Interim Report

Team 38

- Sidharth Giri, 2019101007
- Dhruv Kapur, 2019101038

Problem Description

Semantic Textual Similarity (STS) measures the **degree of equivalence in the underlying semantics of paired snippets of text**. Given two sentences, the model should return a *continuous-valued similarity score on a scale from 0 to 5*, with 0 indicating that the semantics of the sentences are completely independent and *5 signifying semantic equivalence*. Performance is assessed by computing the **Pearson correlation** between machine assigned semantic similarity scores and human judgements.

Approaches Tried

Currently we have worked on creating baselines for the monolingual and cross lingual semantic textual similarity.

Monolingual

- **TF-IDF vectorization**. Used an unsupervised method to create vectors for sentences. To obtain a similarity score:
  - Used cosine similarity between the vectors to find a score between [0, 1], where 0 would indicate towards similar vectors.
  - Scaled the score to the range of [0, 5] to predict the final score.

Cross-Lingual (En-Es)

- Since TF-IDF vectors from two different languages cannot be **compared straight-forwardly by just taking cosine similarity** (*since different languages would have different word token bases, which would not be semantically aligned*). So we used a **supervised predictive model** to find similarity between cross lingual sentences.

- We concatenate the two TF-IDF vectors, to create a new vector embedding which encodes information from both the sentences. Since the vector is quite sparse, we even *tried to use dimensionality reduction methods (Principal Component Analysis* and *Scalar Vector Decomposition)*, to compress the excess dimensions.
- The **vector is then passes through a MLP (Multi Layer Perceptron)** to predict the semantic similarity score. Through this method we don't have to worry about alignment of the tokens of TF-IDF for different languages.

Challenges

For the TF-IDF Approach that we have as our **baseline**, there several drawbacks to it:

- **Lack of reliable vocabulary**, i.e. highly dependent on training vocab set and common words in test set
- **Context,** (*Essential for semantics*) is not captured in TF-IDF
- When it comes to *cross-lingual* sentence semantic similarity, the **lack of alignment between the TF-IDF vectors** from different languages, makes it almost impossible to use any vector similarity technique to find similarity. Moreover, we have to be quite **dependent on the MLP** to learn patterns from concatenation of two semantically unrelated vectors.