Abstractive summarization

Mandate-1

-Aakanksha Rani
MT2022001

1. Problem Statement :

Given a set of tweets pertaining to a trending topic, create an abstractive prose summary of the tweets. Do not just string the tweets together to form the summary. The summary will need to paraphrase and/or say more than what is directly said in the tweets. Propose a rubric to evaluate the accuracy of your summarization.

2. Introduction :

Recently twitter has become one of the most popular networking sites. It enables people to freely post short messages up to 140 characters and express their views on current trending topics. But it is impossible for any twitter user and organizations to get an overview of what's going around by reading each and every tweet out of millions of tweets. Also for extracting important information from a set of information in a short period of time people want an efficient way.

So abstractive summarization can be claimed as an effective key to solve the above problem .A summary that provides representative information of the topic without any biases and with well formed ,easy to understand and clear sentences would be preferred.
NLP, the field of artificial intelligence that studies the interaction between human and computer and helps in designing programs to process large amounts of natural language data, is a domain on which we will work for the above problem statement's solution .

<u>Language -</u> *A system of conventional spoken , manual or written symbols by means of which human beings communicate and express themselves.*

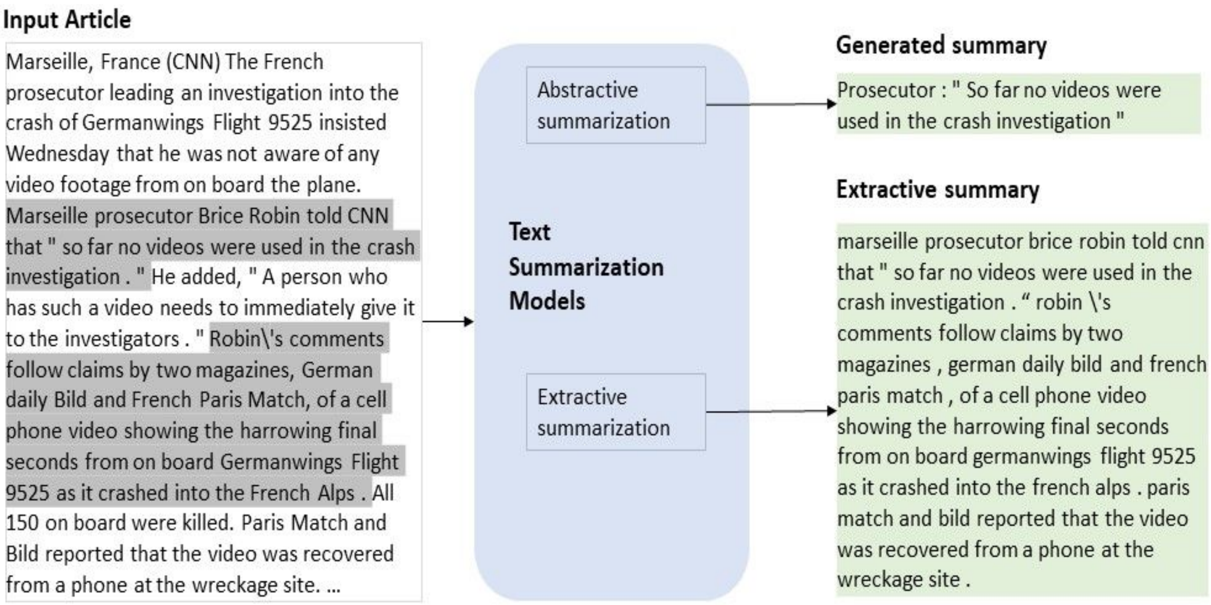
<u>Text Summarization -</u> It reduces the number of sentences and words of a document without changing its meaning. There are different methods to extract information from raw text data and use it for a summarization model.

1- Extractive Approaches -

Using an extractive approach we summarize our document on the basis of frequency method i.e. sentences with most frequently used words are used to form summary.The extractive method employs the use of TextRank to create a summary.

2- Abstractive Approaches -

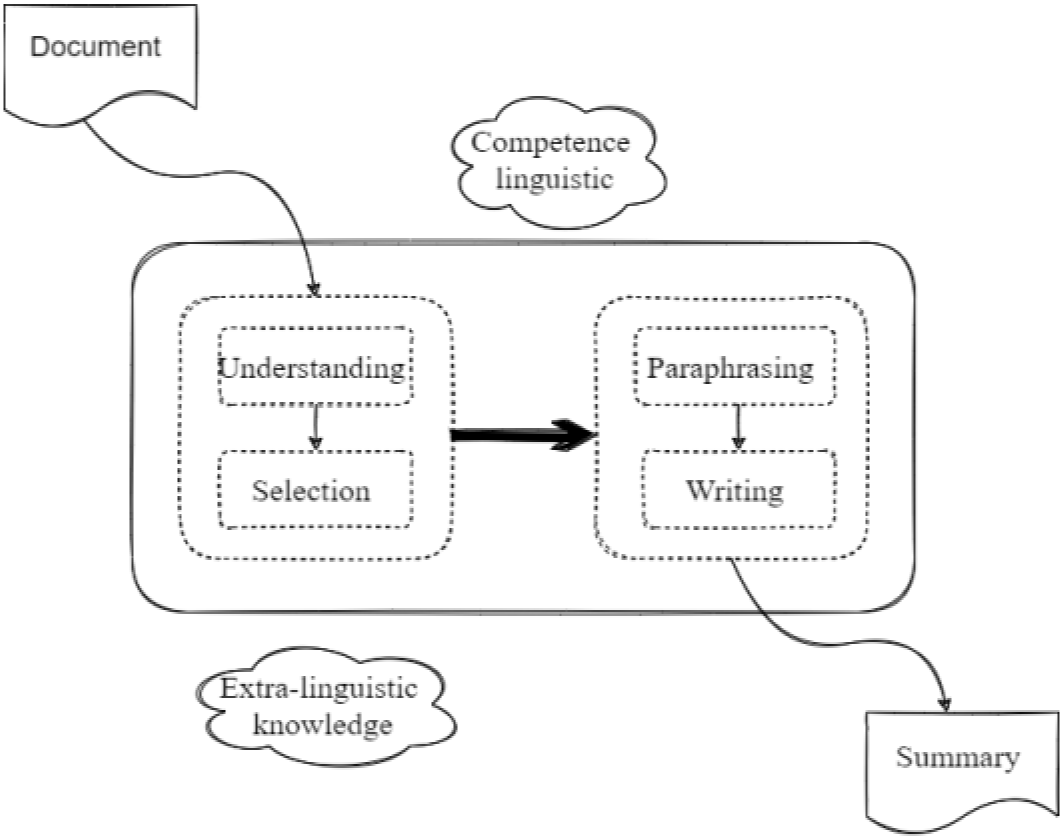
An abstractive approach is more advanced . It does not select sentences from the originally given text passage to create the summary ,instead it produces a paraphrasing of the main contents of the given text. It is similar to what we humans do to summarize.



Marseille, France (CNN) The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane. Marseille prosecutor Brice Robin told CNN that " so far no videos were used in the crash investigation . " He added, " A person who has such a video needs to immediately give it to the investigators . " Robin\'s comments follow claims by two magazines, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings Flight 9525 as it crashed into the French Alps . All 150 on board were killed. Paris Match and Bild reported that the video was recovered from a phone at the wreckage site. ...

Prosecutor : " So far no videos were used in the crash investigation "

marseille prosecutor brice robin told cnn that " so far no videos were used in the crash investigation . “ robin \'s comments follow claims by two magazines , german daily bild and french paris match , of a cell phone video showing the harrowing final seconds from on board germanwings flight 9525 as it crashed into the french alps . paris match and bild reported that the video was recovered from a phone at the wreckage site .

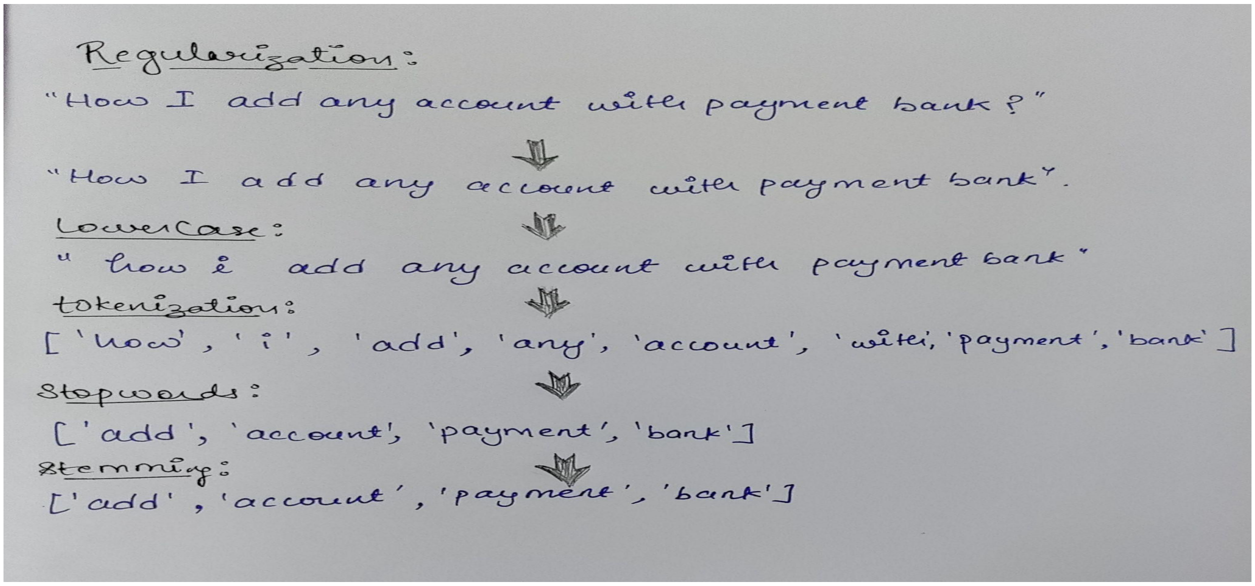


Human’s Process for generating summary

3. Problem Formulation :

1- For this problem statement firstly we need to have a corpus (means the text dataset to train the model) . So a collection of a large number of tweets on a specific topic.
Dataset - Large corpus of tweets

2- Cleaning the text dataset is the next step. Text pre-processing includes various steps .



Regularization:

"How I add any account with payment bank ?"

⇓

"How I add any account with payment bank".

Lowercase: ⇓

" how i add any account with payment bank "

tokenization: ⇓

['how', 'i', 'add', 'any', 'account', 'with', 'payment', 'bank']

Stopwords: ⇓

['add', 'account', 'payment', 'bank']

Stemming: ⇓

['add', 'account', 'payment', 'bank']

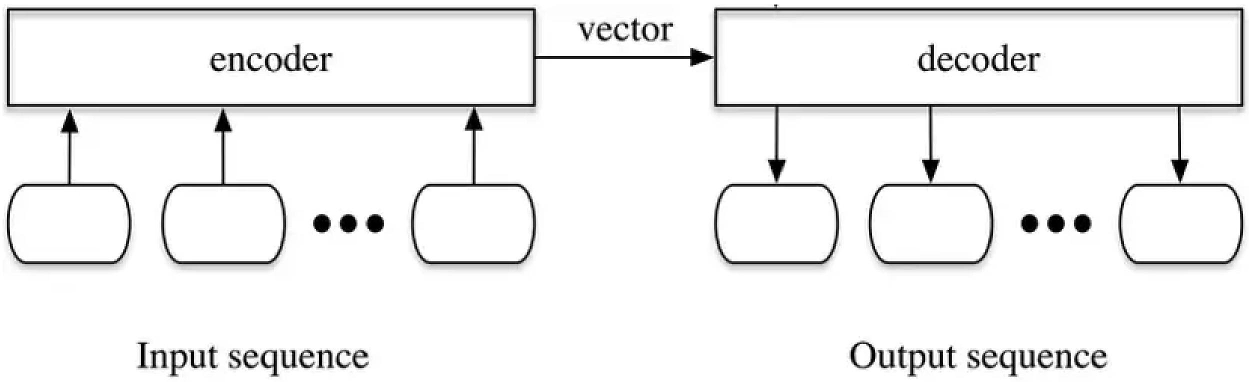
3- Then we will use embeddings. Embeddings represent text into numerical vectors that act as an input for our ML models .It tokenizes each word in a sequence (or sentence) and converts them into a vector space. Word embeddings aim to capture the semantic meaning of words in a sequence of text. It assigns similar numerical representations to words that have similar meanings.
There are various ways to do this like Word2vec, TF-IDF, GloVe, BERT etc.

4- Neural NLP is being preferred here to solve the given problem over statistical NLP.

Neural NLP Vs Statistical NLP

Statistical Language Model is based on assigning probabilities to sequences of words. Statistical NLP requires a high amount of preprocessing lemmatization, stemming, etc) ,feature extraction (NER, phrase identification, etc) and it does not have the capability to remember important things over the long term. But all these problems can be avoided by using the advanced and highly preferred neural nlp. In neural NLP we have amazing libraries such as seq2seq and transformers to build the model.

Seq2seq - *An encoder decoder based model* that takes a sequence as input and outputs another sequence . The encoder is multiple RNN cell and decoder is also a stack of RNN cells .



Transformers - They are a new technique somewhat different from Seq2seq,it is a deep learning model which is also an encoder-decoder based architecture that uses Attention mechanism but without relying on RNNs in order to speed up the model.

5- Text Post Processing- It involves spell checking , grammar and ensuring that the original text's meaning is contained .

6- Finally the last step involves outputting the paraphrased text (summary).

4. Challenges for abstractive text summarization :

We humans have capabilities to understand both conceptual modeling and language instinct . We can build rich abstract models but giving the same capability to a machine needs lots of effort and many challenges we have to face.

1.False Facts
2.Coreference Resolution (e.g. WINOGRAD SCHEMA)
3.Inaccurate tweets
4.Idioms and Phrases
5.Out of vocabulary words

5. Evaluation Metrics :

ROGUE- These metrics compare an automatically produced summary or translation against a manually generated summaries or translations. There are many variations of ROUGE such as one-grams, bi-grams, etc (ROUGE-n, ROUGE-L, ROUGE-SU) but none of them are the best one. Rogue measures recall.

BLEU- Bleu does the same except it measures precision.

6. References :

1-https://aclanthology.org/C12-1047.pdf
2-https://towardsdatascience.com/text-summarization-with-nlp-textrank-vs-seq2seq-vs-bart-474943efeb09
3-https://www.topcoder.com/thrive/articles/text-summarization-in-nlp
4-https://arxiv.org/ftp/arxiv/papers/2204/2204.01849.pdf
5-Classroom Slides- (NLP Mandate-1 (Linguistic Fundamentals))