Team KLM

1. Beeraka Krupa Kiranmai - 2021201022
2. Padigala Lakshman Sai - 2021201069
3. NVSS Maneesh Gupta - 2021201041

Semantic Textual Similarity

Intro to NLP

Overview

Semantic Textual Similarity measures the meaning similarity of the sentences. The main task of the project is to assess the degree of similarity between pairs of sentences, with the goal of developing methods for comparing the meaning of text fragments. The project focuses on three sub-tasks: monolingual, cross-lingual, and semantic relatedness classification. The monolingual sub-task involves evaluating sentence pairs in a given language, while the cross-lingual sub-task requires comparing sentences in different languages.

The semantic relatedness classification sub-task aims to determine whether sentence pairs are related or not In the scale of 0 to 5 also called Gold Standard, where 5 being the two input sentences are of same meaning and 0 being the the input sentences completely differ in the semantic meaning.

Here the monolingual task is between english - english sentences and cross-lingual task is between english and spanish sentences.

Dataset

Data can be obtained by merging data from the year 2012 to 2017 SemEval task. Every data point consists of two sentences and the score of the similarity of the sentence. Further data can be divided into train , validation and test after preprocessing.

Measure of evaluation

PCC

The Pearson's Correlation Coefficient (Pearson's CC) with gold labels provided, PCC a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between -1 and 1 (from least to highly correlated), and scores near to 0 imply poor to no correlation, is the evaluation metric for this task.

Baseline

Baseline approach for the task would be converting the sentences into vectors using a bag of words vectorizer and using cosine similarity to calculate the cosine similarity of the sentence.

Literature Survey and Existing Approaches :

The following papers provide a summary of each team's model which performed well in the Sem-Eval 2017 Task 1 competition.

- **ECNU :** This model is an average aggregation of four deep learning models and three feature engineered models. Three feature tailored models such as Random Forest, Gradient Boosting, and XGBoost employ regression techniques like tree kernels. Each network in neural network models supplies the networks that multiply, subtract, and concatenate paired sentence embeddings element by element.
- **BIT :** The three methods for measuring semantic textual similarity (STS) that are presented in this paper are all based on the jaccard coefficient of information content of the sentence pairs that are used as examples. Calculated the non-overlapping information content (IC) of two sentences to solve the issue. They improved their initial algorithm by adding words sequentially from each tier of the WordNet hierarchy taxonomy and computing information content gain iteratively because no searching for all subsume ideas is required. They attempted to combine sentence alignment and word embedding, respectively, as additional features to train supervised models in order to enhance performance in addition to computing the similarity score using a single IC feature in an unsupervised manner. Their findings indicate that word embedding along with IC produces the greatest outcomes.
- **CompiLIG :** Their approach incorporated supervised and unsupervised syntax-based, dictionary-based, context-based, and MT-based methods. In the context-based method, they use weighted distributed representation of words as sentence embedding and compute cosine similarity, where the weights are computed in dictionaries. In the dictionary-based method, they use two sets of words for a different language that can be obtained from Google Translate; and in the syntactical method, they compute cosine similarity of the n-gram representation of two sentences; the sum of weighted Jaccard distance of such set is then used to compute final score. For MT-based approach, they

use monolingual aligner to get aligned utterances and measure a variation of jaccard distance based on inverse document frequency of aligned utterances.

- **DT_Team :** They developed three different models with various features including similarity scores calculated using word and chunk alignments, word/sentence embeddings, and Gaussian Mixture Model. POS-tagging, name-entity recognition as well as normalization, tokenization, lemmatization are preprocess procedures for word embeddings; This manuscript from word embeddings to sentence embeddings explains word alignment. The similarity score was calculated as the total score for all aligned word pairs divided by the total length of the given sentence pair.

Rough estimate of project milestones:

Date	**Work Progress**
5th March	Literature Survey, Exploring relevant methodologies Data Collection and Preprocessing, Comparative Study
8th March	Interim Submission
25th March	Implementation of Final Model
10th April	Model Training and parameters tuning
20th April	Final Submission

Reference:

1. SemEval-2017 Task 1 : Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation https://arxiv.org/pdf/1708.00055.pdf
2. Semantic Textual Similarity Wiki http://ixa2.si.ehu.eus/stswiki/index.php/Main_Page

Literature survey :

3. ECNU : https://aclanthology.org/S17-2028/
4. BIT : https://aclanthology.org/S17-2007/
5. CompiLIG : https://arxiv.org/pdf/1704.01346.pdf
6. DT_Team : https://aclanthology.org/S17-2014.pdf