Semantic Textual Similarity

16.09.2020

Team 17 (PreNLPians)

R. Guru Ravi Shanker (2018114011)

Jashn Arora (2018114006)

Mentor : Prashant Kodali

Professor : Manish Shrivastava

Problem Statement and Aim

To create a model that given a pair of sentences predicts the semantic relationship between them in terms of paraphrases, entailment, contradiction.

Datasets

1. Quora Question Pairs Dataset
    a. Given two questions: Label is 0 if the questions are not paraphrases and label is 1 if the questions are paraphrases.
2. SICK Natural Language Inference dataset
    a. Given two sentences:Label is 1 if sentences entail one another, Label is -1 if sentence contradicts one another and label is 0 if the sentences are neutral to one another.
3. WikiQA For Question Answering Pair

   a. Given qn and answer 1 if answer follows the question else zero.

NOTE: 1 is our primary datasets and the model will be trained on 2 and 3 if time permits.

BaseLine Model

Bilateral Multi-Perspective Matching for Natural Language Sentences

1. Given two sentences the model has to predict y according to the training dataset.This is done by matching the encoded sentences at all time frames with multiple perspectives.
2. The Network has 5 Layers:
   a. Word Embedding Layer:
      Trained Glove word embeddings to represent the sentences.
   b. Contextual Embedding Layer:
      Used Bi-LSTM model to incorporate context in the word embeddings.
   c. Matching Layer:
      The goal of this layer is to compare each contextual embedding (time-step) of one sentence against all contextual embeddings (time-steps) of the other sentence.
   d. Aggregation Layer:
      It aggregates the two sequences of matching vectors to a fixed layer using another Bi-LSTM model.
   e. Prediction Layer:
      This layer feeds forward the fixed length matching vector and applies a softmax function in the output layer to predict the value y.

Baseline +

1. Use modern state of art models like BERT/XLNET to contextually represent the sentences.
2. Adding an attention matching layer to the model.

TimeLine:

1. **24th September**
    a. Deciding which frameworks to be used for the task and understanding the networks deeply
2. **1st October**
    a. Baseline Implementation will be completed
3. **7th October**
    a. Rectify any errors occurred during the first implementation of the code. Achieving the accuracy stated in the paper.
4. **14th October**
    a. Literature Review of the current state of art developments for Semantic Textual Similarity Task and improving our model.
5. **21 October**
    a. Improving the baseline model by using BERT/XLNET representation of sentences and adding attention layers if required.
6. **28th October**
    a. Improving the accuracy of the baseline model with state of the art implementation.
7. **7th November**
    a. Do qualitative analysis on results
8. **16th November**
    a. Final model delivered.
    b. Read more papers on the STS for completing term papers.