

The first stage of this project is data acquisition, which is done by downloading the dataset via Kaggle. Initial data exploration will be done to identify characteristics of the dataset, as well as taking proper treatment on missing values (if any). During data exploration, observation of characteristics on each record will be done, such as number of unique words, sentence length, and number of unique words, sentence length, and number of common words between question pairs.

After having a glance of the dataset characteristics, features of each question pair will be extracted. Part-of-speech (POS) tagging will be performed on each question, as well as named entity recognitions. Similarity measures from each question pair,, TF-IDF measures and some linguistic measures proposed by Achananuparp et al., as well as word embedding are also considered to be used to enrich the input features. Word embedding and other extracted features will be combined and be used as input to the classification model.

As initial approach, similar model architecture with Cohen[8] will be implemented, which contains embedding matrix as input to LSTM network. Each question on the question pair will be represented as an embedding matrix, and both questions will be processed on separated LSTM network and result in two outputs. Similarity of the outputs will be calculated using exponent of negative Manhattan distance, which will result in 0 or 1:

$$exp(-\left\|h^{left} - h^{right}\right\|_1)$$

The trained model will be cross validated and tested on the test set. If its accuracy is lower than the benchmark model's, another model architecture will be considered in order to gain better accuracy.

---

[6] Shankar and Shenoy. (2017) Identifying Quora question pairs having the same intent.
[7] Homma et al. (2016) Detecting Duplicate Questions with Deep Learning. 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.
[8] How to predict Quora Question Pairs using Siamese Manhattan LSTM. https://medium.com/mlreview/implementing-malstm-on-kaggles-quora-question-pairs-competition-8b31b0b16a07. Accessed on March 10, 2018.