Project Outline

Team 38

- Sidharth Giri, 2019101007
- Dhruv Kapur, 2019101038

Understanding Of Problem

Semantics is how one's lexicon (vocabulary), grammatical structure, tone, and other elements of a sentence combine together to communicate its meaning. Given 2 sentences, we need to quantify how similar the given sentences are in terms of the meaning they are conveying. To quantify this, we have a scale from *0 to 5*, where 0 is lowest level of semantic similarity and 5 implying sentences have the essentially the same meaning.

Scope

References

- Dataset
  - STS 2017 (SemEval 2017 Task 1)
  - Could also use data from previous STS tasks (*additional data will help in training*)
- Papers and Other Readings
  - Task Final Review Paper (Summarizing methods, mistakes and performance to understand what are the approaches used by others for this task)
  - A compilation of papers in this domain, which can be referred later for improving prediction scores.
  - Blog Post by Google AI on Advances in STS
  - Cross Lingual Sentence Embedding
  - Analyzing Cross-Lingual Text Similarity

Interim Deliverables

- Create the baseline, baseline++ approaches for both monolingual and cross-lingual data.
- Create methods to output continuous scores given embeddings.
- Create the pipeline to test and benchmark the approaches (*pipeline includes data cleaning, tokenization, calculating prediction quality using Pearson Correlation, etc.*).

Final Deliverables

- Create a further effective method (based on *deep learning*) to achieve better performance.
- Create a transformers based approach to attempt the problem.
- Working on Cross-Lingual data and devising methods to find STS between English-Spanish sentences.

Implementation Plans

- Baseline using TF-IDF embeddings.
- Baseline ++ using Word2Vec embeddings.

Then further approaches in monolingual *(en-en)* could involve:

- BiLSTM (*or other deep learning approaches, such as using BiLSTM with CNNs*)
- Attention based approaches

For cross-lingual data *(en-es)* approaches could involve:

- Cross Lingual Word Embeddings (using resources like XLM-RoBERTa and mBERT which are a cross-lingual sentence encoder).
- Translation based approaches, using multi-lingual transformers like mT5, where we could translate the text to convert this into a monolingual task.