Project Proposal

Paryul Jain - 20171083
Eesha Dutta - 20171104

Question and Answering

SemEval 2017 -Task 3

Overview

The task requires us to find appropriate answers and rank them according to relevance to a new question which has not already been seen in a given collection of questions and answers.
This is the main task which we call CQA(Community Question Answering).
The task description also suggests alternative paths to achieve the same. This alternative path requires two subtasks to be done.
1. Question Similarity (QS) - Given a question (new question), and a collection of questions, we need to rank those questions according to their similarity from the original question.
2. Relevance Classification (RC) - Given a question from a question answer thread, rank the answer posts according to their relevance with respect to the question.
Task 1 (QS) hopes that the answers to the similar questions should be answering the new questions as well. Once we find a list of similar questions to our original question, we can then rank the best answers from these questions using the tools we develop to solve RC.

Goals

The SemEval task of 2017 is an extension of SemEval 2016 where 4 tasks were covered.
1. Question - Comment Similarity
2. Question - Question Similarity
3. Question - External Comment Similarity
4. ReRank correct answers for a new question
In 2017 an additional task was added but no advancements were made.
We focus on the first 2 tasks and if time permits we will move on to the next 2.

Question - Comment Similarity :- Given a question Q and the first ten comments in its question thread (c1, . . . , c10), the goal is to rank these ten comments according to their relevance with respect to that question.
Question - Question Similarity :- Given a new question Q (aka original question) and the set of the first ten related questions from the forum (Q1, . . . , Q10) retrieved by a search engine, the goal is to rank the related questions according to their similarity with respect to the original question.

Specifications

We will be using the dataset provided from the SemEval task of 2016 and also the CQADupStack dataset which contains over 7 million threads from 12 StackExchange subforums(each focusing on a particular domain). We will also run our model on the Kaggle Quora Dataset as the dataset and results are much easier to interpret and evaluate. Also this widens the space where we can test the robustness of our model.

Similar Word Counts - A very simple approach to detecting similarity between a pair of questions would be to look at unique words in the first question that are also present in the second question as a ratio of the total words in both questions. This number could then be used in a simple model such as logistic regression to predict duplicate versus different questions. This approach has drawbacks since there can be two similar questions with very less words in common but similar meaning. Could be because of sentence structures or use of synonyms etc. From this approach we learn that it is important for us to look beyond comparing words, and to compare the semantic similarity of sentences. Which then leads us to meaning representation of sentences and use of Neural Networks, LSTM's and Attention.

We would like to explore the approaches mentioned in the paper authored by Yallis Chali which explores use of LSTM and Attention for representation of sentences in a fixed dimension vector. And then using similarity techniques like cosine to train the model. We also referred to Ali Fadel's publication which uses pre-trained ELMo embeddings and feeds them to a LSTM with self attention. We will compare the results obtained from the above models and also try to suggest improvements over these.

These will be SVM based classification models which we saw in the SemEval-17 Task-3 top performer’s approach which do specific feature extraction to improve performance.