Semantic Similarity Detection

BTP Project By:
Hrishabh Pandey S20180010064
Ayush Gairola S20180010020
Rakesh Muchimari S20180010109

BTP Code : **B21AP01**
Mentor: **Dr. Amit Praseed**

Contents

1. About Project
   a. What is semantic Similarity
   b. What is our project
   c. Overview of Methods and models utilized
2. Client and Server Side Demonstration
3. Sentence Embeddings using Siamese BERT-Network
4. Data Sets Utilized
5. References

About Project

What is Semantic Similarity

Semantic Textual Similarity (STS) is defined as the measure of “semantic equivalence” between two blocks of text, phrases, sentences, or documents. Semantic similarity methods usually give a ranking or percentage of similarity between texts.

The main objective Semantic Similarity is to measure the distance between the semantic meanings of a pair of words, phrases, sentences, or documents.

What is our project (Deliverables)

For our BTP project, we took inspiration from the below listed implementations which are currently in use.

1. **Customer Support** : Companies can create a corpus of pre seen frequent queries and with our engine they can look for queries which resemble the most similarity, and send an automated response saving lots of unnecessary work-force.

2. **Medical Search Space** : when a new case comes to a practitioner, he/she can look for similar cases in the past and get the most closely resembling case and make better decisions.

We wish to present this technology in the hands of the general consumer with this platform where individuals and organizations will be able to collect and classify text data, which will accelerate their processes.

Client And Server Demonstration.

Sentence Embeddings using Siamese BERT-Networks

BERT

BERT (Devlin et al., 2018) is a pre-trained transformer network (Vaswani et al., 2017), which set for various NLP tasks new state-of-the-art results.

The input for BERT for sentence-pair regression consists of the two sentences, separated by a special [SEP] token. **Multi-head attention** over 12 (base-model) is applied and the output is passed to a simple regression function a, **Feed Forward NN**, to derive the final label.

A large disadvantage of the BERT network structure is that no independent sentence embeddings are computed, which makes it difficult to derive sentence embeddings from BERT.

We use the pre-trained BERT network and only fine-tune it to yield useful sentence embeddings.

SentBERT

SBERT adds a **pooling operation** to the output of BERT to derive a fixed sized sentence embedding.

For our use case, we are using **MEAN** pooling Strategy.

In order to fine-tune BERT we create **siamese and triplet networks** (Schroff et al.,2015) to update the weights such that the produced sentence embeddings are semantically meaningful and can be compared with cosine-similarity.

For the First Model we are using **Regression Objective Function (ROF)** .

In ROF, The cosine similarity between the two sentence embeddings u and v is computed.

We use **Mean-Squared-Error** loss as the objective function.

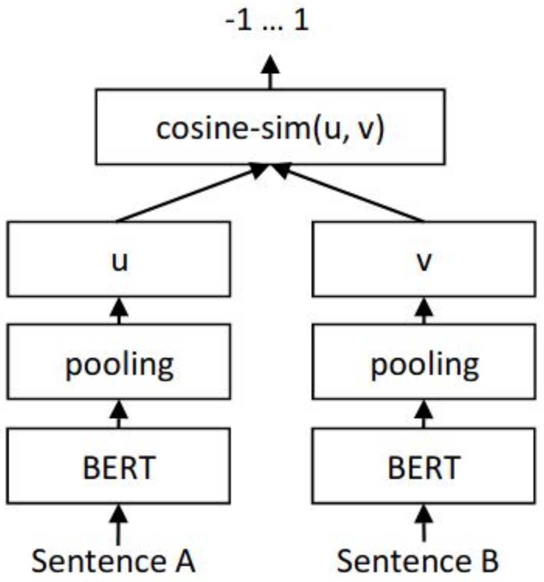


Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

For our current Model we are using Siames Network with Triplet loss.

Siamese Network

A Siamese Neural Network is a class of neural network architectures that contain two or more identical subnetworks. Parameter updating is mirrored across both sub-networks. It is used to find the similarity of the inputs by comparing its feature vectors, so these networks are used in many applications

Triplet Ranking Loss

Given an **anchor sentence a**, a **positive sentence p**, and a **negative sentence n,** triplet loss tunes the network such that the distance between a and p is smaller than the distance between a and n. Margin $\boldsymbol{\epsilon}$ ensures that $\mathbf{s_p}$ is at least $\boldsymbol{\epsilon}$ closer to $\mathbf{s_a}$ than $\mathbf{s_n}$.

$$max(||s_a - s_p|| - ||s_a - s_n|| + \epsilon, 0)$$

Datasets Used

Sick DataSet

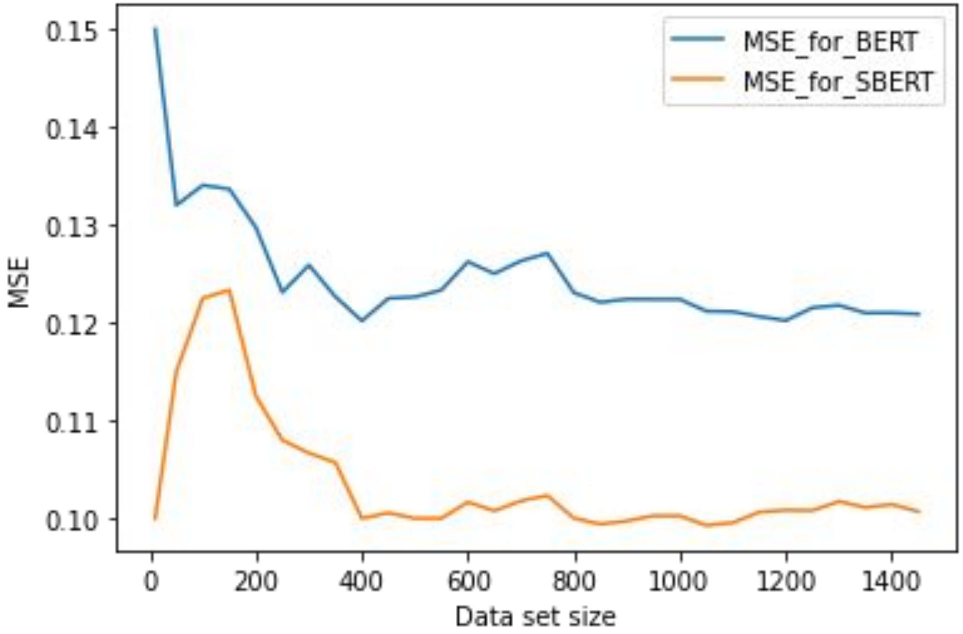
Marelli et al.compiled the SICK dataset for sentence level semantic similarity/relatedness in 2014 composed of 10,000 sentence pairs obtained from the Image Flickr 8 and MSR-Video descriptions dataset. The sentence pairs were derived from image descriptions by various annotators. 750 random sentence pairs from the two datasets were selected,followed by three steps to obtain the final SICK dataset: sentence normalisation, sentence expansion and sentence pairing.

STS DataSet

In order to encourage research in the field of semantic similarity, semantic textual similarity tasks called SemEval have been conducted from 2012. The organizers of the SemEval tasks collected sentences from a wide variety of sources and compiled them to form a benchmark data set against which the performance of the models submitted by the participants in the task was measured

Experimental Results

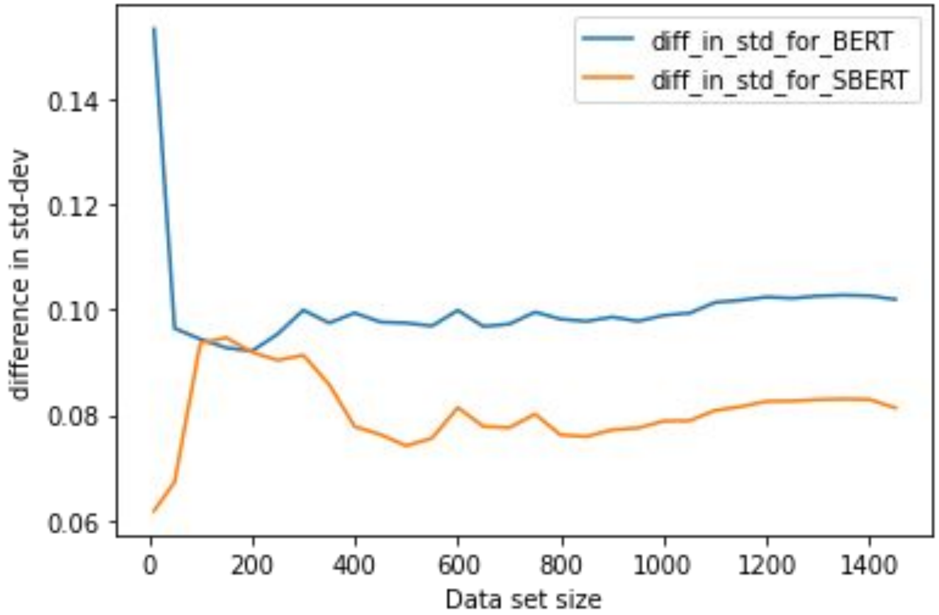
Mean Squared Error



For test dataset of size 1500 **MSE** are observed as below

BERT	SBERT
0.12088	0.10068

Difference in Standard Deviation



For test dataset of size 1500, **STD-DEV** difference 's are observed as below

BERT	SBERT
0.10196	0.08135

References

1.. Reimers, N. and Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

2. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

4. Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering.arXiv preprint arXiv:1503.03832, abs/1503.03832.

5. https://github.com/huggingface/transformers (hugging face, open-source, Transformers Implementation and Base BERT Models)

Thank You.

BTP Project By:

Hrishabh Pandey S20180010064
Ayush Gairola S20180010020
Rakesh Muchimari S20180010109

BTP Code : **B21AP01**
Mentor : Dr. Amit Praseed