**Sentence Correction using RNN’s(Deep learning)**



**Sentence Correction using RNN’s**

**Instigation:**

The information exchange between people rapidly increasing through different social media channels (Facebook, Twitter etc.). People started texting short form of messages and using different terminologies to convey the information. This short form of messages(Fast text) creating big impact and drastically reducing the performance of machine learning models which runs on this data. And it slowly led to make wrong predictions and wrong interpretation from ML models. **Gene Lewis** published a research paper from Stanford University and proposed an approach to overcome this problem in the field of AI.

**Research Paper:**

<https://cs224d.stanford.edu/reports/Lewis.pdf>

**Source Dataset:**

<http://www.comp.nus.edu.sg/~nlp/sw/sm_norm_mt.tar.gz>

In this article i am trying to implement Gene Lewis proposal using recurrent neural networks.

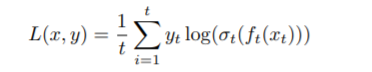
**Contents**

1. Map to Deep Learning Problem
2. Synthetic Data Generation(Artificial Data)
3. Preprocessing and Tokenization
4. Design Encoder and Decoder Neural Network
5. Beam Search
6. Final Pipeline

**1.Map to Deep Learning Problem**

 The source dataset contains only 2K records and each record having Social media text message, Chinese message and original English text. Extracted social media text messages and original English text messages for this problem.

Using encoder and decoder models, to convert corrupted social media text messages into proper English text messages. Converted English messages feed to machine learning models and thus increase the model performance. Here used categorical cross entropy as metric. Cross-entropy loss increases as the predicted probability diverges from the actual class.



**Cross Entropy**

2. **Synthetic Data Generation (Artificial Data)**

Training a neural network with insufficient data may not help in achieving the good accuracy. As dataset size is very small, generated synthetic data using nlpaug library. Used synonym augmentation to keep the semantic meaning of the sentence along with fast text word embedding model to generate the artificial data.

**Gist Link:** https://gist.github.com/Madhusudhan77/5e9488bc551166323719ab1ca87b5d43

**3. Preprocessing and Tokenization**

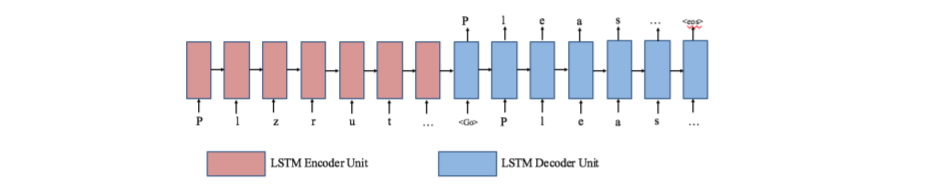
Preprocessing step involved in removing special character and de-contracted the strings present in dataset. Splitting entire preprocessed data into train, test and validation sets .Performed tokenization and created word embedding for both encoder and decoder using fast text Model .

**4. Design Encoder and Decoder Neural Network**

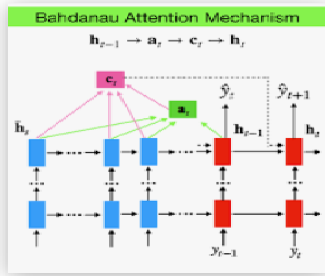
***1.Simple Encoder Decoder neural network***

***2. Encoder and Decoder with Bahdanau’s Attention***

***1.Simple Encoder Decoder neural network:*** Tried simple encoder and decoder with 300 LSTM units and tanh as activation function. Used soft max as last layer of neural network to get the output probabilities for good interpretation. Trained neural network with 50 epoch's and at the end of the training, got trained accuracy(85%) and loss(0.2602%).

**LSTM Encoder and Decoder**

***2. Encoder and Decoder with Bahdanau’s Attention:*** Designed Encoder and Decoder with different attention mechanism's and trained neural network. Encoder and Decoder with Bahdanau’s attention with one step decoder given good train and test accuracy. Used LSTM with 300 units and trained model with 25 epoch's and got 95% train accuracy and cross entropy loss 0.0850.



**Bahdanau’s Attention Mechanism**

**5.Beam Search:**

NMT is using a simple left-to-right beam-search decoder to generate new translations that approximately maximize the trained conditional probability. The current beam search strategy generates the target sentence word by word from left-to- right while keeping a fixed amount of active candidates at each time step. Beam search decoder increased the performance of Encoder and Decoder model and helped in making good language translation.

**Gist Link:** https://gist.github.com/Madhusudhan77/3cd7926b67fe76188ba823f48aece6f6

**6. Final Pipeline:**

**Link :** https://vimeo.com/559388689

**Future Works**

1. As source dataset size is very less(2K), can use more augmentation technics and NLP mechanism's to generate the artificial data .
2. There might be a possibility to use different RNN’s to increase the performance .Ex : bidirectional LSTM’s.

**Reference Links:**

1. <https://www.appliedaicourse.com/>
2. <https://arxiv.org/abs/1702.01806>
3. <https://machinelearningmastery.com/beam-search-decoder-natural-language-processing/>

**Git code:**

[**Madhusudhan77/Sentence-Correction-Using-RNN-s**  
*Sentence Correction using RNN's(Deep learning) Instigation: The information exchange between people rapidly increasing…*github.com](https://github.com/Madhusudhan77/Sentence-Correction-Using-RNN-s)

**LinkedIn:**

[www.linkedin.com/in/reddivari-madhusudhan-42a002195](http://www.linkedin.com/in/reddivari-madhusudhan-42a002195)