CS 491 NLP Project Report 2

Group No.: **G18**

Group Members:

Sr No.	Name	Enrollment No.	Email
1	Akshat Khanna	U101115FCS046	akshat.khanna@st.niituniversity.in
2	Chinmaya Bansal	U101115FCS077	chinmaya.k.bansal@st.niituniversity.in
3	Mayank Singh	U101115FCS200	mayank.singh@st.niituniversity.in
4	Sangamesh Kotalwar	U101115FCS210	sangameshn.kotalwar@st.niituniversity.in

Reference Paper Title: **Summarizing Lengthy Questions**

Authors: Tatsuya Ishigaki, Hiroya Takamura and Manabu Okumura

Goal of your work

Summarizing lengthy questions is the problem of creating a short, accurate, and fluent summary of a longer questions.The goal of this project is to summarize a question which can often be lengthy, helps respondents understand the question. Approaches used in generic summarization tasks are often classified into two different types: extractive and abstractive.

1. **Extractive text summarization** involves the selection of phrases and sentences from the source document to make up the new summary. Techniques involve ranking the relevance of phrases in order to choose only those most relevant to the meaning of the source.
2. **Abstractive text summarization** involves generating entirely new phrases and sentences to capture the meaning of the source document. This is a more challenging approach, but is also the approach ultimately used by humans. Classical methods operate by selecting and compressing content from the source document.

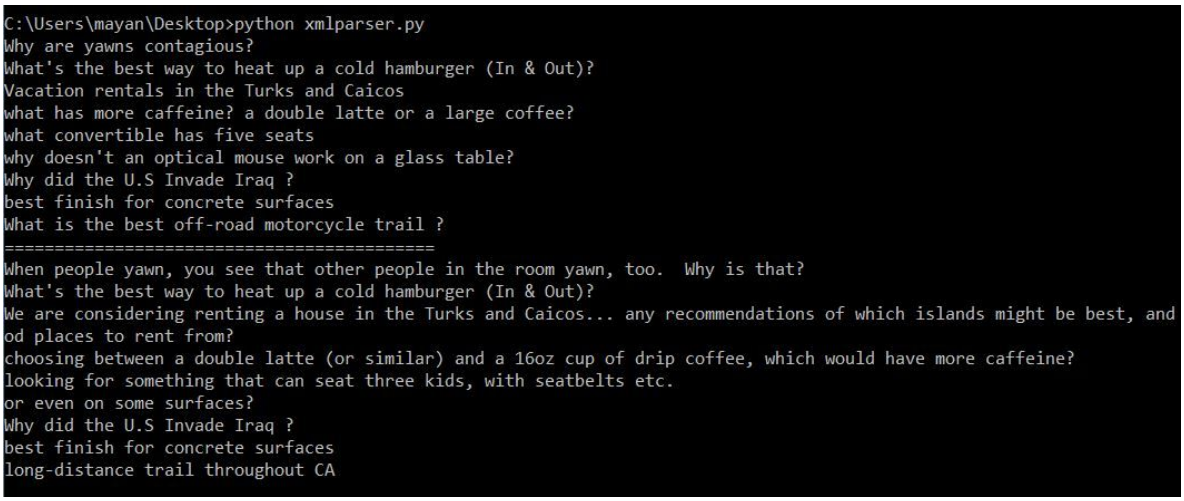
We are going to run these two approaches on same dataset and show that some of the summarization cannot be done by extractive approach but requires abstractive

approach.And the evaluation of the summarisation would be done using ROUGE-N metric.

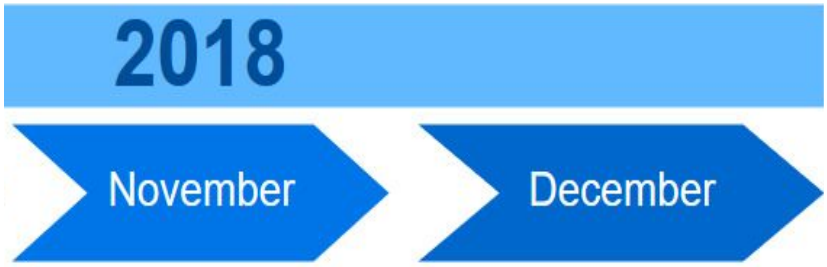
Work done so far

1. Reading of paper and going through references to understand the concept more.
2. Since the paper was based on the dataset *Yahoo! Answers Comprehensive Questions and Answers version 1.0,* which was not publicly available, so we applied for the dataset as research purpose. Now we have got access to the dataset so we will be proceeding with the analysis.
3. The size of dataset was around 5.7GB and we faced difficulty while parsing the data into an xml parser so we narrowed down the data to only the required xml tags which are <subject> and <content>.The new size was brought down to 216MB. We will be using the <content> data in extractive approach and both <subject> and <content> data will be used as training dataset for abstractive approach.
4. Extractive approach to summarisation which involves tasks like
    a. Reading data from source and performing cleanup and formatting.
    b. Tokenizing Input
    c. Create a frequency scoring system
    d. Selection of top N sentences based on their score

    Our data after parsing (Subject and content of the text segregated)



Plan of work and responsibilities:



1. Analysis using extractive approach
   -> Rule based approach
   -> Machine learning based approach
2. Analysis using abstractive approach
   -> Vanilla encoder - decoder
   -> Encoder - decoder with attention
   -> Encoder - decoder with copying mechanism

Responsibilities:
Akshat - Rule based approach, vanilla encoder - decoder, encoder - decoder with copying mechanism

Chinmaya - Rule based approach, vanilla encoder - decoder, encoder - decoder with attention

Mayank - Machine learning based approach, encoder - decoder with attention

Sangamesh - encoder - decoder with copying mechanism, machine learning based approach.

1. Experiments
2. Evaluation
   -> Evaluation with ROUGE
   -> Manual Evaluation
3. Qualitative Analysis
4. Final report submission

Responsibilities:
Akshat - Experiment and evaluation with ROUGE

Chinmaya - Evaluation with ROUGE and manual evaluation

Mayank - Manual evaluation and qualitative analysis.

Sangamesh - Qualitative analysis, cross checking evaluations and final report