Abhas Goyal[1], **Anshul Agarwal**[2], **Harshit Itondia**[3], **Vedant Joshi**[4], **Vijay Kumar**[5]
[1]18817013, [2]190159, [3]190369, [4]18816856, [3]18817858
[1]ECO, [2]ME, [3]MSE, [4]ECO, [5]ECO,
{abhasg, anshulag, itondia, vedantj, vijayk}@iitk.ac.in

Abstract

With the advent of Natural Language Generation (NLG) tasks, thousand of models are trained on terabytes of data to generate text. Evaluating the quality of generated text is a crucial task of NLG. A variety of evaluation metrics for NLG are available, each with its own strengths and weaknesses. Several evaluation metrics, including conventional metrics like BLEU (bilingual evaluation understudy), ROUGE (recall-oriented understudy for Gisting Evaluation), and METEOR (Metric for Evaluation of Translation with Explicit Ordering), have been presented for evaluating NLG models. However, these metrics correlates poorly with human judgements. There are other machine learned metrics like BERTScore which uses heavy model like BERT to calculate similarity between human annotated and machine generation text.Apart from this there is gap in research with regards to text generation of regional Indian languages . We propose a composite evaluation metric that captures semantic, syntactic and pragmatics of text using various kinds of deep learning models,specifically for English to Hindi Machine translation task,with an aim to be able to reduce this gap significantly.

1 Introduction

NLG deals with creating human-readable text from input text or even images. It has wide applications in translating text from one language to another, summarizing input text in a concise way, question answering as used in chatbots, image captioning, and story generation. The main challenge lies in the evaluation of text generated by the model. Human evaluation is considered to be the gold standard for all NLG applications (Celikyilmaz et al., 2020). So, one way is to assign a group of linguists and language experts and ask them to evaluate the machine-generated output. But applications generate trillions of lines of text which makes this method time-consuming and expensive. In the past few years, advancements have been made in NLP, which resulted in improvements in NLG applications. So, automatic evaluation of model-generated text is required.

Human evaluation technique cannot be applied to day-to-day NLG applications. NLG researchers use automatic evaluation metrics. These evaluation metrics give similarity score between *reference text* (human-annotated text) and *candidate text* (machine-generated text). Two of the earliest metric were BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). These were based on the heuristic method of n-gram overlap. ROUGE has different variants and calculates recall instead of precision. Since both these metrics are more focused on capturing word matching, these do not capture syntactic and semantic similarity between *candidate* and *reference text*.

Recently semantic similarity-based metrics have been used, which correlate better with human judgements of similarity between machine output and ground truth. These approaches use embeddings or vec- tors of words/ tokens/ subtokens of the sentences (both *candidate* and *reference text*). These embeddings are created using machine learning models.Similarity score like cosine similarity is used to get the similarity score. These embeddings can be static, i.e., vector representation for the word or token remains the same for every text or can be dynamic like in BERTScore (a BERT-based similarity score metric), which uses contextual embeddings where the vector representation of the token can change according to the context where it is used. These machine-learned metric approaches capture semantic, syntactic, and grammatical features between *candidate* and *reference text*. These are more correlated with human judgements of similarity. BLEURT a BERT-based evaluation metric gave better absolute Kendall Tau score on BLUERT than BLEU, METEOR and BERTScore with human judgements on WebNLG dataset (Sellam et al.,

2020).

In our approach, we will try to extract semantics, syntactic and grammatical features score from the *candidate* and *reference text* in the form of vectors and then will pass it through composite metric along with embedding of candidate and reference text. Additionally, we employed Principal Component Analysis (PCA) to minimise the data dimensionality in order to lessen the computational cost of the model. For rank classification, the processed input is subsequently delivered to a deep neural network along with the corresponding rank label. Our experimental results demonstrate that including assessment criteria greatly enhances model performance, leading to higher accuracy in rank classification tasks.

2 Problem Definition

Design an automatic machine learned evaluation metric which perform mainstream task of evaluating quality of machine Translation system from English to Hindi. In this paper we explore various methods to get our evaluating metric to be as similar as possible to human annotated score which continue to be gold standard for such task. This problem becomes more relevant as indigenous languages' text generation has not been evaluated as extensively as compared to text generation of other popular languages such as English, Spanish, French, German, Dutch,Portuguese and other European languages.

3 Related Work

Celikyilmaz et al. (2020) have divided NLG evaluation into three categories in their survey paper, namely: (1) Human-Centric, (2) Untrained and (3) Machine-Learned evaluation metrics.

3.1 Human-Centric Evaluation Methods

Since natural language is complex and constantly evolving, judging it with a constrained set of parameters is hard. Therefore, human evaluation becomes the most critical part of evaluating any NLG system and is often considered a gold standard for judging untrained and machine-learned evaluation metrics.

However, the human evaluation also faces some challenges. It can be very time-consuming and expensive as various tasks may need a certain level of domain expertise. Furthermore, particular evaluation dimensions such as diversity may not be suitable for a large group of human annotators (Hashimoto et al., 2019).

Belz and Reiter (2006) divides human evaluation in two categories, *intrinsic evaluation* and *extrinsic evaluation*. The intrinsic evaluation focuses on certain dimensions of text such as *adequacy, fluency, diversity, actuality, coherence and consistency*. Extrinsic evaluation measures the performance of the NLG system in the downstream task (Hastie and Belz, 2014).

It is also important to have a high inter-evaluator agreement for a reliable human evaluation. The higher inter-evaluator agreement indicates a well-defined evaluation task, and differences in the text are noticeable. Artstein and Poesio (2008) have given various agreement measures such as *Percent agreement, Cohen's* κ*, Fleiss'* κ*, Krippendorff's* α. However, it is also worth noting that aiming for a higher inter-evaluator agreement may not be desirable (Amidei et al., 2019).

3.2 Untrained Automatic Evaluation Metrics

3.2.1 n-gram overlap based

The n-gram overlap-based metrics are the most commonly used metrics in NLG systems. **BLEU** is the first metric to evaluate the similarity between human reference text and candidate text. BLEU score was the product of *brevity penalty* and *geometric mean* of precision of n-grams (1gram, 2gram, ...). However, it cannot evaluate the syntactic and semantic properties of text. It also struggles with morphologically rich languages. Caccia et al. (2018) mentions that texts having higher BLEU scores were mostly grammatically correct but often lacked global coherence.

Many other n-gram overlap-based metrics are iterative improvements over BLEU for addressing its specific drawbacks. Like **ROUGE** is recall based measure, which makes the result more interpretable (Callison-Burch et al., 2006). **METEOR** uses stemming and synonym matching to improve upon semantic similarity evaluation. **CIDEr** uses tf-idf weights to give more weights to rarely used words in reference text. **RIBES** is not reliant on word boundaries and also incorporates word order for evaluation.

There are also metrics like **MASI**, which uses the Jaccard coefficient and **DICE**, which uses the Dice coefficient.

3.2.2 Distance based

Edit Distance-Based Metric:
Word error rate (WER): It is the fraction of words that must be added, removed, or substituted in the candidate sentence to acquire the reference text, such that a distance metric between hypothesis and reference sentences is produced
Translation edit rate (TER): TER is the minimal count of edits required to modify a machine produced sentence such that it matches one of the reference sentences precisely, standardised by the mean length of the reference text.
Vector Based Evaluation Metrics:
Sentence Mover's Distance (SMD): Each document is represented by an SMD, which is a collection of sentences, words, etc., with sentence embeddings weighted according to their length. SMD quantifies the cumulative distance required to align the sentence embeddings of one text with those of another.

3.2.3 Content overlap

Spice: In Spice, to assess the quality of captions, we convert both candidate captions (C) and reference captions (R) into a scene graph. For example, take the G(R1), G(R1),... G(Rn) graphs for reference sentences and the union of all these graphs, which is a graph semantic representation of an image scene. This model encodes directly for objects, properties, and relationships (Anderson et al., 2016).
Creating a scene graph from an image description is a two-step procedure. First, syntactic links between words in the caption are identified using a dependency parser that is trained on a large dataset. A rule-based framework then converts these relationships to scene graphs. The F-score is determined by comparing the candidate graph to the reference scene graph and assessing the combination of tuples reflecting the semantic propositions included inside the graph.

3.3 Machine-Learned Evaluation Metrics

3.3.1 Regression-Based Evaluation:

The model receives a pair of sentences—the reference sentence and the candidate sentence and generates a score that indicates the quality of the candidate sentence relative to the reference sentence. (Shimanaka et al., 2018) The model understands semantic meaning by using sentence embeddings that have already been trained. The embeddings are made with a language model like BERT or GPT that has been trained on a huge amount of text data. It learns the relationship between the reference and candidate sentence embeddings by using a multilayer perceptron (MLP) with several hidden layers.

4 Corpus/Data Description

The English-Hindi Human Judgement Corpus from the WMT14 shared task is a dataset containing human judgments of the quality of machine-generated translations from English to Hindi. The WMT14 shared task was a machine translation evaluation campaign that focused on translation between European languages and Hindi (Bojar et al., 2014).

The purpose of the corpus is to provide a high-quality dataset for evaluating English-to-Hindi machine translation systems. The human annotations provide a reliable measure of the quality of the translations, which can be used to compare various approaches to machine translation and to identify enhancement opportunities.

The WMT14 Hindi Human Judgement Corpus comprises a set of human judgements that rank candidate sentences relative to reference translations, as opposed to designating numerical scores. Consequently, it is a ranking dataset. It comprises of 1698 sets of 5 candidate sentences with their reference sentence.

The corpus is downloadable from the WMT website and is available for research purposes. It has been utilised in a number of machine translation research studies and remains a valuable resource for scholars in this field.

Apart from this we also generate another corpus with human annotations which gives scores based on factors such as semantics, syntactic, grammatical and overall quality of translation. For this, we use IIT Bombay English to Hindi dataset which has candidate reference pairs. We then annotate these sentences ourselves. For generalised annotations, we average 5 distinct human annotations. For annotation purposes, we create a Label Studio dashboard which allows us to simplify the process.

5 Proposed Approach

Since we have two different types of datasets, this allows us to create two different metrics, one which allows us to rank the translated sentences against other candidates' sentences, the other allows us to give a rating to the quality of translation.

Predicting the rank of a sentence based on reference and candidate sentences has caught the attention of researchers in the field of Natural Language Processing (NLP). We use the WMT dataset for this purpose, which provides us with 5 different candidate sentences for a single reference sentence, and their rankings. Tokenization, stemming, and concatenation of reference and candidate sentences were all part of the input data preprocessing.

We added additional characteristics to the input data to accommodate for various evaluation criteria such as Bert, Blue, and Meteor. In addition, to lower the computational complexity of the model, we used Principal Component Analysis (PCA) to reduce the data dimensionality. The processed input is then sent into a deep neural network with the associated rank label for rank classification. Our experimental findings show that integrating evaluation metrics improves model performance significantly, resulting in greater accuracy in rank classification tasks.

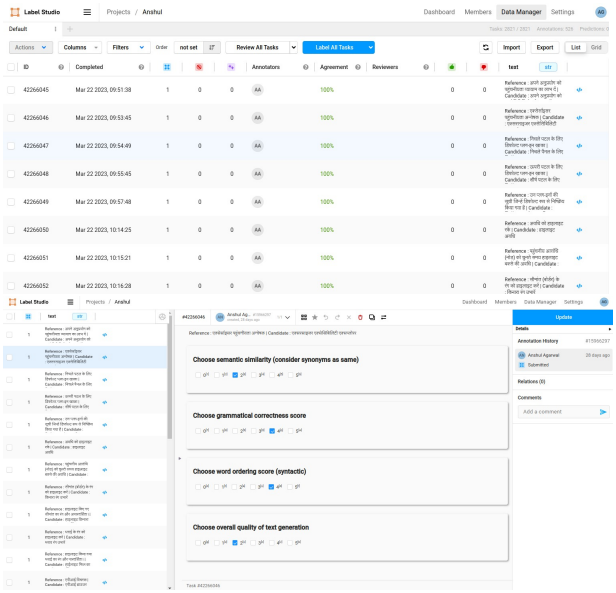
We also experiment this same model setting for different types of neural networks, including Multi-Layer Perceptron, Vanilla RNN, CNN and LSTM. Data processing and feature engineering is done similarly to the above model. The output for this model is the average of the 4 scores given, for the semantic, syntactic, grammatical and overall correctness.

We apply a similar model setting to create the score based metric. For this we use our own self-annotated dataset, where annotations have been made on the grammatic, syntactic, semantic and overall quality of transaction. The reference and candidate sentences for this have taken from the IITB-hindi-english corpus.

6 Experiments and Results

Key points (Following can be in any order):

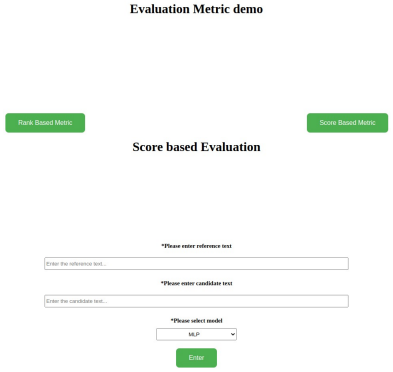
1. The first part of the project was to create the human annotations for the iitb-hindi-english dataset. After proper cleaning of the dataset, we go on to create a label-studio dashboard, which allows us all to give scores to the quality of translation for different sentences.



We use this generated dataset to create the score-based metric.

2. Next, we obtain the WMT dataset, and use it for creating the rank based metric.
3. **Rank based model**: The first model that we experiment with is the multilayer perceptron. Here we use the Bert embeddings of the candidate and target sentences, and concatenate them. Then, we perform Priciple Component Analysis(PCA) on these embeddings. After this, we also include existing metrics such are Bleu, Meteor and Ter, and include them as features. The output here is 5x1 one hot vector, where the 1 denotes the rank assigned to that sentence. This gives a very good accuracy on the train dataset.
4. We then experiment with different neural networks Vanilla RNN, LSTM and CNN. All these models give accuracy similar to the case of the Multilayer perceptron.
5. **Score based model**: For this, the only thing we change is the outputs, which are the annotations that we had generated. This is a 4 dimensional output, where each dimension denotes the score given based on syntactic, semantic, grammatical and overall quality of translation.
6. Here, again we try the different neural networks, which give similar results. One think that can be observed is that the accuracy we get here is lesser than that of the Rank based metric.
7. Once, the various models have been created, we create an interface for presentation purposes. This interface allows us to chose which

kind of metric we want to generate (rank-based or score-based), and then allows us to chose which model we want to use to generate our metric from the various models we tried

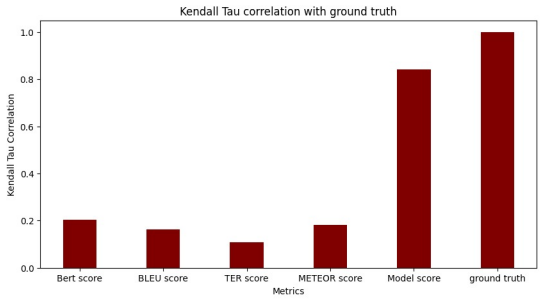


8. GitHub contains all the code files for the models, data files, and codes

7 Error Analysis

Model	Train Accuracy	Test Accuracy
HindVal	0.84	0.67

Table 1: Rank Based metrics Result



8 Future Directions

One of the biggest challenge that we faced in this project was the lack of human annotations available for hindi-english translations. The WMT-14 dataset had 8000 sentence pairs, but in the next editions, hindi-english pair was ommited. We also tried generating our own annotations using the IITB corpus, but it was a tedious task, and we ended up with around only 800 sentence pairs. Future research can focus on generating human annotations, and increase the volume of datasets available, so that we can properly judge the quality using evaluation metrics.

Also, some transformer based models are also being explored to create and generate metrics. A recent paper called BARTScore had also used tried to use Seq2Seq for this task. An approach that is being propsed is to incorportate the method used to generate the evaluation metric using the same models used for text generation.

The quality of the produced output in many NLP tasks may rely not only on the text but also on other modalities like photos, videos, and audio. Future studies might concentrate on creating assessment measures that can measure the output quality of multimodal systems while accounting for text and other modalities. The relevance, coherence, and informativeness of the summary should be taken into consideration while assessing its quality. Future study might concentrate on creating assessment measures that are task-specific and capture the essential components of phrase quality for each activity.

9 Individual Contribution

Name	Contribution
Abhas Goyal	Dataset, Models, Ppt, Report
Anshul Agarwal	L.R.,Models, Ppt, Report, UI
Harshit Itondia	L.R.,Models, Ppt, Report
Vedant Joshi	L.R.,Models, Ppt, Report
Vijay Kumar	L.R.,Models, Ppt, Report

10 Conclusion

This study proposes a composite evaluation metric for English to Hindi machine translation that incorporates various deep learning models to capture semantic, syntactic, and pragmatic aspects of generated text in order to address the limitations of existing evaluation metrics for Natural Language Generation tasks. The WMT14 Hindi Human Judgement Corpus is used in the first measure to predict the rank of translated sentences based on reference and candidate sentences. With the inclusion of new assessment measures, we have employed deep neural networks to considerably boost model performance. Additionally, we have experimented with other neural network architectures, such as the Multi-Layer Perceptron, Vanilla RNN, CNN, and LSTM. The second measure entails utilising our own self-annotated dataset created from the IITB-hindi-english corpus to provide a score based on the grammatical, syntactic, semantic, and overall quality of translation. For researchers in the area of machine translation, both metrics provide useful resources for future research, especially for indigenous languages such as Hindi, which have received

less attention than European languages. Overall, the work intends to increase machine translation quality and close a research gap in text creation research for regional Indian languages.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354, Tokyo, Japan. Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2018. Language gans falling short.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Helen Hastie and Anja Belz. 2014. A comparative evaluation methodology for NLG in interactive systems. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4004–4011, Reykjavik, Iceland. European Language Resources Association (ELRA).

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.