**Project Name: Semantic Textual Similarity**

|  |  |
| --- | --- |
| Project Name | Semantic Textual Similarity |
| Baseline Model | Siamese Recurrent Architectures for Learning Sentence Similarity +  Augment word representations with character level features |
| Team Members | Sachin Raja, Arvin Goyal |
| Mentor | Prashant Kodali |
| Version | Draft |

Table of Contents

[**Problem Statement Description** 3](#_Toc50907818)

[**Describe Datasets** 3](#_Toc50907819)

[MSRPC 3](#_Toc50907820)

[Semeval STS 3](#_Toc50907821)

[SICK 3](#_Toc50907822)

[**Summaries Base Line Model** 4](#_Toc50907823)

[**Addition to Base Line Model** 4](#_Toc50907824)

[**Instream Dead Lines** 4](#_Toc50907825)

**Problem Statement Description**:

Paraphrasing methods identify, generate, or extract phrases, sentences that convey almost the same information. Different worded sentences may bear the similar meaning and can be identified by paraphrase identification. Paraphrase detection has importance as it contributes to various NLP tasks like Text summarization, Document Clustering, Question Answering, natural language inference, information Retrieval, Plagiarism Detection, Text Simplification.

**Describe Datasets**:

We will be using MSRPC, Semeval STS and SICK and in our projects.

MSRPC:

Microsoft Research Paraphrase Corpus (MSRPC) is text file containing 5800 pairs of sentences which have been extracted from news sources on the web, along with human annotations indicating whether each pair captures a paraphrase/semantic equivalence relationship in binary form. Last published: March 3, 2005.

The dataset consists of 5,801 sentence pairs. The average sentence length is 21, the shortest sentence has 7 words while 36 is the longest sentence length. 3,900 are labeled as being in the paraphrase relationship. 4,076 training pairs are present in training file while 1,725 pairs present in test file. All sentences were labeled by two annotators who agreed in 83% of the cases. A third annotator resolved conflicts.

Semeval STS:

Nearly 14 thousand sentence pairs are linked above with gold human annotated STS labels. It contains test sets with gold standard labels for the 2016 Semantic Textual Similarity (STS) shared task. Each evaluation set input file has the following tab-separated format:

* One STS pair per line
* Each line contains the following fields: STS Sent1, STS Sent2, Sent 1 Source Notes, Sent 2 Source Notes.

Each evaluation set gold standard label file has the following format:

* One numeric STS label between 0 and 5 per line
* Blank lines indicate that the pair was not included in the official scoring

Above are primarily based on the references in

<http://alt.qcri.org/semeval2016/task1/index.php?id=data-and-tools>

SICK:

The Sentences Involving Compositional Knowledge (SICK) data set consists of 10,000 English sentence pairs, built starting from two existing paraphrase sets: the 8K ImageFlickr data set (http://nlp.cs.illinois.edu/HockenmaierGroup/data.html) and the SEMEVAL-2012 Semantic Textual Similarity Video Descriptions data set (http://www.cs.york.ac.uk/semeval-2012/task6/index.php?id=data). Each sentence pair is annotated for relatedness in meaning and for the entailment relation between the two elements.

File Structure: tab-separated text file

Fields:

- pair\_ID: sentence pair ID

- sentence\_A: sentence A

- sentence\_B: sentence B

- entailment\_label: textual entailment gold label (NEUTRAL, ENTAILMENT, or CONTRADICTION)

- relatedness\_score: semantic relatedness gold score (on a 1-5 continuous scale)

- entailment\_AB: entailment for the A-B order (A\_neutral\_B, A\_entails\_B, or A\_contradicts\_B)

- entailment\_BA: entailment for the B-A order (B\_neutral\_A, B\_entails\_A, or B\_contradicts\_A)

- sentence\_A\_original: original sentence from which sentence A is derived

- sentence\_B\_original: original sentence from which sentence B is derived

- sentence\_A\_dataset: dataset from which the original sentence A was extracted (FLICKR vs. SEMEVAL)

- sentence\_B\_dataset: dataset from which the original sentence B was extracted (FLICKR vs. SEMEVAL)

- SemEval\_set: set including the sentence pair in SemEval 2014 Task 1 (TRIAL, TRAIN, or TEST)

Above are primarily based on the reference from

<http://clic.cimec.unitn.it/composes/sick/>

<http://clic.cimec.unitn.it/composes/materials/SICK.zip>

**Summaries Base Line Model**:

**Addition to Base Line Model**:

**Instream Dead Lines**: