INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

HYDERABAD

International Institute of Technology - Hyderabad
SMAI - SPRING 2019 Course Project

PREDICT MOVIE TAGLINE FROM TEXT SYNOPSIS

Project ID - 11

Team 31- PGSSP ML TECHNIES

Project guide : Satyam mittal

Project members : Jyothsna – Madan – Surekha - Karthikeyan

ABSTRACT

Predicting movie taglines from the context of movie plot synopsis. Our model is a bidirectional encoder-decoder RNN. It was optimized for best performance with tests utilizing Long short-term memory (LSTM) cells as building blocks. Attention model for predicting each word of the tagline conditioned on the movie plot. While the model is structurally simple, it can easily be trained end to end and scales to large amount of training data. We apply our model to the IMDB movies dataset. We evaluate the predicted tagline using standard metrics like ROGUE score, showing that model can encode texts in a way that preserve its syntactic and semantic coherence

GIVEN PROBLEM STATEMENT

- Predicting Movie Tagline for any Given Movie Plot
- It deals with the context of movie synopsis -NLP Problem

CHALLENGES

Data:

- Missing values
- Non UTF8 Text contents
- Non English Text Contents

CHALLENGES – CONT.

Training Model:

- Huge Volume of data to be processed
- Computation complexity leads to Long hours of training
- Risk of Losing training data while executing training for long hours – Alternatively we used Checkpoints to save progress of the training.

CHALLENGES – CONT.

Unknown words in Data set:

- There is always possibility of getting a Unknown word which is not part of word embedding matrix which we created. i.e Out of Vocabulary Words

- Examples : Name of Characters in the movie like Sivagami, Baahubali etc.

CHALLENGES – CONT.

Training Model:

- Huge Volume of data to be processed
- Computation complexity leads to Long hours of training
- Risk of Losing training data while executing training for long hours – Alternatively we used Checkpoints to save progress of the training.

IMPLEMENTATION DETAILS

- Bi Directional – Encoder decoder RNN LSTM
- Attention Mechanism
- TensorFlow based implementation.
- Batch processing of Data

IMPLEMENTATION DETAILS , CONT.

- Word Vector Embedding .
  - Numberbatch word embedding* is built on:
    - ConceptNet 5.5,
    - GloVe,
    - word2vec,
    - Parallel text from OpenSubtitles 2016
  - * Refer GitHub Repository mentioned in the reference section
-

IMPLEMENTATION DETAILS , CONT.

- Intermediate Training Data/Weights is saved as Check points.
  - Can be Loaded back to system later.
  - More training data can be trained on top of existing Checkpoint.
  - Saved Check point can be distributed as Binary file to others to deploy the model

- Training Stop Criteria – Early stopping when the error did not improve for continuous iterations.

DEMO

FUTURE WORKS

- Tuning of model to avoid the <UNK> words and repetitions of words

- Example:Point Generator model

REFERENCES – PART 1

- Predicting Movie Genres Based on Plot Summaries : https://arxiv.org/pdf/1801.04813.pdf

- Folksonomication: Predicting Tags for Movies from Plot Synopses UsingEmotion Flow Encoded Neural Network - https://aclweb.org/anthology/C18-1244

REFERENCES – PART 2

- Data Source:

- Found in Kaggle - https://www.kaggle.com/rounakbanik/the-movies-dataset

- Filename : movies_metadata.csv

- Concept Net Number batch – https://github.com/commonsense/conceptnet-numberbatch

- Glove - https://nlp.stanford.edu/projects/glove/

REFERENCES – PART 3

- Why We used Rouge scores : https://stats.stackexchange.com/questions/301626/interpreting-rouge-scores

- RNN – Sequence to Sequence Model : https://towardsdatascience.com/seq2seq-model-in-tensorflow-ec0c557e560f

Thank you!