INDIAN INSTITUTE OF TECHNOLOGY JODHPUR

# Grammatical Error Correction for Indian Language

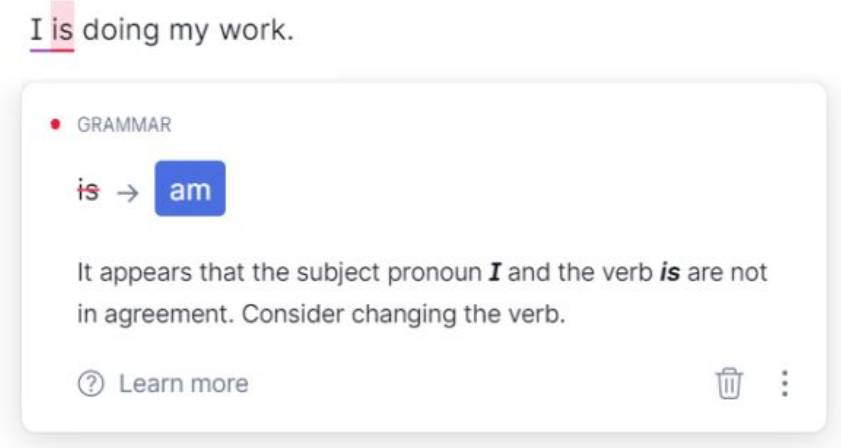
Presented by: Puja Gupta (M21MA004)
Supervisors: Dr. Gaurav Harit and Dr. Kirankumar Hiremath

# Project Outline

- Grammatical Error Correction (GEC)
- Motivation
- Problem Formulation
- Experiments on English GEC
- Proposed method for Hindi GEC
- Dataset Used
- Results
- Prediction
- Conclusion
- Future Wok

# What is GEC task in NLP?

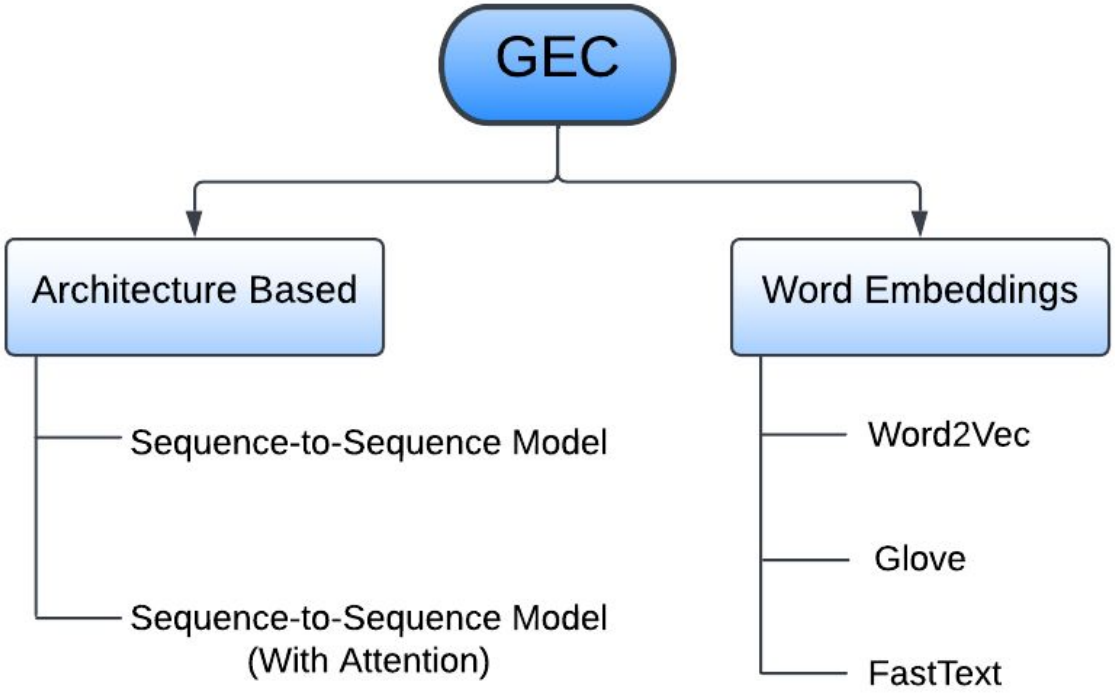
- Grammatical Error Correction (GEC) is the task of error detection and correction in the text.
- In NLP, GEC can be formulated as a sequence-to-sequence task, where a model is trained using grammatically incorrect sentences as input and return a grammatically correct sentence.



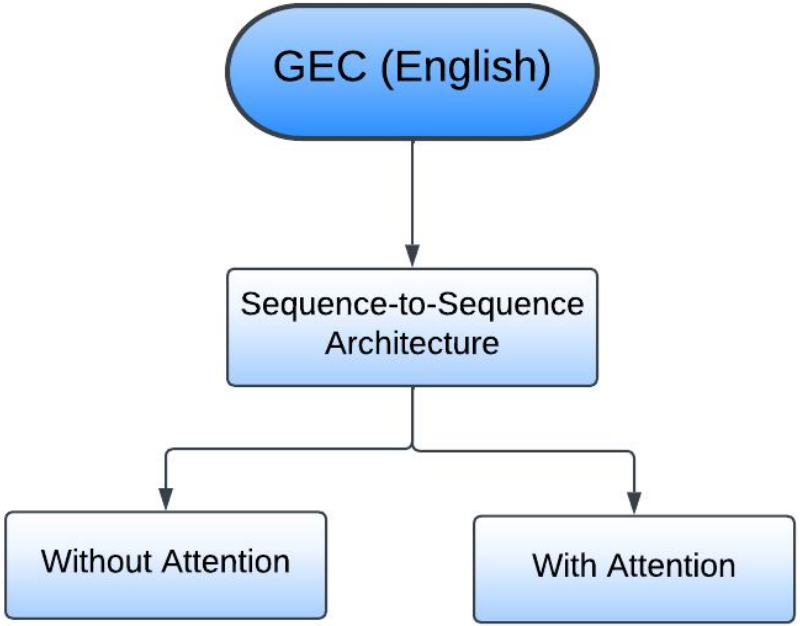
# Motivation

- A lot of research and trained models are available for English Language.
- Minimal amount of work on GEC for  Indian Languages.
- No state-of-the-art trained model available for the GEC task in Hindi.
- Need a well trained model for implementing GEC tasks for Indian languages which can help the writers.
- Proposed the work towards GEC for Hindi Language.

# Problem Formulation



# Experiments on English Language




LANG8 dataset
(80624 training, 20156 validation, 2000 test sentences)

# Sequence-to-Sequence Model

- Consists of two sub models: Encoder and Decoder.
- Both i/p and o/p seq can be of same or different lengths.
- The submodels can be implemented using any architecture : RNN, LSTM, Bidirectional LSTM, etc.
- Final Hidden state of Encoder can be viewed as “Context Embedding vector” of the entire sentence and is fed to decoder.

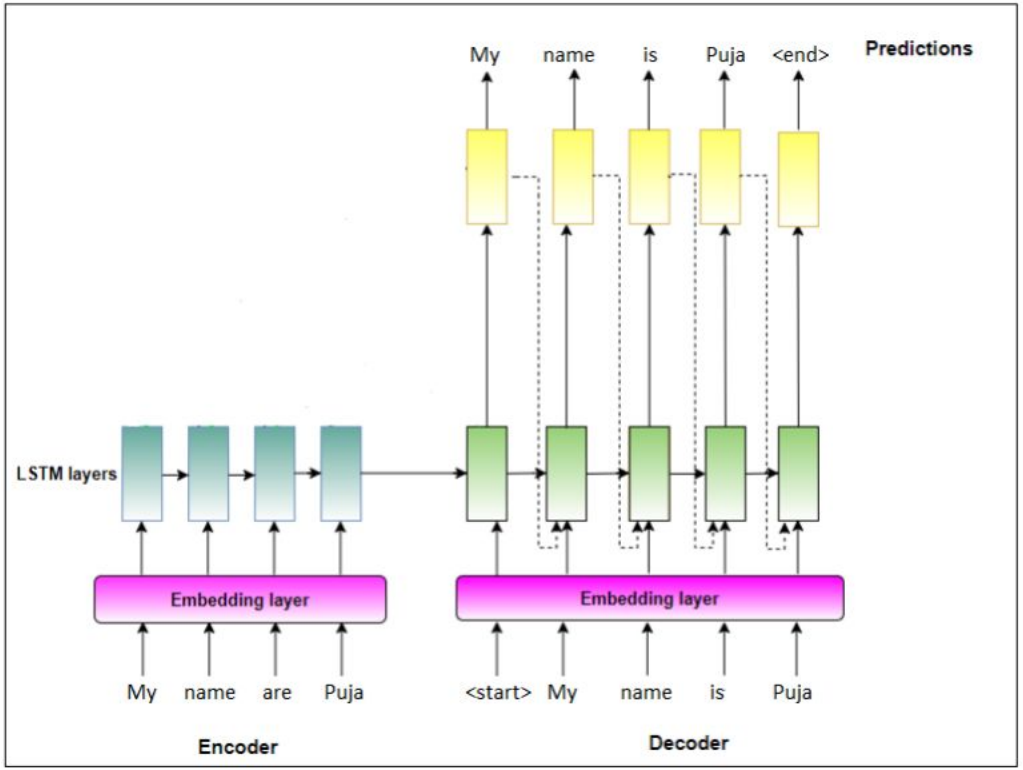
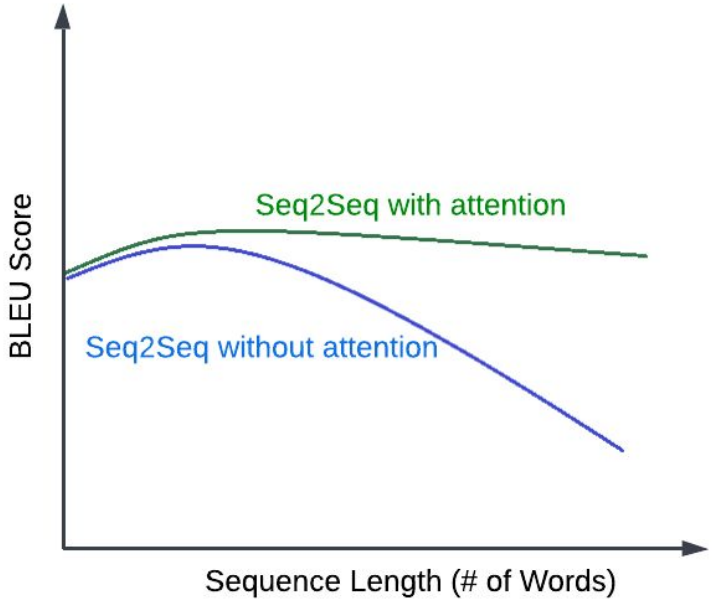



Figure 2.2: Seq2Seq Encoder Decoder model without Attention

# Shortcomings of Sequence-to-Sequence Model

- Compress the information about the i/p sentence into a fixed length vector “Embedding”.
- The longer the input sentence the greater the information loss leading to information bottleneck.



# Sequence-to-Sequence (With Attention)

- Let the decoder look back to the source & learn which part of the i/p to focus at each time stamp.
- Self attention ($c_t$) to understand the context.
- At each decoding timestep, the decoder accesses the encoder's hidden state.
- Attention scores are calculated, how much "attention" should each encoder hidden state be given for this decoder timestep.
- Concat context vector, self attention vector ($c_t$), attention vector ($a_t$) with i/p embedding and give as input to decoder.

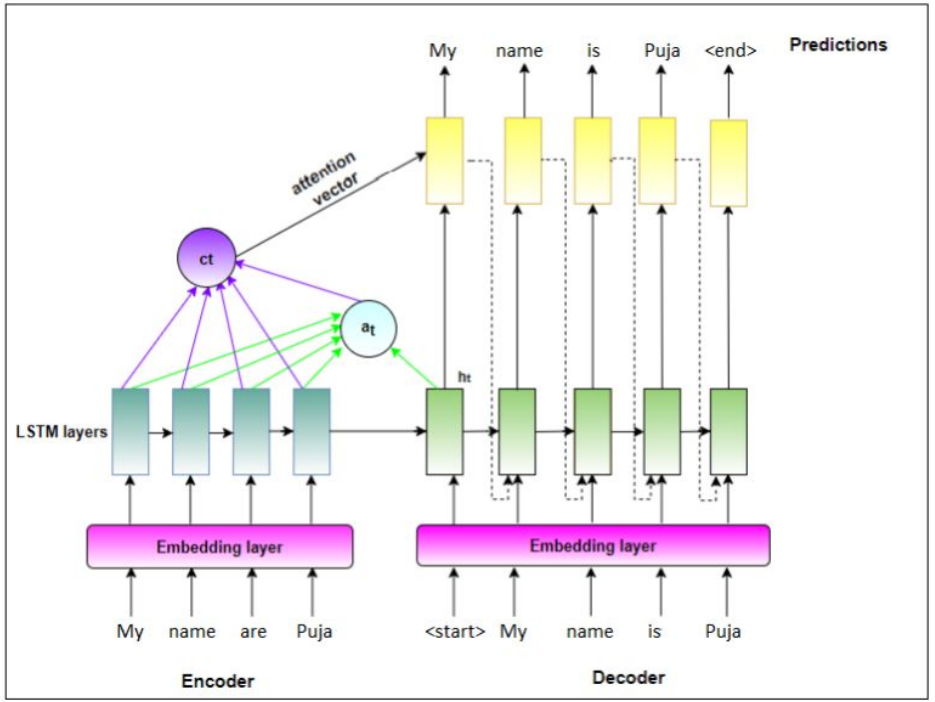
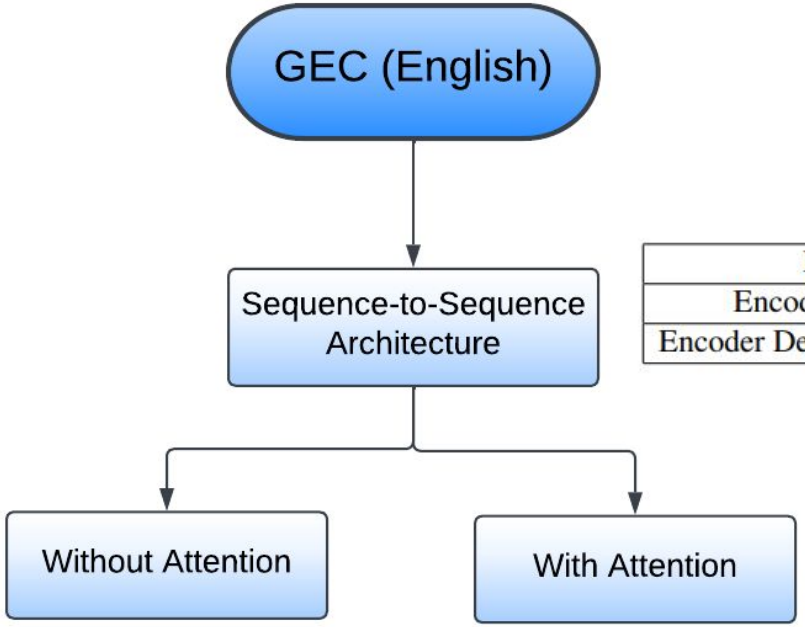



Figure 2.3: Seq2Seq Encoder Decoder model with Attention mechanism for GEC in English

# Benefits of Attention

- Attention significantly improves Encoder Decoder based Seq2Seq model.
- With Attention, Seq2Seq model does not forget the source input.
- Attention mechanism keeps the context of input sentences to encoder with importance of a particular word.
- Downside: much more computation.

# Results on English Language




LANG8 dataset
(80624 training, 20156 validation, 2000 test sentences)

| Model | Embedding | Training Loss | Validation Loss | BLEU Score |
|---|---|---|---|---|
| Encoder Decoder | Trainable | 0.5163 | 1.1652 | 0.1294 |
| Encoder Decoder (Attention) | Trainable | 0.2610 | 0.6519 | 0.4426 |

Table 5.1: Test results obtained for different models for English

# Proposed Method for Hindi GEC

Seq2Seq model with attention produced a decent result for English. These results motivated us to implement the same method for GEC in Hindi as well.

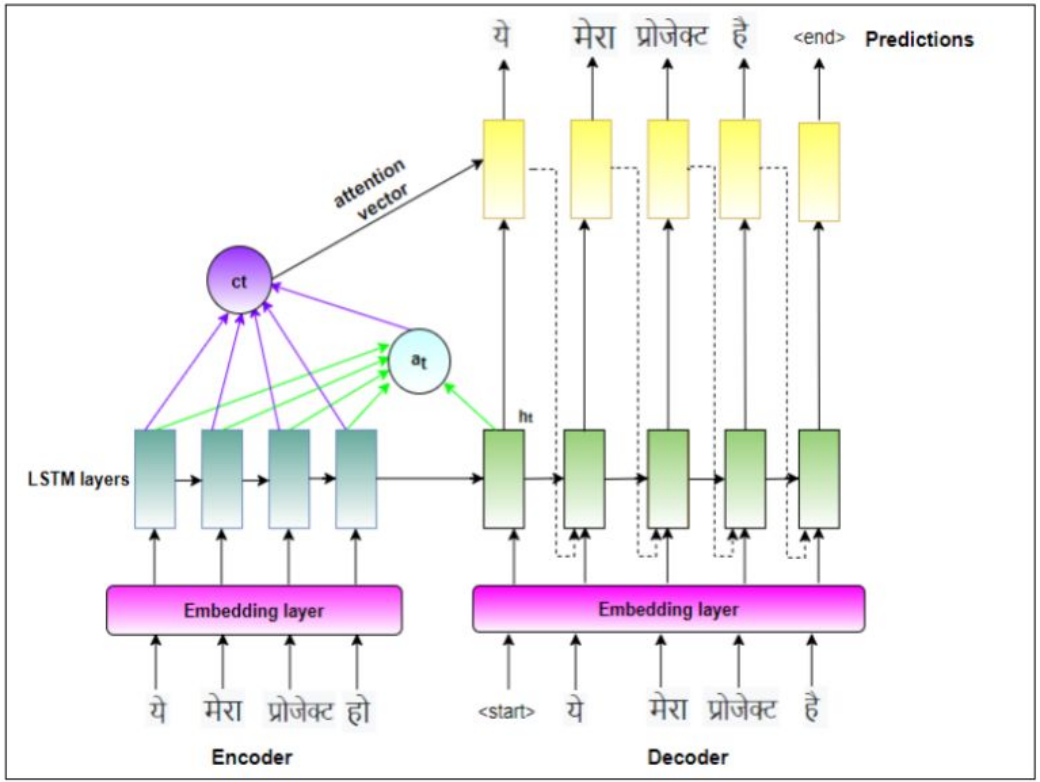
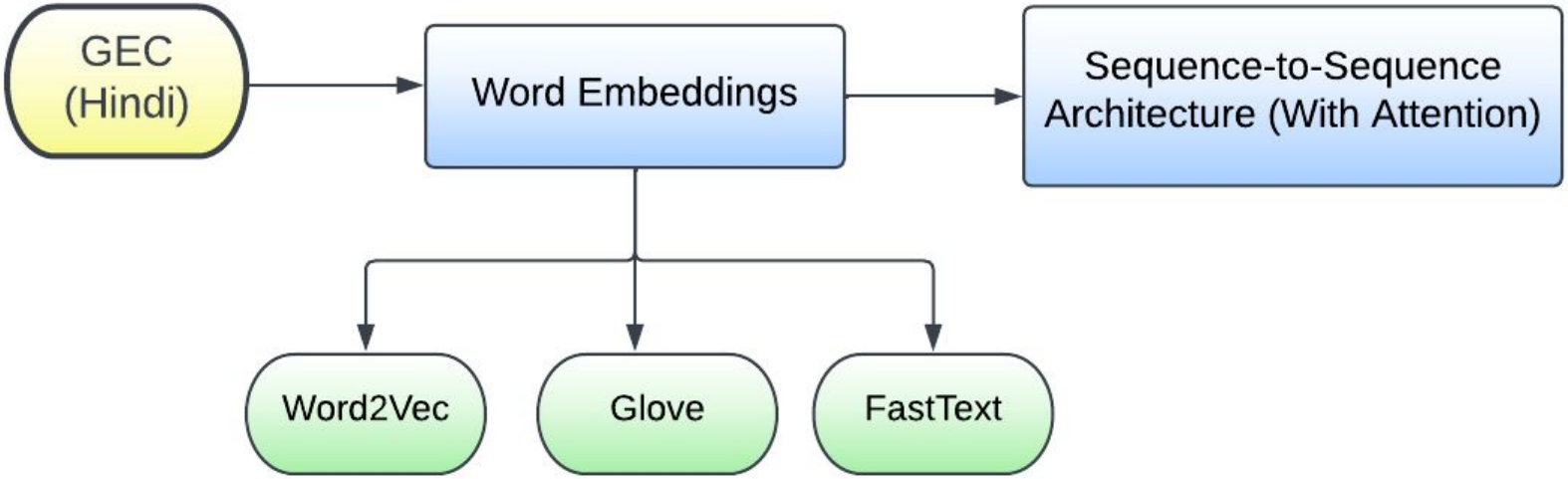



Figure 4.1: Seq2Seq Encoder Decoder model with Attention for GEC in Hindi

# DataSet Used : Etoori's Dataset

| | enc_input | dec_input |
|---|---|---|
| 0 | परन्तु वे दोनों उन बातों को ज़्यादा समय तक अप ... | परन्तु वे दोनों उन बातों को ज़्यादा समय तक अपन... |
| 1 | देश में हिन्दी को विस्थापित कर का षड़यंत्र चल ... | देश में हिन्दी को विस्थापित करने का षड़यंत्र च... |
| 2 | तीन साल पहले कातिलाना हमले के प्रकरण में एफआर ... | तीन साल पहले कातिलाना हमले के प्रकरण में एफआर ... |
| 3 | रामायण रिविजिटेड अ टेल ऑफ लव एंड एडवेंचर नाम स... | रामायण रिविजिटेड अ टेल ऑफ लव एंड एडवेंचर नाम स... |
| 4 | तब तक के लिए हमें विराम ले की अनुमति दीजिए । | तब तक के लिए हमें विराम लेने की अनुमति दीजिए । |

# Proposed method for Hindi




Etoori dataset
(112K training, 28K validation, 10K test sentences)

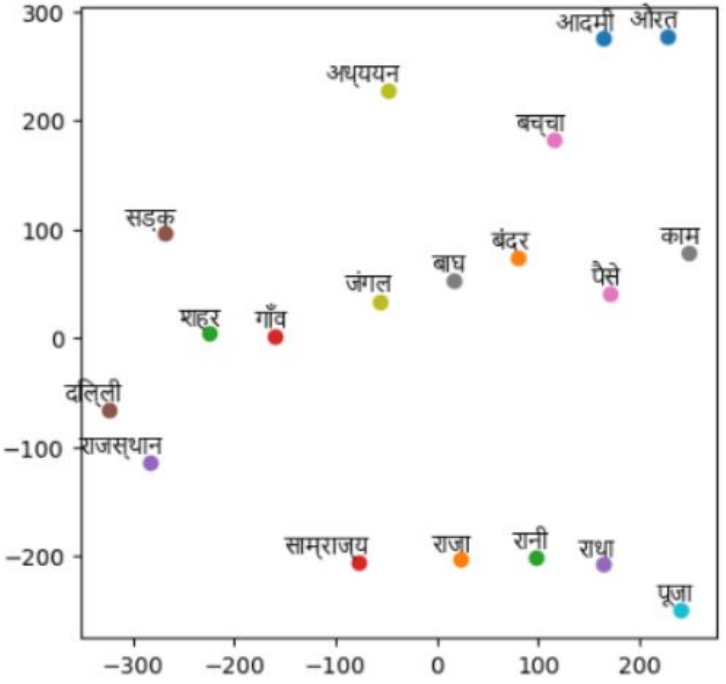
# Word Embeddings

- Machines does not understand the textual data directly.
- Need to convert the textual data into some word vector representation and feed that to the model for processing.
- Word Embeddings enable to store contextual information in a low dimensional vector.
- Words that occur in similar contexts tend to have similar embeddings.
- The vectors are learnt by models, through unsupervised learning.

# Word2Vec

- Word2Vec is a popular method to generate word embeddings.
- The Word2Vec objective function causes the words that occur in similar context to have similar embeddings.
- Words with the same meaning comes closer.

```
Nearest Neighbors to पूजा using Word2Vec are:
[('आराधना', 0.7864888310432434),
 ('पुजा', 0.7753461599349976),
 ('पूजन', 0.7515121102333069),
 ('अराधना', 0.7468945384025574),
 ('उपासना', 0.7327789664268494),
 ('पुजन', 0.6703079342842102),
 ('आरती', 0.6687861680984497),
 ('प्राणप्रतिष्ठा', 0.6564066410064697),
 ('पूजापाठ', 0.6346961259841919),
 ('अर्चना', 0.6229457259178162)]
```



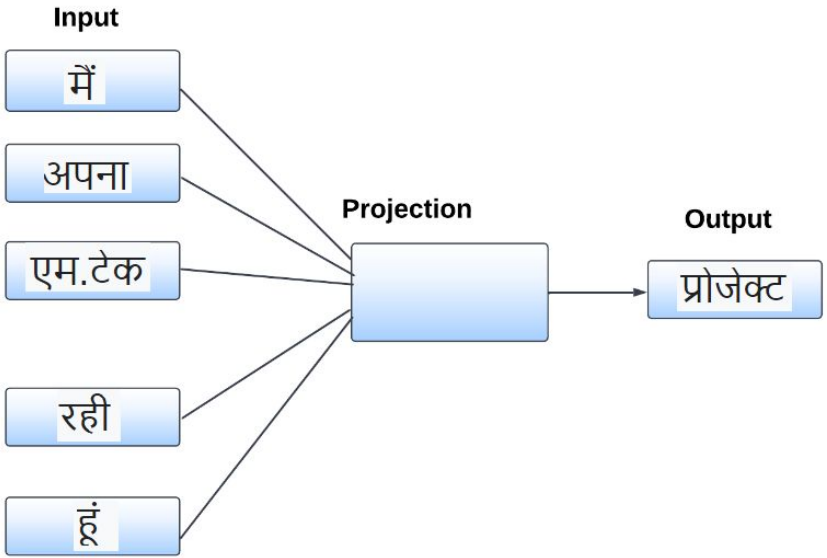



Fig: CBOW : Predicts central word based on window of words surrounding it.

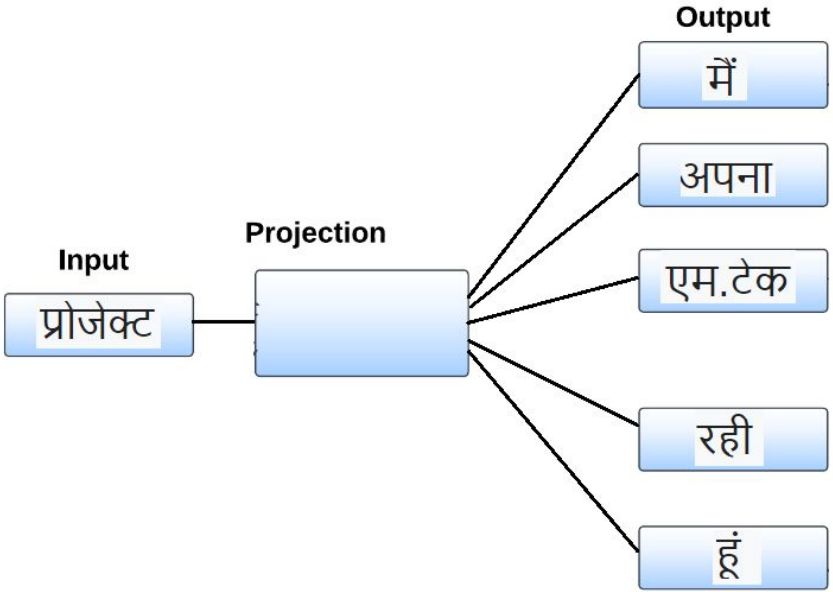
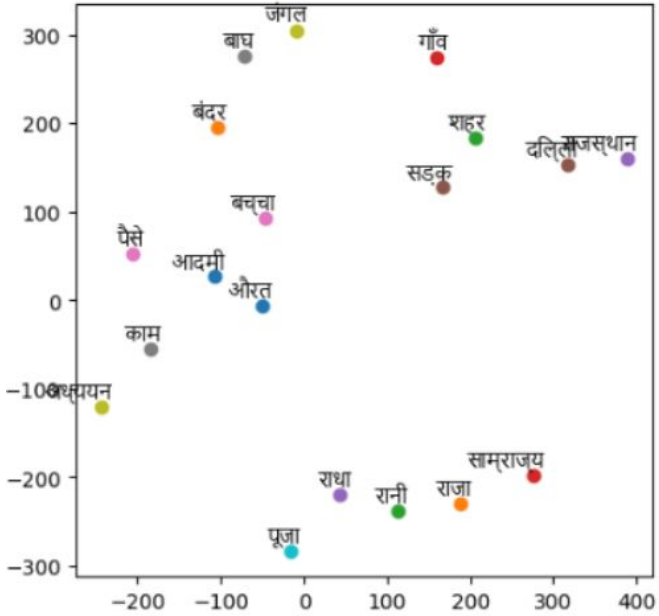



Fig: Skip Gram : Predicts context based on central word.

Pre Trained Word2Vec Model (CBOW) on Hindi is used for generating word embeddings of dimension 300.

# Glove

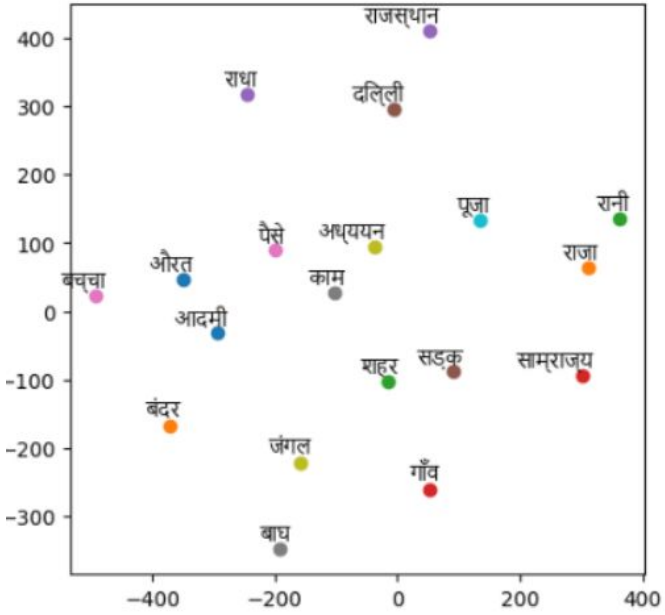
- It is also another popular method to find the word embedding.
- It captures global information as well, in the form of word co occurrence matrix.
- Relationships between word pairs are also taken into account rather than capturing just word and word relationships.
- Less weight is given to the most frequent word pairs to prevent meaningless stop words, from dominating the training progress.
- Huge Computation is required.



# FastText

- FastText was created by the Facebook Research Team and it is helpful for word representations.
- It is different from gensim word2vec model as it takes in a word representation using the character n-gram technique.
- Unlike Word2Vec and Glove, it can give word vectors for OOV (Out of Vocabulary) and rare words because a single word is further broken into character n-grams.

```
Nearest Neighbors to पूजा using Fasttext are:
[(0.7876037955284119, 'पूजां'),
 (0.7792069911956787, 'पूजा।'),
 (0.7434489130973816, 'पूजारी'),
 (0.7403855919837952, 'देवपूजा'),
 (0.736215353012085, 'पूजापाठ'),
 (0.7296294569969177, 'पूजाओं'),
 (0.7163258790969849, 'अर्चना'),
 (0.7134318947792053, 'पूजन'),
 (0.7024132609367371, 'पूजक'),
 (0.6988179683685303, 'पूजती')]
```



# Hyperparameters used

| | |
|---|---|
| Loss used | Sparse Categorical Cross Entropy Loss |
| Epochs | 51 |
| Encoder Vocabulary Size | 73829 |
| Decoder Vocabulary Size | 72971 |
| Embedding Dimension | 300 |
| Maximum Sequence Length | 35 |
| Evaluation metrics | Accuracy, Bleu score |
| Trained on devices | 172.25.0.209 Server |

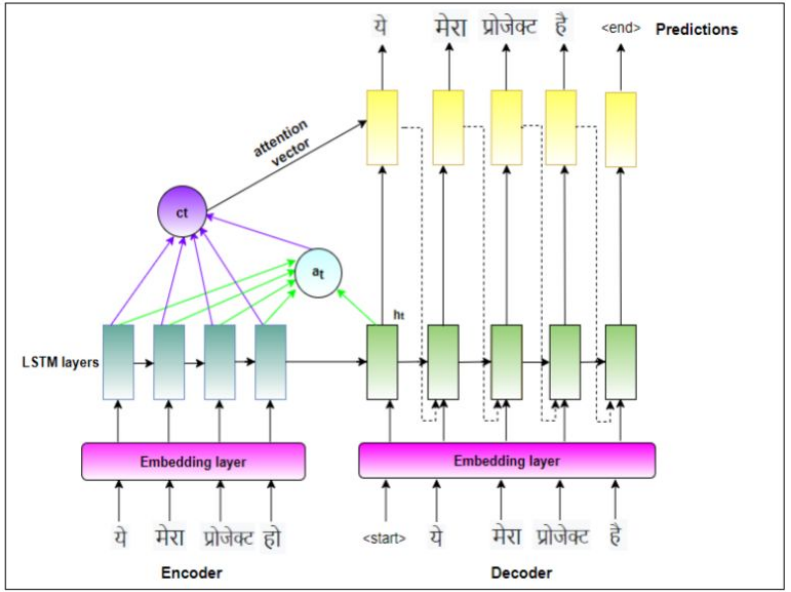



Figure 4.1: Seq2Seq Encoder Decoder model with Attention for GEC in Hindi

The experiment and analysis code repository is available on the 172.25.0.209 server under the user id gupta92.

# Result on Proposed Method (Hindi GEC)

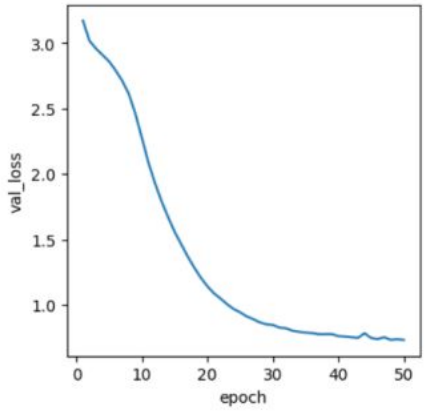
| Model (Seq2Seq) | Validation Accuracy (%) | BLEU Score | Model Size (Bytes) | Inference Time(ms) |
|---|---|---|---|---|
| Encoder Decoder | 69.55 | 0.30497 | 255761972 | 177 |
| Encoder Decoder (Attention) | 99.02 | 0.76631 | 331019284 | 186 |

Table 2: Encoder Decoder model with attention outperformed Encoder Decoder model without attention

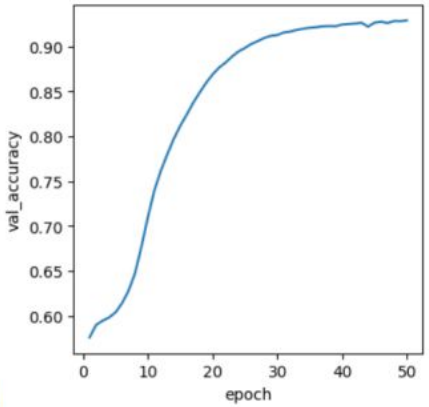
# Result on Proposed Method (Hindi GEC)

| Model (Seq2Seq) | Embedding | Training Accuracy (%) | Validation Accuracy (%) | BLEU Score (Greedy Search) |
|---|---|---|---|---|
| Encoder Decoder (Attention) | Word2Vec | 98.08 | 95.44 | 0.7796 |
| Encoder Decoder (Attention) | Glove | 98.86 | 96.15 | 0.82002 |
| Encoder Decoder (Attention) | FastText | 99.30 | 96.55 | 0.8458 |

Table 3: Results obtained for different word embeddings

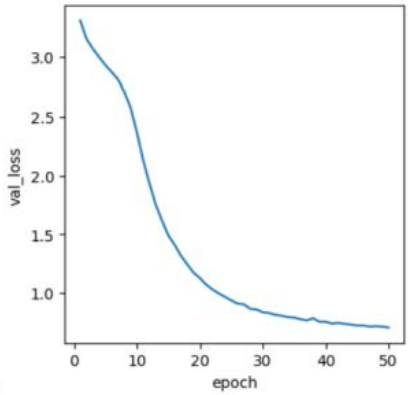



[a]

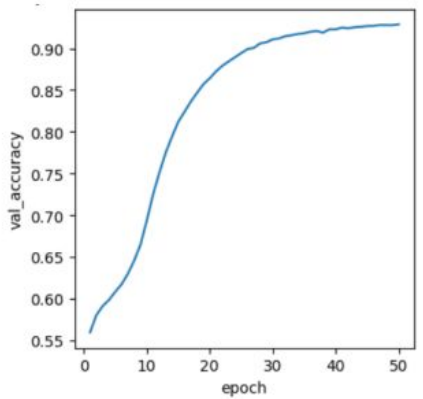



[b]

Figure 5.5: [a] Validation Loss Vs Epoch and [b] Validation Accuracy Vs Epoch for Encoder Decoder Architecture with Attention using Word2Vec Word Embedding

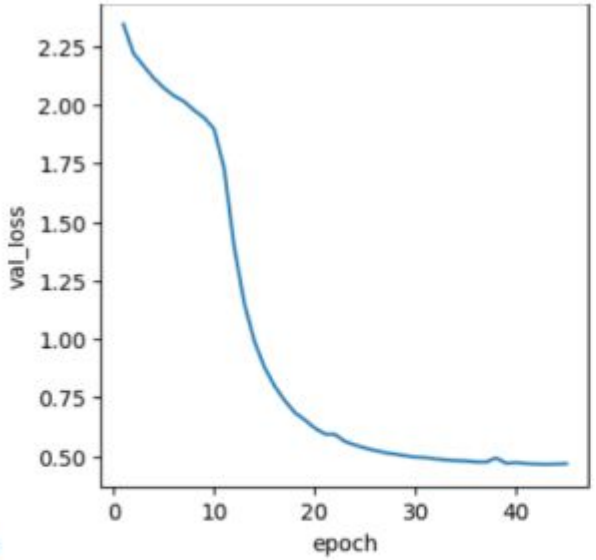



[a]

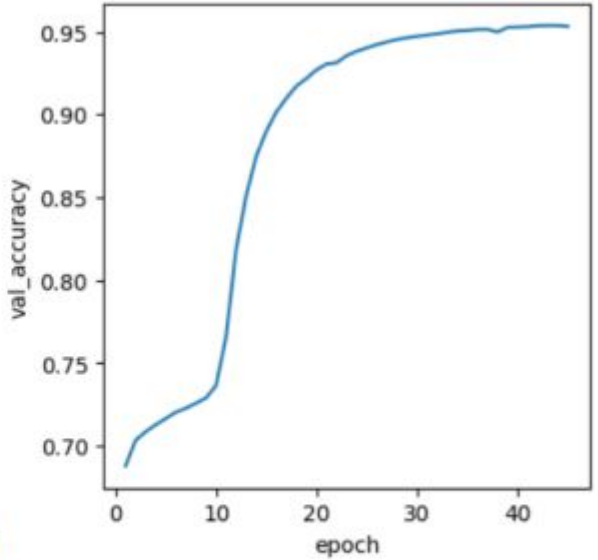



[b]

Figure 5.6: [a] Validation Loss Vs Epoch and [b] Validation Accuracy Vs Epoch for Encoder Decoder Architecture with Attention using Glove Word Embedding




[a]




[b]

Figure 5.7: [a] Validation Loss Vs Epoch and [b] Validation Accuracy Vs Epoch for Encoder Decoder Architecture with Attention using Fasttext Word Embedding

# Result on Proposed Method (Hindi GEC)

| Model (Seq2Seq) | Embedding | Model Size (Bytes) | Inference Time (ms) |
|---|---|---|---|
| Encoder Decoder (Attention) | Word2Vec | 48870508 | 273 |
| Encoder Decoder (Attention) | Glove | 48870508 | 253 |
| Encoder Decoder (Attention) | FastText | 48870508 | 345 |

Table 2: Results obtained for different word embeddings

# GEC on Proposed Method (Hindi GEC)

| | |
|---|---|
| OUTPUT 1 | INPUT SENTENCE ===> घर में घुस के बाद भी सुकून नहीं ।<br>PREDICTED SENTENCE ===> घर में घुसने के बाद भी सुकून नहीं । <end><br>ACTUAL SENTENCE ===> घर में घुसने के बाद भी सुकून नहीं । <end> |
| OUTPUT 2 | INPUT SENTENCE ===> कइयों के साम कठिनाइयाँ होती हैं ।<br>PREDICTED SENTENCE ===> कइयों के सामने कठिनाइयाँ होती हैं । <end><br>ACTUAL SENTENCE ===> कइयों के सामने कठिनाइयाँ होती हैं । <end> |
| OUTPUT 3 | INPUT SENTENCE ===> वह डब्बा मैंने संभाल कर नहीं रख पाया ।<br>PREDICTED SENTENCE ===> वह डब्बा मैं संभाल कर नहीं रख पाया । <end><br>ACTUAL SENTENCE ===> वह डब्बा मैं संभाल कर नहीं रख पाया । <end> |
| OUTPUT 4 | INPUT SENTENCE ===> इस हादसे के शिकार हुओं में अधिकांश बीस से तीस साल के युवा थी ।<br>PREDICTED SENTENCE ===> इस हादसे के शिकार हुओं में अधिकांश बीस से तीस साल के युवा थे । <end><br>ACTUAL SENTENCE ===> इस हादसे के शिकार हुओं में अधिकांश बीस से तीस साल के युवा थे । <end> |

# GEC on Proposed Method (Hindi GEC)

| | |
|---|---|
| OUTPUT 5 | INPUT SENTENCE ===> सीआरपीएफ अप ट्रेनिंग में लड़कियों को आत्मरक्षा की पूरी जानकारी दी है ।<br>PREDICTED SENTENCE ===> सीआरपीएफ ने अपने ट्रेनिंग में लड़कियों को आत्मरक्षा की पूरी जानकारी दी है । <end><br>ACTUAL SENTENCE ===> सीआरपीएफ ने अपने ट्रेनिंग में लड़कियों को आत्मरक्षा की पूरी जानकारी दी है । <end> |
| OUTPUT 6 | INPUT SENTENCE ===> हर इक सोच में अन्तर होता हैं<br>PREDICTED SENTENCE ===> हर इक सोच में अन्तर होता है <end><br>ACTUAL SENTENCE ===> हर इक सोच में अन्तर होता है <end> |
| OUTPUT 7 | INPUT SENTENCE ===> लेकिन अब समय पलटी मारी है और फिर सचिन के शून्य में आउट हो के रिकार्ड की तरह शेयर बाजार में भी गिरावट का रिकार्ड जारी है ।<br>PREDICTED SENTENCE ===> लेकिन अब समय ने पलटी मारी है और फिर सचिन के शून्य में आउट होने के रिकार्ड की तरह शेयर बाजार में भी गिरावट का रिकार्ड जारी है । <end><br>ACTUAL SENTENCE ===> लेकिन अब समय ने पलटी मारी है और फिर सचिन के शून्य में आउट होने के रिकार्ड की तरह शेयर बाजार में भी गिरावट का रिकार्ड जारी है । <end> |
| OUTPUT 8 | INPUT SENTENCE ===> उसनें बिपाशा ने यह स्वीकार किया है कि उन्हें कभी करीना को समझने का मौका ही नहीं मिला ।<br>PREDICTED SENTENCE ===> वहीं बिपाशा ने यह स्वीकार किया है कि उन्हें कभी करीना को समझने का मौका ही नहीं मिला । <end><br>ACTUAL SENTENCE ===> वहीं बिपाशा ने यह स्वीकार किया है कि उन्हें कभी करीना को समझने का मौका ही नहीं मिला । <end> |

# Conclusion

- The grammatically incorrect sentences are getting corrected by this model.
- The model trained using the attention mechanism outperformed the one without attention.
- The best results were obtained from the model trained using the Fasttext word embedding.

# Future Work

- Etoori dataset have only one error per sentence. A more error inclusive dataset (dataset with multiple form of errors and more than one error per sentence) can be prepared.
- Spelling Correction can also be added in the proposed architecture.
- Dynamic embeddings like BERT, ELMo etc., can also be incorporated to get the contextual vector representation of a word, thus helping in better learning.
- Other Evaluation techniques can be incorporated.
- The process can be generalized for other low-resource Indian languages for GEC tasks and made open source.

# Thank You