# ABEER AHMAD (38) – MOHAMMED DEIFALLAH (59)

# TEXT SUMMARIZATION

**Tuesday, March 20th 2019**

## OVERVIEW

Text summarization is the process of generating a concise and logical version of a larger original document; since reading the whole corpus or summarizing it manually is absolutely a tedious job and waste of both effort and time. The main idea of summarization is to capture a subset of data that contains the “information” of the entire set.

It is a challenging problem as it processes a very large amount of data. It is even more complex to deal with unstructured data into a traditional database.

## PROJECT SCOPE

- **NLP:** Applying *Extractive* or *Abstractive* method, whichever is suitable and feasible.
- **Encoding:** Application of *Universal Sentence Encoder*.

## HIGH-LEVEL REQUIREMENTS

The system should support the following functionalities:

- It can extract features from the given sentences, understanding them well.
- It can generate a semantically and linguistically correct text.

## INPUT/OUTPUT EXAMPLE

Following is an illustrative example of already-existing summarization model:

- **Input:** Wikipedia Article on Artificial Intelligence.
- **Output:** “Artificial intelligence (AI), sometimes called machine intelligence, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans and other animals. Many tools are used in AI, including versions of search and mathematical optimization, artificial neural networks, and methods based on statistics, probability and economics. The traditional problems (or goals) of AI research include reasoning, knowledge representation, planning, learning, natural language processing, perception and the ability to move and manipulate objects. When access to digital computers became possible in the middle 1950s, AI research began to explore the possibility that human intelligence could be reduced to symbol manipulation. One proposal to deal with this is to ensure that the first generally intelligent AI is 'Friendly AI', and will then be able to control subsequently developed AIs. Nowadays, the vast majority of current AI researchers work instead on tractable "narrow AI" applications (such as medical diagnosis or automobile navigation). Machine learning, a fundamental concept of AI research since the field's inception, is the study of computer algorithms that improve automatically through experience.”

## DATASET

We decide to work on either *CNN or DailyMail* dataset, as a part of ***DeepMind Q&A*** dataset. In details:

- ***CNN*** dataset approximately contains 92,570 articles in “*.story”* files. Moreover, each article has some highlighted meaningful bullet points.
- ***DailyMail*** contains 219,503 articles with 3-4 bullet points as the gold label sentences for each one.

## MODEL ARCHITECTURE

Most probably will be an encoder-decoder model; where:

- ***The Encoder*** reads the complete input sequence and encodes it into an internal representation, usually a fixed-length vector, described as the context vector. Various types of encoders can be used—more commonly, bidirectional RNNs, such as LSTMs, are used.
- ***The Decoder,*** on the other hand, reads the encoded input sequence from the encoder and generates the output sequence.

## STATE-OF-THE-ART MODELS

1. Pretraining-Based Encoder-Decoder Framework
2. Generative Adversarial Network
3. Pointer-Generator Networks

| | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|
| **Model 1** | **41.71** | **19.49** | **38.79** |
| **Model 2** | 39.92 | 17.65 | 36.71 |
| **Model 3** | 39.53 | 17.28 | 36.38 |

## EVALUATION METRIC

***ROUGE*** (Recall-Oriented Understudy for Gisting Evaluation) is best known as an evaluation metric for automatic summarization and natural language processing tasks in general. We will use it as a metric to compare the output text to the gold label sentences.

## APPLICATIONS

Possible applications of the proposed model:

- Summarization of relatively long paragraphs for people with cognitive disorders.
- Shortening input sentences for search engines.
- Summarization of lectures, reference chapters, articles, etc.

## GRADUATION PROJECTS

| | **Visual Question Answering** | **Video Question Answering** |
|---|---|---|
| **Supervisors** | Dr. Mohamed Ismail<br>Dr. Nagia Ghanem | Dr. Nagwa El-Makky<br>Dr. Marwan Torki |
| **Relevancy** | Not relevant | Not relevant |

## RESOURCES

- Automatic Text Summarization: Past, Present and Future.
- A Gentle Introduction to Text Summarization.
- Encoder-Decoder Deep Learning Models for Text Summarization.
- Text Summarization Using Keras Models.
- Text Summarization Using Deep Learning Made Easy.
- Unsupervised Text Summarization Using Sentence Embeddings.
- Understand Text Summarization and Create Your Own Summarizer in Python.
- An Introduction to Text Summarization Using the TextRank Algorithm.