Survey on Different NLP models for Semantic Similarity

Raj Awate
Information Technology
MIT Academy of Engineering, Pune
rpawate@mitaoe.ac.in

Keshav Bajaj
Information Technology
MIT Academy of Engineering, Pune
kjbajaj@mitaoe.ac.in

Anilkumar Gupta
CDAC
Senior Member IEEE
akgupta5592@gmail.com

Abstract:

Calculating Semantic Similarity of Sentences helps in many Real-life application Which Includes developing an automatic grading system, Determining Repeatedly asked Questions on Quora, Stack overflow, and some of the other similar platforms. There are various Machine Learning models and Deep Learning techniques used to find out semantic similarity of sentences and this article presents accuracies, advantages, and disadvantages of various models such as Siamese network model, BERT, T5, used with different datasets to calculate similarity in sentences. This study will help various machine learning engineers while fine-tuning and using pre-trained machine learning models.

Keywords — transformers, semantic similarity

----------**********----------

I. INTRODUCTION

Most of the time questions that are being asked on platforms like Quora, stack-overflow, stack exchange is the same and they are being repeatedly asked by some other people also, so to sort this out NLP comes into play. Also, most of the university exams are becoming online these days, and this trend took exponential increases after this covid pandemic and if the university wants to ask any subjective questions to students, then again NLP comes into play. Linguistics is the study of the sounds, grammar and meaning in languages. The final objective in linguistics is to explain why patterns in languages are as they are. Linguistics tries to explain why phenomenon occur in languages using a descriptive approach, and find the rules people unconsciously follow when they speak and write. On the other hand, prescriptive approaches try to describe how people should speak and write and what rules of language people should know. In linguistics, semantics is the study of the meaning of words, their structure, and their relationships with other words. Semantic similarity has seen many solutions in a small amount of time. The central idea behind the most solution is that, the identification and alignment of semantically similar or related words across the two sentences and the aggregation of these similarities to generate an overall similarity. Most of the time questions that are being asked on platforms like Quora, stack-overflow, stack exchange is the same and they are being repeatedly asked by some other people also, so to sort this out NLP comes into play. Also, most of the university exams are becoming online these days, and this trend took exponential increases after this covid pandemic and if the university wants to ask any subjective questions to students, then again NLP comes into play.

Linguistics is the study of the sounds, grammar and meaning in languages. The final objective in linguistics is to explain why patterns in languages are as they are. Linguistics tries to explain why phenomenon occur in languages using a descriptive approach, and find the rules people unconsciously follow when they speak and write. On the other hand, prescriptive approaches try to describe how people should speak and write and what rules of language people should know. In linguistics, semantics is the study of the meaning of words, their structure, and their relationships with other words. Semantic similarity has seen many solutions in a small amount of time. The central idea behind the most solution is that, the identification and alignment of semantically similar or related words across the two sentences and the aggregation of these similarities to generate an overall similarity.

II. SEMANTIC SIMILARITY

A. State of the art approaches for finding the semantic textual similarity

After the introduction of BERT, we will discuss Some of the state-of-the-art methods for finding the semantic similarity in sentences. BERT is a pre-trained Transformers network model, which is used for various NLP tasks and it provides state-of-the-art results in those tasks, one of which is the sentence pair regression task. The input for the sentence pair regression consists of the pair of two sentences separated by [SEP] token, multi-head attention over 12 (base-model) or 24 layers (large-model) is applied and the output is passed to a simple regression function to de-rive the final label, BERT give a state-of-the-art performance on semantic textual similarity Benchmark. while Roberta showed improved performance on BERT [3]

RoBERTa:
RoBERTa stands for robustly optimized BERT pretraining approach. Plain design changes made in architecture and pre-training procedure gave significantly improved results. Those changes were:
Training with bigger batch sizes and longer sequences:
While training RoBERTa researchers trained model with 125 steps of 2k sequences and 31k steps with 8k sequences of batch size as larger batch size improves the perplexity on masked language modelling objective as well as end task accuracy

Dynamically changing the masking pattern:
In original BERT model the training data is duplicated 10 times and then masked with different strategy every time to avoid single static masking as masking is performed once during pre-processing.
Instead in RoBERTa Dynamic masking is being used in which masking is done differently every time we pass data in model

Removing the next sentence prediction objective:
In BERT, the model is trained to predict whether the next sentence is from same or different context with the help of auxiliary NSP (loss) and when researchers experimented with it, they concluded that removing this NSP loss increases performance of model for downstream tasks.

SBERT:
SBERT makes use of the BERT model for sentence pair regression task and Semantic textual similarity. moreover, it adds a pooling layer at the output of BERT and Roberta to get fixed sized outputs. Cosine similarity between two sentence embeddings u and v is computed and mean squared error loss function is used as objective function.[6]

$$\text{Similarity}(A, B) = \frac{A.B}{\|A\| x \|B\|}$$

A and B are sentence embedding vectors

Dataset used for finding semantic textual similarity by SBERT and RoBERTa
The STS benchmark (STSb) (Cer et al., 2017) pro-vides is a popular dataset to evaluate supervised STS systems. The data includes 8,628 sentence pairs from the three categories captions, news, and forums.

T5-11B
The T5 model i.e., Text to Text Transfer Transformer model was built while conducting a large-scale study for exploring the limits of transfer learning. The architectures like BERT,RoBERT,GPT,etc were the underlying models of T5 which performed good in Transfer Learning. The model like BERT were successful in various kinds of NLP task's but there we needed to connect a layer which was specific to the task. Suppose I want to do a sentiment analysis of a Tweet then in BERT then we would need to add a layer of Neural networks on top of BERT whose output would be a vector which will gives us the probability of each corresponding sentiment analysis but in T5 we directly convert this task in text to text i.e., the input would be the tweet in text and output would be in text i.e., "0" string for negative sentiment or "1" string for positive sentiment. In case of semantic similarity, we would give the T5 Model the input in the from "Text1:Text2" and it will output it as "0.98" i.e., the semantic similarity between the text [5]

Big Bird
As the name suggest this was a transformer model designed to understand the longer sequence of text like Articles, Books, etc.There are models like BERT which are remarkably successful but as we have longer sequenced the quadratic dependency on the input length is the biggest drawback. The Big Bird model was introduced to solve the problem and is helpful in Longer Sequence of Data.
In attention mechanism in BERT what we do is we give every token a higher level of representation using other tokens of the sequence length but this is the reason which causes the quadratic dependency on input length. The BIGBIRD consist of four types of Attention Mechanisms as follows
The above diagram represents an adjacency matrix representation of attention i.e., the coloured part will represent the connection between the input token and output token for attention mechanism
1. Random Attention:
In Random attention Mechanism makes connection between input and output token for each token constant times.
2. Window Attention:
The Window Attention mechanism makes connection between neighbouring tokens for each token constant times
3. Global Attention
The Global Attention Mechanism forms an intermediary connection between every token which with the help of given above mechanism by using multiple layers give a sense of full attention mechanism.

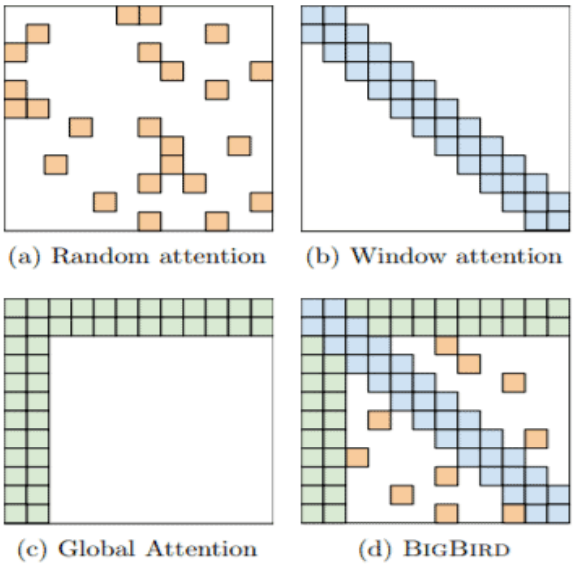

Figure 1: Building blocks of the attention mechanism used in BIGBIRD. White color indicates absence of attention. (a) random attention with r = 2, (b) sliding window attention with w = 3 (c) global attention with g = 2. (d) the combined BIGBIRD model. [4]

Sr No	Model	Pearson	Spearman	Advantages	Drawbacks
1	SBERT(STS-base)	----	0.8467	Computational time to get sentence embeddings required is much less than that of BERT	Its Structure makes it difficult to use for many other NLP tasks
2	RoBERTa	0.922	----	More Optimised in terms of design and architecture	While the replication study is well appreciated, the novelty contribution of the Roberta is marginally incremental as the model structure is unchanged from BERT. The other techniques applied also are trivial. - Little can be deduced from the experiments, as performance is often improved by training over more data.
3	T5-11B	0.925	0.921	It is a text-to-text transformer model hence it is generalized way for doing Downstream NLP tasks	T5 is a huge model with 11B parameters and hence would require huge computational cost. There are also different smaller versions of T5 but the model's accuracy will suffer with the lower number of parameters
4	BigBird	----	0.878	Big Bird was developed to so that we can work on input text which have a larger input length. It is helpful in the tasks where the we need to work on input texts like Long Articles,etc	In a worst-case scenario sparse attention might require larger number of layers to be stacked and might result into quadratic size dependency

CONCLUSIONS

After reviewing those many papers and with the help of spearman and pearson score, we get for semantic similarity of sentences it can be concluded that t5-11b can perform better for measuring similarity on increasing the cost of computation.

REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, JakobUszkoreit, Llion Jones, Aidan N Gomez, ŁukaszKaiser, and Illia Polosukhin. 2017. Attention is Allyou Need. In I. Guyon, U. V. Luxburg, S. Bengio,H. Wallach, R. Fergus, S. Vishwanathan, and R. Gar-nett, editors,Advances in Neural Information Pro-cessing Systems 30, pages 5998–6008.

[2] Alexis Conneau, Douwe Kiela, Holger Schwenk, Lo ̈icBarrault, and Antoine Bordes. 2017. SupervisedLearning of Universal Sentence Representationsfrom Natural Language Inference Data. InProceed-ings of the 2017 Conference on Empirical Methodsin Natural Language Processing, pages 670–680, Copenhagen, Denmark. Association for Computa-tional Linguistics.

[3] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[4] Zaheer, Manzil, et al. "Big bird: Transformers for longer sequences." *arXiv preprint arXiv:2007.14062* (2020).

[5] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." arXiv preprint arXiv:1910.10683 (2019).

[6] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint arXiv:1908.10084* (2019).

\