**Text\_Summarization\_CNN\_DailyMail**

**Project Proposal**

**Author(s): Hemangi Kinger, Varun Shah, Ishan Kuchroo**

**11/7/2022**

**Abstract**

According to IBM, Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

The final project details the implementation of tools, libraries and techniques acquired during the course.

**Introduction**

Abstractive text summarization is the task of generating a headline or a short summary consisting

of a few sentences that captures the salient ideas of an article or a passage. We use the adjective ‘abstractive’ to denote a summary that is not a mere selection of a few existing passages or sentences

extracted from the source, but a compressed paraphrasing of the main contents of the document,

potentially using vocabulary unseen in the source document.

CNN/Daily Mail is a dataset for text summarization. Human generated abstractive summary bullets were generated from news stories in CNN and Daily Mail websites as questions (with one of the entities hidden), and stories as the corresponding passages from which the system is expected to answer the fill-in the-blank question. The authors released the scripts that crawl, extract, and generate pairs of passages and questions from these websites.

**Data Overview**

**Overview**:

In all, the corpus has 286,817 training pairs, 13,368 validation pairs and 11,487 test pairs, as defined by their scripts. The source documents in the training set have 766 words spanning 29.74 sentences on an average while the summaries consist of 53 words and 3.72 sentences. The dataset is inspired from the one used in the Association for Computational Linguistics (ACL) 2017 paper [Get To The Point: Summarization with Pointer-Generator Networks](https://arxiv.org/pdf/1704.04368.pdf).

**Source**:

<https://cs.nyu.edu/~kcho/DMQA/>

**Snapshot**:

Text

Description automatically generated

**NLP – Methods and Procedures**

For this project, we model abstractive text summarization using Recurrent Neural Networks, and show that they achieve state-of-the-art performance on two different corpora.

**Recurrent Neural Networks**:

According to Wikipedia, a recurrent neural network is a class of artificial neural networks where connections between nodes can create a cycle, allowing output from some nodes to affect subsequent input to the same nodes.

Diagram

Description automatically generated

**Procedures**:

We will try and customize the RNN’s according to the dataset. We will study the existing procedures done on the dataset (given in <https://paperswithcode.com/dataset/cnn-daily-mail-1> ) and try to build our own Attentional Encoder Decoder.

**Packages**: Tensorflow / Pytorch, Numpy, Pandasm etc.

**Metrics**: Paperwithcode website for this dataset has various benchmarks. We’ll try to build/improve existing models and compare the accuracy of the model with existing benchmarks.

**Appendix**

GitHub Link - <https://github.com/IshanKuchroo/Text-Summarization-for-CNN-and-DailyMail>

**References**

<https://paperswithcode.com/dataset/cnn-daily-mail-1>

<https://cs.nyu.edu/~kcho/DMQA/>