IASNLP 2022

The 11th Advanced Summer School on NLP (IASNLP-2022)

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

PROJECT REPORT

TITLE : Feature Based MT Evaluation

Project Member:

1) Shubham Varma

2) Suman Saurabh

Mentor:

Ananya Mukherjee (LTRC)

TABLE OF CONTENT

Introduction

Machine Translation:

Machine translation (MT) is automated translation. It is the process by which computer software is used to translate a text from one natural language (such as English) to another (such as Spanish).To process any translation, human or automated, the meaning of a text in the original (source) language must be fully restored in the target language, i.e. the translation. While on the surface this seems straightforward, it is far more complex. Translation is not a mere word-for-word substitution. A translator must interpret and analyze all of the elements in the text and know how each word may influence another. This requires extensive expertise in grammar, syntax (sentence structure), semantics (meanings), etc., in the source and target languages, as well as familiarity with each local region.Human and machine translation each have their share of challenges. For example, no two individual translators can produce identical translations of the same text in the same language pair, and it may take several rounds of revisions to meet customer satisfaction. But the greater challenge lies in how machine translation can produce publishable quality translations.

Machine translation Evaluation

Evaluation of translations, which resulted from Machine Translation. The evaluation can be manual or automatic. Automatic evaluation is done using an evaluation metric. The evaluation metric assigns a quality score to a translation by comparing it to the reference translation.

Parameters of Machine Translation Evaluation

There is a need for evaluating a translation for the end user or reader, irrespective of whether the translation is produced by a human being or a machine. The following are the parameters of evaluation.

- Comprehensibility– Is the reader able to comprehend the translation correctly and easily?
  The above may be broken up into two parameters:
  1.Adequacy– Correctness or accuracy of translation
  2.Fluency– Ease of understanding or readability of translation by the reader.

The first one pertains to whether the information is given correctly in the translated sen-tence/text. The second one is about whether the reader is able to grasp the information, which is why fluency is important besides convenience to the reader. The former may be judged using information theoretic aspects, whereas the latter is cognitive.

For evaluating a translation for system development the primary aim of the evaluation is to help the developer of the MT system. This would require identifying the weakness in the MT system, and then suggesting ways to overcome it.It normally requires either the analysis of the source or target sentences/text to identify areas where the system is performing weakly, or the analysis of the weakness in the approach adopted to build the MT system.

Different Types of MT Evaluation Metrics

BLEU

BLEU was one of the first metrics to report a high correlation with human judgments of quality. The metric is currently one of the most popular in the field. The central idea behind the metric is that "the closer a machine translation is to a professional human translation, the better it is". The metric calculates scores for individual segments, generally sentences — then averages these scores over the whole corpus for a final score. It has been shown to correlate highly with human judgments of quality at the corpus level.BLEU uses a modified form of precision to compare a candidate translation against multiple reference translations. The metric modifies simple precision since machine translation systems have been known to generate more words than appear in a reference text. No other machine translation metric is yet to significantly outperform BLEU with respect to correlation with human judgment across language pairs.

NIST

The NIST metric is based on the BLEU metric, but with some alterations. Where BLEU simply calculates n-gram precision adding equal weight to each one, NIST also calculates how informative a particular n-gram is. That is to say, when a correct n-gram is found, the rarer that n-gram is, the more weight it is given.For example, if the bigram "on the" correctly matches, it receives lower weight than the correct matching of bigram "interesting calculations," as this is less likely to occur. NIST also differs from BLEU in its calculation of the brevity penalty, insofar as small variations in translation length do not impact the overall score as much.

Word error rate

The Word error rate (WER) is a metric based on the Levenshtein distance, where the Levenshtein distance works at the character level, WER works at the word level. It was originally used for measuring the performance of speech recognition systems but is also used in the evaluation of machine translation. The metric is based on the calculation of the number of words that differ between a piece of machine-translated text and a reference translation.

A related metric is the Position-independent word error rate (PER), which allows for the re-ordering of words and sequences of words between a translated text and a reference translation.

METEOR

The METEOR metric is designed to address some of the deficiencies inherent in the BLEU metric. The metric is based on the weighted harmonic mean of unigram precision and unigram recall. The metric was designed after research by Lavie (2004) into the significance of recall in evaluation metrics. Their research showed that metrics based on recall consistently achieved higher correlation than those based on precision alone, cf. BLEU and NIST. METEOR also includes some other features not found in other metrics, such as synonym matching, where instead of matching only on the exact word form, the metric also matches on synonyms. For example, the word "good" in the reference rendering as "well" in the translation counts as a match. The metric also includes a stemmer, which lemmatises words and matches on the lemmatized forms. The implementation of the metric is modular insofar as the algorithms that match words are implemented as modules, and new modules that implement different matching strategies may easily be added.

LEPOR

A new MT evaluation metric LEPOR was proposed as the combination of many evaluation factors including existing ones (precision, recall) and modified ones (sentence-length penalty and n-gram based word order penalty). The experiments were tested on eight language pairs from ACL-WMT2011 including English-to-other (Spanish, French, German, and Czech) and the inverse, and showed that LEPOR yielded higher system-level correlation with human judgments than several existing metrics such as BLEU, Meteor-1.3, TER, AMBER and MP4IBM1.An enhanced version of LEPOR metric, hLEPOR, is introduced in the paper.hLEPOR utilizes the harmonic mean to combine the sub-factors of the designed metric. Furthermore, they design a set of parameters to tune the weights of the sub-factors according to different language pairs. The ACL-WMT13 Metrics shared task results show that hLEPOR yields the highest Pearson correlation score with human judgment on the English-to-Russian language pair, in addition to the highest average-score on five language pairs (English-to-German, French, Spanish, Czech, Russian). The detailed results of the WMT13 Metrics Task is introduced in the paper.

Approach

As we saw above there are several evaluation metrics which compute the final translation score by computing individual feature scores (lexical, morphological, synonym similarity (word level) and sentence semantics similarity). Our aim is to train our model such that optimal weights are learned for these linguistic features, resulting in having high correlation with humans.

In our approach we are using word embedding to find out the similarity between the reference sentence and the machine translated sentences on the basis of the similarity score if the score is above 0.7 then it is an synonym match, if the score is between 0.5 and 0.7 then it is a root match and if the words are matching exactly then it’s an exact match, F-beta(3) score for the above features are calculated. We will treat these features as our input and train the weights of these input features against the COMET score generated for the input sentences. We have tried different algorithms such as LSTM, ANN, Random Forest, Decision Tree, SVR, Ensemble learning algorithms such as XGBoost, Adaboost, etc and tried to tune there hyperparameters to get the best pearson correlation coefficient possible.The testing dataset consists of 500 sentences on which these different algorithms are tested and the pearson correlation coefficient with the human is calculated.

$$F_{mean} = \frac{P \cdot R}{\alpha P + (1 - \alpha) R}$$

Example: (For demo purpose we are taking a English Example)

Suppose our Sentence:

Reference sentence: India is on the verge of becoming third largest economy in few years.

MT output sentence: India is going to become third largest economy in coming years.

1. Lexical Matching: (India, is, third, largest, economy,years,in)

   Remaining Words in Both Sentences:

   Reference sentence: on the verge of becoming few

   MT output sentence: going,to,become,coming

2. Stem Matching: (become from [becoming,become])

   Reference sentence: on the verge of few

   MT output sentence: going to,coming

3. Synonym Matching: on the verge -> going to (Similar Phrases).

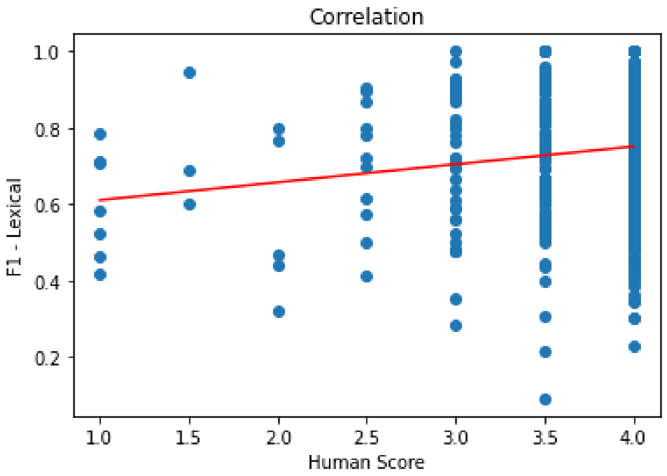
   But ‘coming’ and ‘few’ are not synonyms but they convey the same intent in the context of a given sentence which we will deal with using different advanced methods.

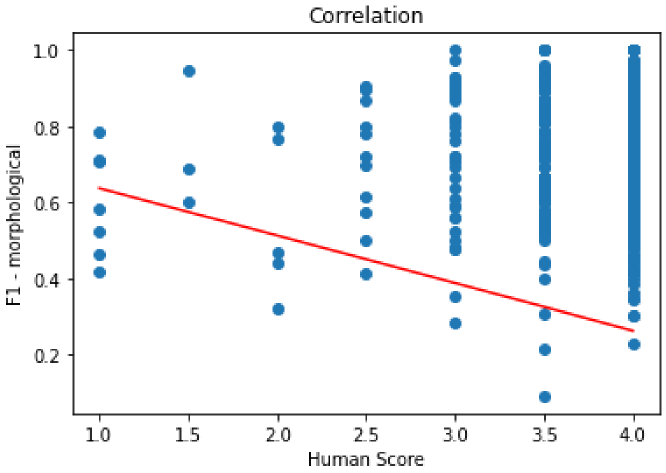
Tools & technology use

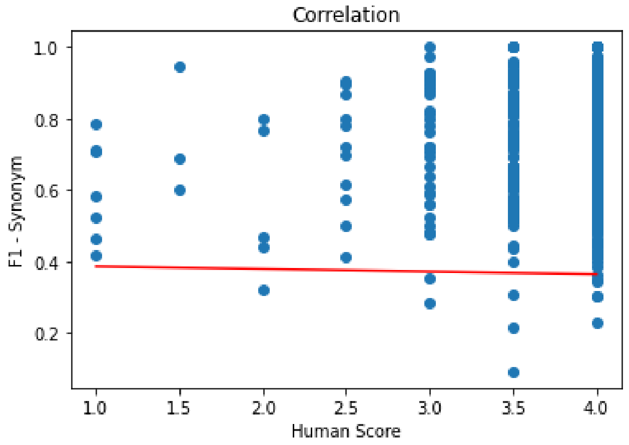
- Python Version - 3.9
- TensorFlow Version - 2.9
- Pip Version - 22.1.2
- PyTorch Version - 21.02
- GPU - Nvidia K80/T4
- GPU Memory - 12 GB
- Platform - Google Colab, Kaggle
- Embedding - FastText Embedding
- Evaluation Metric (Gold Standard) - COMET

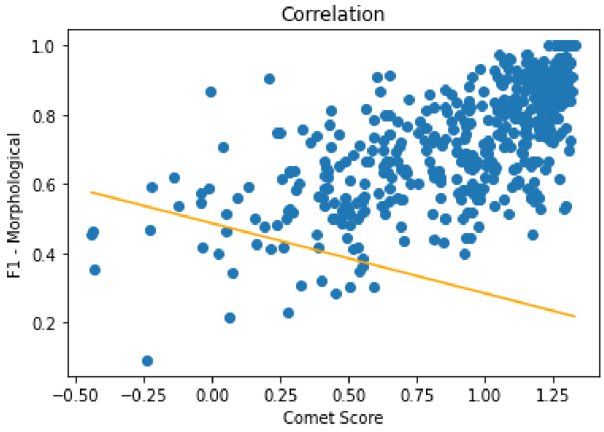
Analysis of dataset

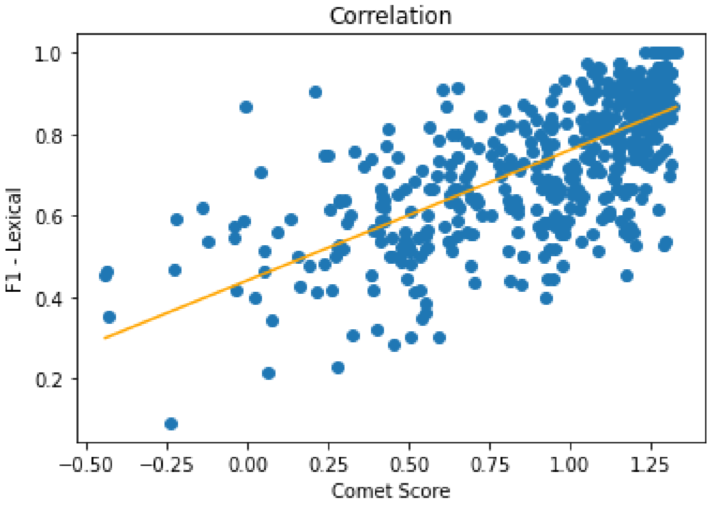
Testing data:

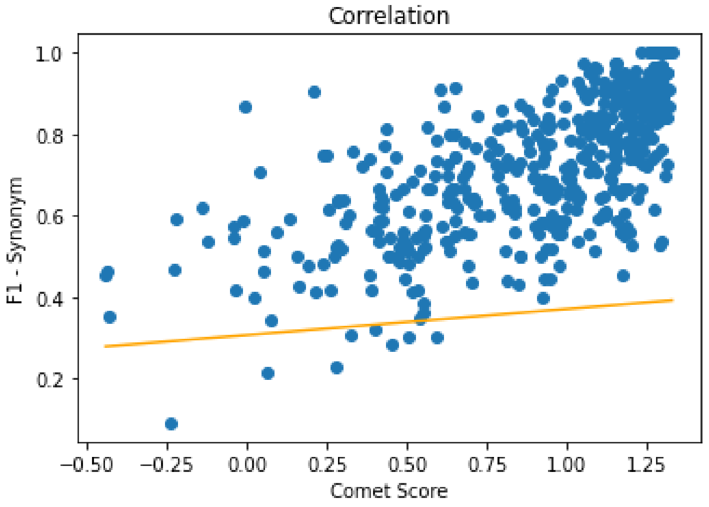


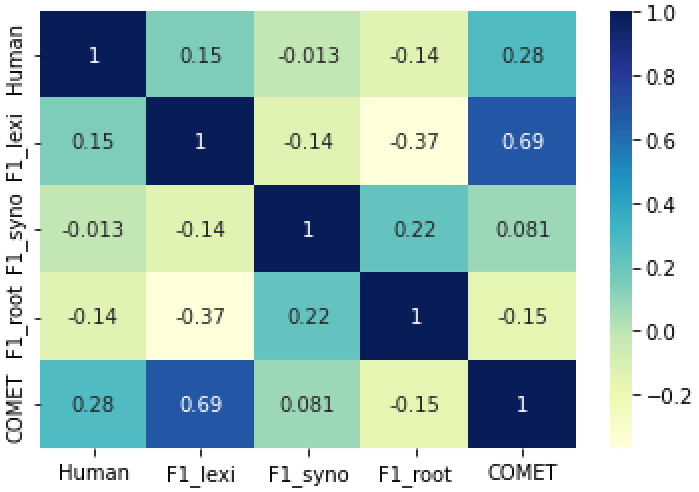




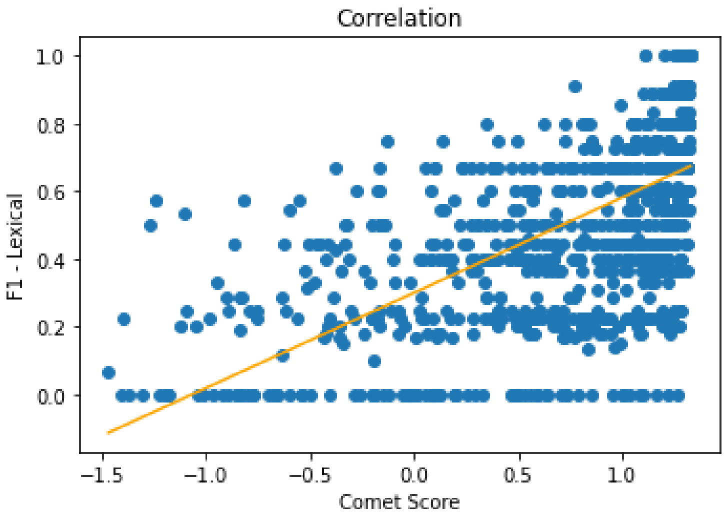


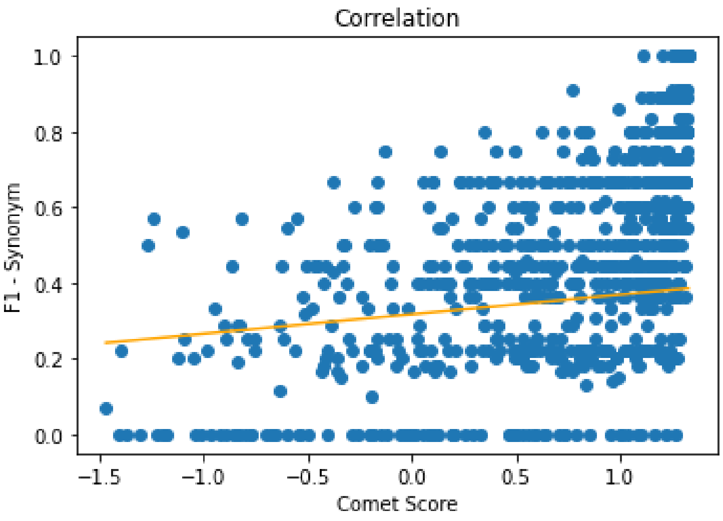


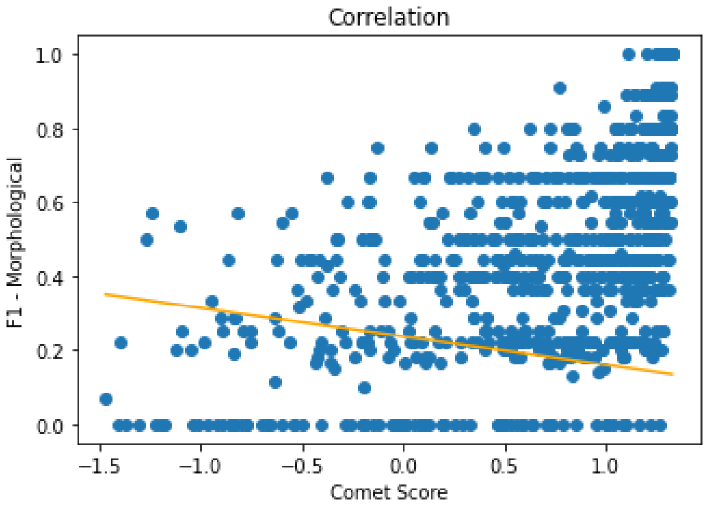




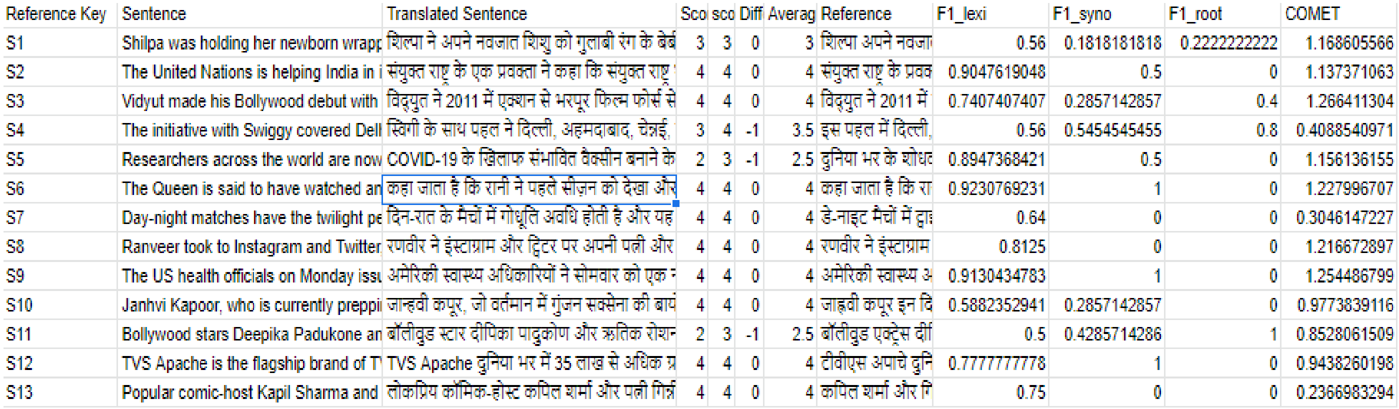
Training Data:







Sample of Dataset:



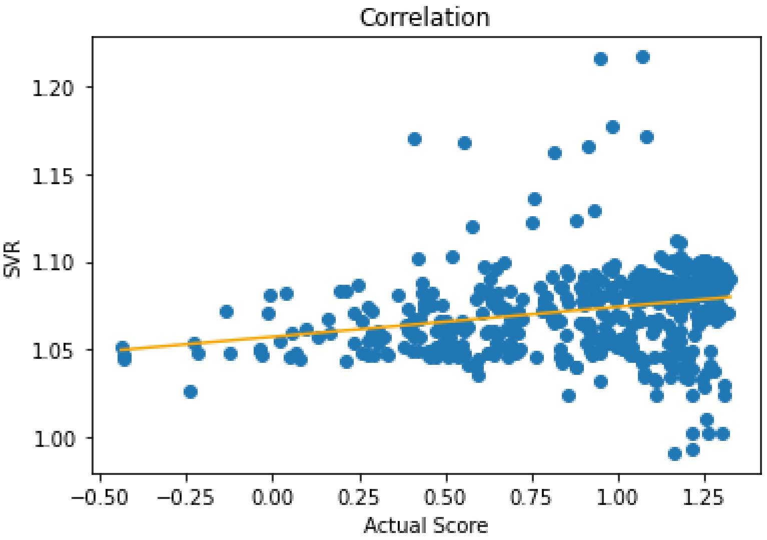
Reference Key	Sentence	Translated Sentence	Sco	sco	Diff	Averag	Reference	F1_lexi	F1_syno	F1_root	COMET
S1	Shilpa was holding her newborn wrapp	शिल्पा ने अपने नवजात शिशु को गुलाबी रंग के बेब	3	3	0	3	शिल्पा अपने नवजा	0.56	0.1818181818	0.2222222222	1.168605566
S2	The United Nations is helping India in i	संयुक्त राष्ट्र के एक प्रवक्ता ने कहा कि संयुक्त राष्ट्र	4	4	0	4	संयुक्त राष्ट्र के प्रवक	0.9047619048	0.5	0	1.137371063
S3	Vidyut made his Bollywood debut with	विद्युत ने 2011 में एक्शन से भरपूर फिल्म फोर्स से	4	4	0	4	विद्युत ने 2011 में	0.7407407407	0.2857142857	0.4	1.266411304
S4	The initiative with Swiggy covered Dell	स्विगी के साथ पहल ने दिल्ली, अहमदाबाद, चेन्नई,	3	4	-1	3.5	इस पहल में दिल्ली,	0.56	0.5454545455	0.8	0.4088540971
S5	Researchers across the world are now	COVID-19 के खिलाफ संभावित वैक्सीन बनाने के	2	3	-1	2.5	दुनिया भर के शोध	0.8947368421	0.5	0	1.156136155
S6	The Queen is said to have watched an	कहा जाता है कि रानी ने पहले सीज़न को देखा और	4	4	0	4	कहा जाता है कि रा	0.9230769231	1	0	1.227996707
S7	Day-night matches have the twilight pe	दिन-रात के मैचों में गोधूलि अवधि होती है और यह	4	4	0	4	डे-नाइट मैचों में ट्वा	0.64	0	0	0.3046147227
S8	Ranveer took to Instagram and Twitter	रणवीर ने इंस्टाग्राम और ट्विटर पर अपनी पत्नी और	4	4	0	4	रणवीर ने इंस्टाग्राम	0.8125	0	0	1.216672897
S9	The US health officials on Monday issu	अमेरिकी स्वास्थ्य अधिकारियों ने सोमवार को एक न	4	4	0	4	अमेरिकी स्वास्थ्य अ	0.9130434783	1	0	1.254486799
S10	Janhvi Kapoor, who is currently preppi	जान्हवी कपूर, जो वर्तमान में गुंजन सक्सेना की बायो	4	4	0	4	जाह्नवी कपूर इन दि	0.5882352941	0.2857142857	0	0.9773839116
S11	Bollywood stars Deepika Padukone an	बॉलीवुड स्टार दीपिका पादुकोण और ऋतिक रोशन	2	3	-1	2.5	बॉलीवुड एक्ट्रेस दी	0.5	0.4285714286	1	0.8528061509
S12	TVS Apache is the flagship brand of T	TVS Apache दुनिया भर में 35 लाख से अधिक प्र	4	4	0	4	टीवीएस अपाचे दुनि	0.7777777778	1	0	0.9438260198
S13	Popular comic-host Kapil Sharma and	लोकप्रिय कॉमिक-होस्ट कपिल शर्मा और पत्नी गिन्नी	4	4	0	4	कपिल शर्मा और गि	0.75	0	0	0.2366983294

- We can see there is a maximum correlation of lexical matching with human score, obviously we understand the reason behind this maximum correlation, as if there are more common words between translated sentence and reference sentence there will be higher chances of correct translation.
- However there is poor correlation between morphological matching and human rating, the reason behind this may be due to fewer cases where words with the same root word are in different morphological form.
- However we also observed COMET score correlation with human score for sentences is only 0.28. Which also made us think about the results which we may get on training with a COMET score.
- There can be other reasons for low morphological correlation with human score, like if some word is coming in different morphological forms in two different sentences so there may be high chances of wrong translation because morphologically same words carry different meaning in the sentences, like if some words are appearing as running and in translated sentence it is ran then definitely both the words are carrying different meaning to the sentence, therefore this explains the reason of low morphological correlation.
- Now if we see semantic correlation with human rating there is some positive correlation with human rating which is obvious since here our cosine similarity will give a good score for words carrying similar meaning but they are not morphologically the same.

Result and Analysis

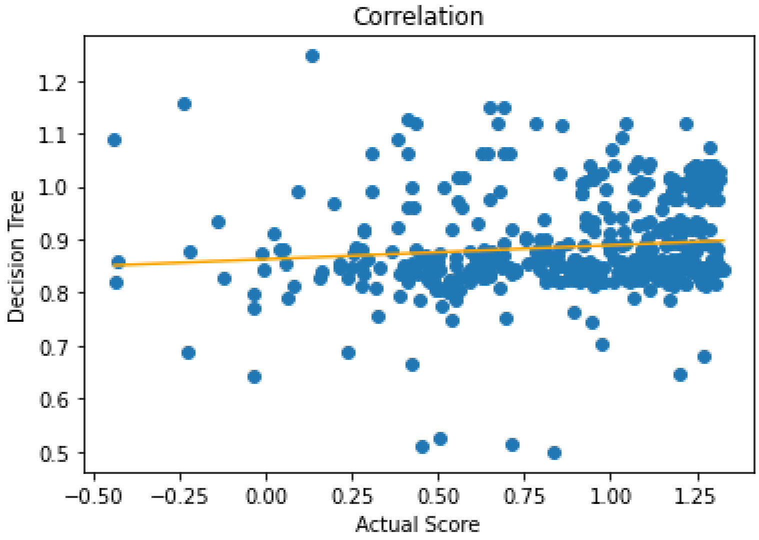
1. SVR

- kernel='rbf'
- C= 1e3
- gamma= 0.1
- Rest all parameters are default parameters
- After trying for different hyper parameters we got the best results of prediction with **Pearson' s Correlation Coefficient as 0.25**



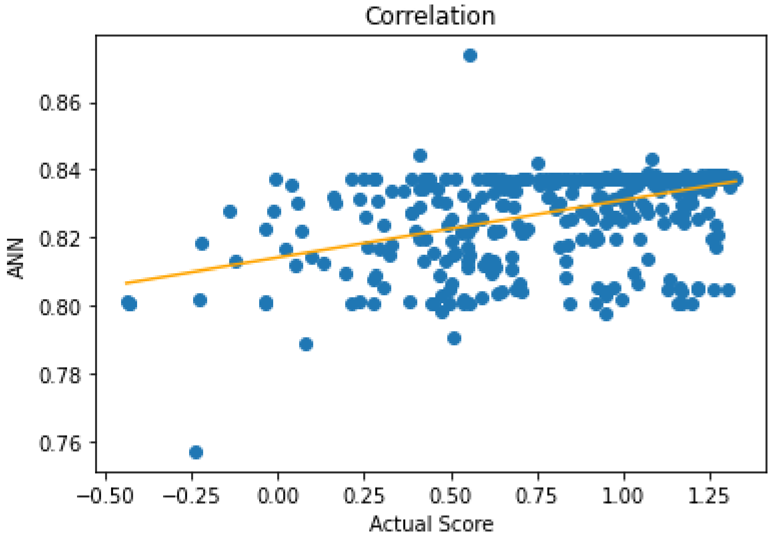
2. Decision Tree

- *criterion='squared_error', splitter='best', max_depth=None, min_samples_split=2, min_samples_leaf=1,*
- *Rest default parameter*
- ***Obtained Pearson's correlation coefficient of 0.2***



3. ANN

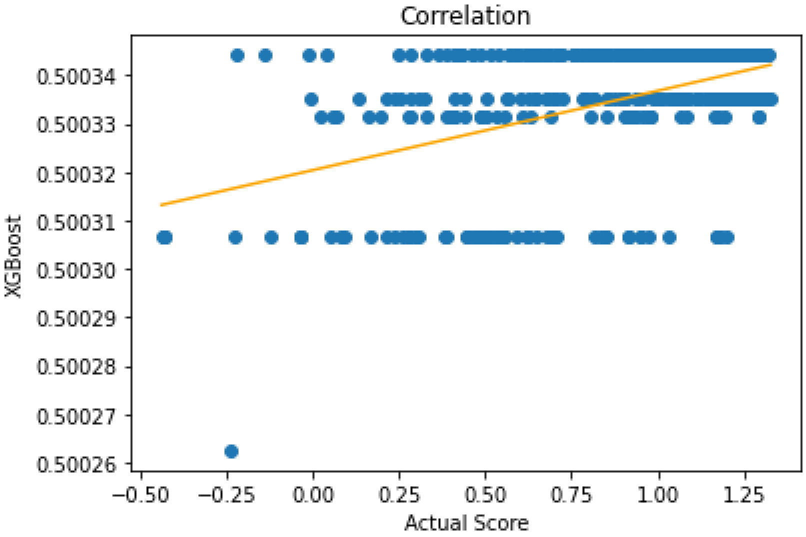
- 4 Hidden Layer(Dense Layer)) with 300-300-100-56 nodes in each layer respectively
- Optimizer = Adam
- Loss Function: Mean Square Error
- Batch Size: 30
- MSE:0.3599
- ***Obtained Pearson's correlation coefficient of 0.55***



4. **XGBoost**

Parameters:

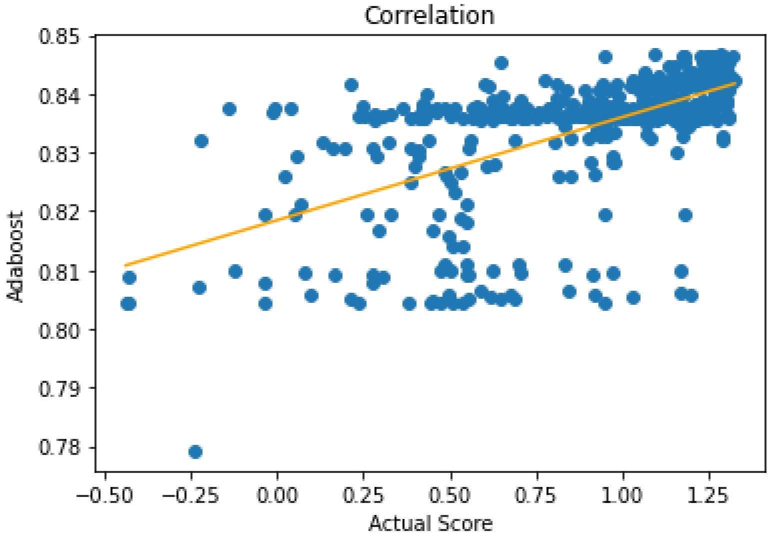
- Objective: reg:linear
- random_state = 0
- learning_rate = 0.001
- N_estimator = 100
- Other are constant
- **Pearson Correlation coefficient = 0.482**



5. **AdaBoost**

Hyper Parameters:

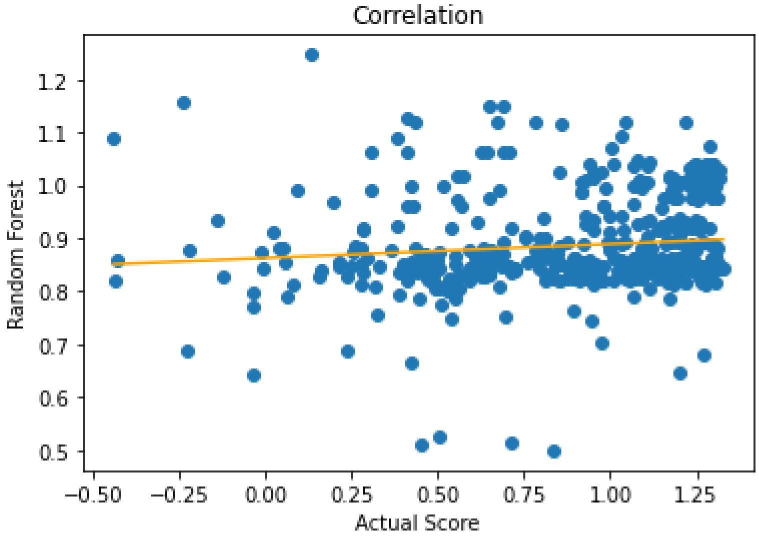
- random_state = 0
- n_estimator = 100
- Learning_rate = 0.001
- loss = linear
- Other are constant
- **Pearson Correlation coefficient = 0.482**



6. **Random Forest**

Hyper parameters

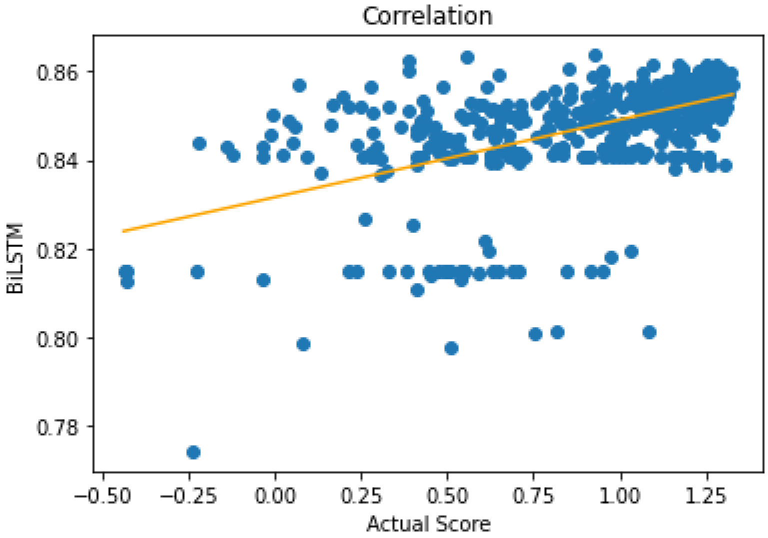
- n_estimators=150
- random_seed = 42
- criterion='squared_error'
- Rest of all the parameters were default.
- We tried with different number of estimators like 50,100,150,200,250 but the best result we get on n_estimators = 100 with **Pearson Correlation coefficient = 0.15**



7. BiLSTM

Hyperparameters:

- Two Bi-Directional LSTM Layers with number of nodes: 64-20
- Dropout: 0.5
- Optimizer : rmsprop
- Loss Function: Mean Square Error
- batch_size=30



Conclusion

- Our Dataset is of complex nature however there are only 3 features but still to find the best model which can give maximum pearson correlation coefficient value with human rating we tried 7 different models.
- Lexical features are most correlated with human rating and COMET score while morphological features are showing very poor correlation with Human and COMET score.
- Deep Learning models are performing better than Machine learning models, with their dense architecture we are able to capture the hidden pattern in data and the predicted scores are better correlated with human rating.
- Machine learning models have capability to learn the function but our dataset have some complex pattern which need rigorous training, therefore advance machine learning models such as AdaBoost, Xgboost are performing well on training data and test data

Future Work

- While evaluating features and evaluating F1 scores, we observed there may be some cases where if reference sentence consist of many words and the other translated sentence have few words,cases may arise where after doing lexical matching, Morphological matching,Semantic matching, many few words may remain unmatched which may indicate poor translation.
- Therefore we can introduce penalty features in our metric while evaluating matching and evaluating features value(especially for free word language families), which will definitely improve the model.

References

https://machinelearningmastery.com/adaboost-ensemble-in-python/

https://towardsdatascience.com/predictive-analysis-rnn-lstm-and-gru-to-predict-water-consumption-e6bb3c2b4b02

https://towardsdatascience.com/lstm-and-bidirectional-lstm-for-regression-4fddf910c655

https://towardsdatascience.com/predictive-analysis-rnn-lstm-and-gru-to-predict-water-consumption-e6bb3c2b4b02

https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe

https://www.mygreatlearning.com/blog/adaboost-algorithm/

https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/

https://www.analyticsvidhya.com/blog/2021/08/a-walk-through-of-regression-analysis-using-artificial-neural-networks-in-tensorflow/

https://machinelearningmastery.com/regression-tutorial-keras-deep-learning-library-python/

https://www.tensorflow.org/tutorials/keras/

https://www.tensorflow.org/text/guide/word_embeddings

https://www.tensorflow.org/text/guide/bert_preprocessing_guide

https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda

https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680

https://towardsdatascience.com/random-forest-hyperparameters-and-how-to-fine-tune-them-17aee785ee0d

https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/

https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0#:~:text=Support%20Vector%20Regression%20is%20a,the%20maximum%20number%20of%20points.

https://iiitaphyd-my.sharepoint.com/personal/iasnlp_iiit_ac_in/_layouts/15/onedrive.aspx?ga=1&id=%2Fpersonal%2Fiasnlp%5Fiiit%5Fac%5Fin%2FDocuments%2FIASNLP%2D2022%2F29%2D06%2D2022%2FMachine%2DTranslation%2DEvaluation%2DBy%2DProf%2ERajeev%5FSangal%2Fmteval%2Dframework%2Dfeb2022%2Dsangal%2Darxive%2Epdf&parent=%2Fpersonal%2Fiasnlp%5Fiiit%5Fac%5Fin%2FDocuments%2FIASNLP%2D2022%2F29%2D06%2D2022%2FMachine%2DTranslation%2DEvaluation%2DBy%2DProf%2ERajeev%5FSangal

https://en.wikipedia.org/wiki/Evaluation_of_machine_translation#:~:text=It%20was%20originally%20used%20for,text%20and%20a%20reference%20translation.