Corpus-Based Machine Translation System with Deep Neural Network for Sanskrit to Hindi Translation

* Translation from Sanskrit to Hindi
* Bhagavad Gita would be used as an input data
* Deep Neural Network would be used for training
* Data would be passed through a neural network after data analysis and processing
* The solution proposes a better BLEU Score (Billingual Evaluation Understudy) and Word error rate.
* Steps of MTS:-

1. Decoding: where the meaning of the text is decoded
2. Re-encoding: where the decoded text is re-encoded into the target language.

* Different approaches adopted for MTS are:

1. Rule-Based Approach
2. Dictionary Based Approach
3. Corpus-Based Approach
4. Knowledge-Based Approach
5. Hybrid Approaches

* Part 2 discusses some of the previous works done in the field of MTS
* This technique does not require any rules or dictionaries because they automatically learn about the language from the large set of corpora
* The work is done in three phases, which are again subdivided into various small steps, namely,
  + - * 1. Data Analysis

Data Collection

Data Cleaning

Data Preprocessing

Data Visualization

* + - * 1. Data Processing

Tokenization

Parsing

Splitting

Pass to Hidden Layers

* + - * 1. Target Generation

Training

Testing

* Data Analysis
  + Collecting data to form a big dataset, containing all the word meanings systematically, arranged in a proper format, in a single .csv file.
  + Cleaning the dataset by removing all the redundant or unused words and extra symbols.
  + Visualization of the data to check the structure and the co-relation of the dataset.
  + If no relation is found, then preprocessing is done from the starting.
* Data Processing
  + Tokenization is the chopping the data into words
  + Store all the dataset into matrix format and then Tokenize the data into words and label it into the numeric format.
  + Use Count Vectorise technique which converts all data into numeric format
  + Then the data is analyzed using lexical and semantic analysis. Arrange the data into the grammatical structure of Sanskrit.
  + Data is split into two parts—one for model development for predictive analysis and the other for performance analysis.
  + Divide the data for training and testing purposes.
* Target Text Generation
  + Training is performed through the deep neural network.
  + Data is passed through a huge number of hidden layers, and the number of hidden layers will increase automatically through auto-tuning.
* The performance of the Corpus-Based Approach is 24% better than the Rule-Based Approach, as calculated through the BLEU score.
* Word Error Rate – is a metric used to calculate error rate by comparing machine-translated output with the human-translated output.
* Lesser the WER, the better the model is.
* The performance of the Corpus-Based Approach is 39.6% better than the Rule-Based Approach, as calculated by the WER.
* In the model, Keras is used in the front end, and Tensor Flow is used in the back end.