**Bilingual Semantic Text Similarity**

MT Evaluation

(Bilingual Evaluation)

Submitted By

**Binod Kumar Mishra, Pursotam Kumar, Vaibhavi Kailash Kothadi**

Mentor by

**Ms. Ananya Mukherjee, Dr. Manish Shrivastava**



**11th Advanced Summer School on NLP**

**(IASNLP-2022)**

**IIIT Hyderabad**

1. **Introduction:**

Bilingual semantic text similarity measures the equivalence of meanings between two languages. Generally, With the help of bilingual distributed representation of words, we try to capture semantic meaning. It can be divided into three phases, Preprocessing, Similarity Measure, and Score estimation [1].

When we try to build a Machine Translation system at that time, we should focus more on how much it is accurately translated to other languages. For translating text from one language to other languages different approach is available, which include Rule-based, Statistical based, and Neural based approach. In this report, we are focusing only on the neural-based approach.

For building a Neural network-based Machine Translation model, we require numbers to train the model because the machine can understand only numbers. And for this thing, we can take help from different word Embedding methods. For training Neural networks, we can build our own embedding model or can use a pre-trained model.

1. **Problem Statement:**

Our work is divided into three parts.

* The first part is to study various word-level embeddings and sentence-level embeddings.
* Second part is to fine-tune COMET machine translation evaluation model using our own dataset.
* Third part is to find out Pearson correlation using Normalized score and other sentence embedding approaches.

1. **Brief Literature Survey:**

There are various Pre-Trained Word Embedding Models available, like Word2Vec [2, 3], Glove[4], Fasttext [5, 6], ELMO[7], BERT[8], etc. These are word-level vector representations.

Sentence level embeddings are SBERT [9], LaBSE [10], and LASER [11] which provide a single vector representation for each sentence.

COMET[12]: It is a machine translation Evaluation method to find out how much it is correctly translated in other languages.

Eneko Agirre [13] focused on Semantic textual similarity between English and Spanish language.

Nils Reimers & Iryna Gurevych [9] concluded in their paper that, the construction of BERT makes it unsuitable for semantic similarity search as well as for unsupervised tasks like clustering.

We should extend BERT to Sentence-BERT (SBERT). It is a modified model of BERT. It uses Siamese and Triplet network to capture semantic meaningful of any sentence. This will not help to reduce the computational resources to find distance between two sense-pair but also reduce the time from hours to seconds.

1. **Dataset:**

We have collected dataset from [14] containing 1,00,000 parallel English and translated Hindi sentences to fine-tune COMET machine translation model .

1. **Proposed Approach to solve the problem:**

To fine-tune COMET machine translation evaluation models, we require Training and validation dataset. So, we split dataset into two parts and then train our model. The flow diagram is shown in fig 1.

Parallel Corpus

Input sentence

English-Hindi

Validation data

Training data

Pre trained COMET Model

Score

Fig 1: Fine-tune COMET machine translation evaluation model

1. **Languages:** Hindi, English.
2. **Technology Used:** Python, Numpy, Pandas, Py-torch/Tensor Flow & Keras
3. **Result:**

We fine-tune the COMET machine translation evaluation model for better accuracy. For this thing, we used 100,000 parallel English and translated Hindi sentences. After training with our own dataset, we get a better result. The results in terms of Loss and Accuracy are shown in table 1.

Table 1: Result of Fine tune COMET

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No.** | **Epoch** | **Loss** | **Accuracy** |
|  | Epoch 0 | 0.000346 | 0.798 |
|  | Epoch 1 | 0.00017 | 0.849 |
|  | Epoch 2 | 0.0000771 | 0.870 |

The result of cosine distance between vector representation of English sentences and translated Hindi sentences is shown in fig. 2 using LaBSE, SBERT, COMET and LASER.

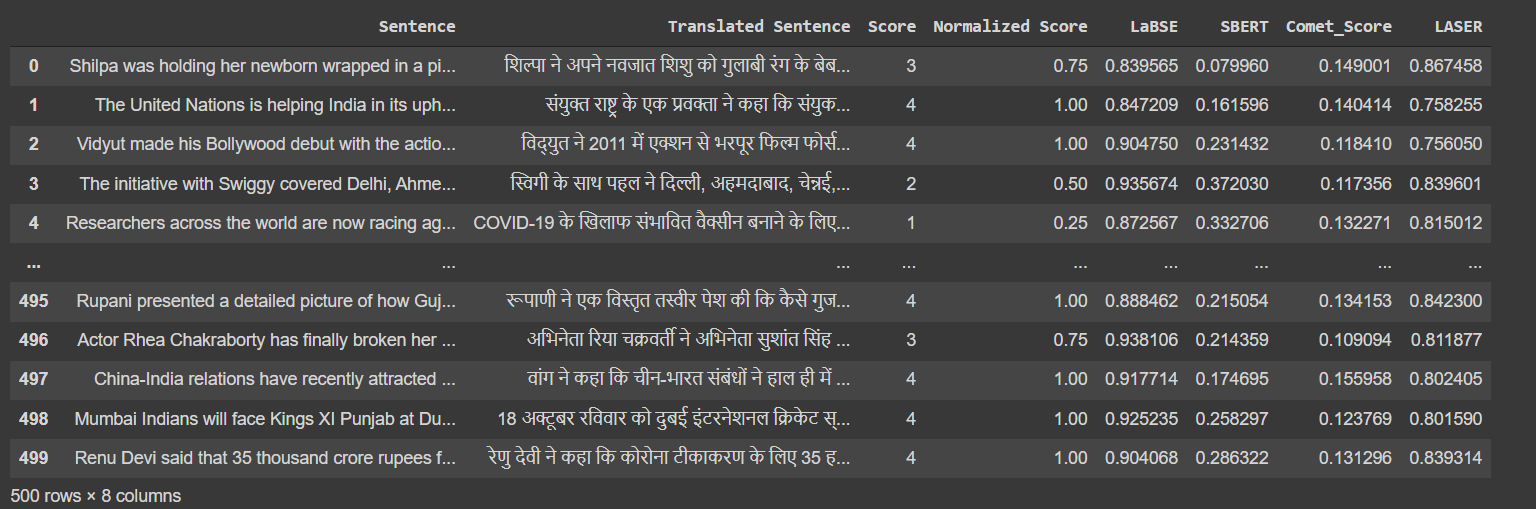


Fig 2: Cosine distance between pair of English Sentence and

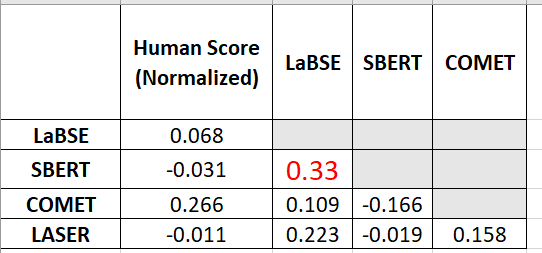
translated Hindi sentences using LaBSE, SBERT, and LASER.

We also find out the Correlation Pearson value between Each sentence level embeddings. With the help of this, we can find out how much both are strongly associated. The Correlation Pearson value (r) can be calculated by using equation (i)

…….. (i)

Whereas x and y are two vectors of length n and mx and my corresponds to the means of x and y respectively. If the value r is negative, it means that Negative Correlation, positive means Positive Correlation and 0 means no correlation. The comparison result of correlation pearson is shown on table 2.

**Table 2: Comparison Correlation Pearson**



1. **Conclusion & Future Work:**

In this report we focus three important thing, first one is related to Embedding approach, second is fine tune COMET machine translation evaluation model and find out which two approach gives better correlation pearson result. We concluded that out of all LaBSE and SBERT gives better result.

The future work is that if we Fine tune with huge data then it may be possible that we get better result.

**References:**

1. Shajalal, M. and M. Aono, *Semantic textual similarity between sentences using bilingual word semantics.* Progress in Artificial Intelligence, 2019. **8**(2): p. 263-272.

2. Mikolov, T., et al., *Efficient estimation of word representations in vector space.* 2013.

3. Mikolov, T., et al. *Distributed representations of words and phrases and their compositionality*. in *Advances in neural information processing systems*. 2013.

4. Pennington, J., R. Socher, and C.D. Manning. *Glove: Global vectors for word representation*. in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.

5. Joulin, A., et al., *Fasttext. zip: Compressing text classification models.* 2016.

6. Bojanowski, P., et al., *Enriching word vectors with subword information.* 2017. **5**: p. 135-146.

7. Ulčar, M. and M.J.a.p.a. Robnik-Šikonja, *High quality ELMo embeddings for seven less-resourced languages.* 2019.

8. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.

9. Reimers, N. and I.J.a.p.a. Gurevych, *Sentence-bert: Sentence embeddings using siamese bert-networks.* 2019.

10. Feng, F., et al., *Language-agnostic bert sentence embedding.* 2020.

11. Artetxe, M. and H.J.T.o.t.A.f.C.L. Schwenk, *Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond.* 2019. **7**: p. 597-610.

12. Rei, R., et al., *COMET: A neural framework for MT evaluation.* 2020.

13. Agirre, E., et al. *SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability*. 2015. Denver, Colorado: Association for Computational Linguistics.

14. Kunchukuttan, A., et al., *Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages.* 2020.