



**T.C.
MARMARA UNIVERSITY
FACULTY OF ENGINEERING
COMPUTER ENGINEERING DEPARTMENT**

CSE4078 Introduction to NLP

Alpaca Style Report

April 25th, 2024

Group 8 Members

Alper Özdemir

Eren Başpınar

Emirhan Erdoğan

Faruk Akdemir

Şule Koca

Text Summarization

Text Summarization is a natural language processing (NLP) task that involves condensing a lengthy text document into a shorter, more compact version while still retaining the most important information and meaning. The goal is to produce a summary that accurately represents the content of the original text in a concise form.

This can be achieved through various techniques, broadly categorized into extractive and abstractive summarization.

Extractive Summarization

It involves selecting the most relevant sentences or phrases from the original text and combining them to create a summary.

Techniques used in extractive summarization include:

TF-IDF (Term Frequency-Inverse Document Frequency): This method evaluates the importance of each word in a document relative to a collection of documents. Sentences containing the most significant words are selected for the summary.

TextRank: Inspired by Google's PageRank algorithm, TextRank assigns scores to sentences based on their importance within the document's context and connectivity to other sentences. The top-ranked sentences form the summary.

Graph-based approaches: These methods represent sentences as nodes in a graph, where edges between nodes represent the relationship between sentences. By analyzing the graph structure, important sentences are identified for summarization.

Abstractive Summarization

It aims to generate a concise summary in the system's own words, potentially rephrasing and restructuring sentences to convey the main ideas.

Techniques used in abstractive summarization include NLP, attention mechanisms and language generation models. NLP techniques such as sequence-to-sequence models, recurrent neural networks (RNNs), and transformers can be used to generate summaries by understanding the context of the text and generating new sentences to capture the main points. Attention Mechanisms allow the model to focus on relevant parts of the input text when generating the summary, improving coherence and relevance. Language Generation Models are advanced language models like GPT (Generative Pre-trained Transformer) are trained on vast amounts of text data and can generate human-like summaries by predicting the next sequence of words given an input text.

Hybrid Approaches

Some summarization techniques combine elements of both extractive and abstractive methods to produce summaries. For instance, a system might first extract key sentences using extractive techniques and then employ abstractive methods to rewrite and refine the summary.

Dataset Identification

Row	Dataset Name	URL	Source	Description	Instance Number
1	TR-News	TR-News	Hugging Face	News Description	
2	lr-sum	lr-sum	Hugging Face	lr-sum Description	
3	wikipedia-tr-summarization	wikipedia-tr-summarization	Hugging Face	wikipedia-tr-summarization Description	

TR-News

The TR-News dataset is a collection of Turkish news articles designed for use in natural language processing tasks such as text summarization. It contains a total of 307,562 rows, each representing a news article. The dataset includes various fields such as abstract, author, content, date, source, tags, title, topic, and URL.

- **Abstract:** A brief summary or description of the news article.
- **Author:** The author(s) of the news article.
- **Content:** The main body of the news article, containing detailed information about the news story.

- **Date:** The date when the news article was published.
- **Source:** The source or publication from which the news article originates.
- **Tags:** Keywords or labels associated with the news article, indicating its topics or themes.
- **Title:** The headline or title of the news article.
- **Topic:** The broader category or subject matter to which the news article belongs.
- **URL:** The web address or link to access the full news article online.

The dataset was introduced in the paper "Abstractive text summarization and new large-scale datasets for agglutinative languages Turkish and Hungarian" by Batuhan Baykara and Tunga Güngör, published in the journal Language Resources and Evaluation in 2022.

Lr-sum


LR-Sum is an automatic summarization dataset that focuses on less-resourced languages. It contains human-written summaries for news articles in 39 languages, sourced from the Multilingual Open Text corpus based on Voice of America newswire text. The dataset is released under a Creative Commons license (CC BY 4.0), making it openly accessible for research in automatic summarization. Curated by the BLT Lab and shared by Chester Palen-Michel, this dataset includes summaries for news articles in languages such as Albanian, Amharic, Armenian, Azerbaijani, Bengali, Turkish and many others.

The dataset's structure includes essential fields such as:

- **'id':** Unique identifier for each article
- **'url':** URL linking to the original news article
- **'title':** Title of the news article
- **'summary':** Human-written summary of the article
- **'text':** Full text of the news article (excluding the title)

LR-Sum is sourced from the Multilingual Open Text v1.6 corpus, specifically from Voice of America newswire texts.

Wikipedia-tr-summarization

This is a Turkish summarization dataset  prepared from the 2023 Wikipedia dump. The dataset has been cleaned, tokenized, and summarized using Huggingface Wikipedia

dataset cleaner script, custom cleaning scripts, and OpenAI's gpt3.5-turbo API. The dataset includes text and summary. The data size is 341,537,415 bytes.

The dataset was introduced in the paper " Wikipedia Turkish Summarization Dataset" by Musab Gültekin, published in the Hugging Face in 2023.

New Datasets For Iteration 2

Turkish News Summarization Dataset from XL-Sum in Hugging Face web page:

<https://huggingface.co/datasets/csebuetnlp/xlsum/viewer/turkish>

- **id**
string lengths
854
- **url**
string lengths
35337
- **title**
string lengths
14163
- **summary**
string lengths
41.01k
- **text**
string lengths
11939.3k

MLSUM-Multilingual Summarization Dataset from Kaggle web page.

<https://www.kaggle.com/datasets/thedevastator/mlsam-multilingual-summarization-dataset>

This tu_train.csv is a file in the MLSAM dataset that contains Turkish articles and their corresponding summaries, along with additional information such as topic, URL, title, and date. All text has 246498 unique values so input numbers are 246490 and output numbers are 245094.

Alpaca Style Dataset Representation

In the picture below, you can see the example of how our datasets are represented as json file format. You can access the whole file from our documents.

```
CSE4078S24_Grp8_AlpaStyle_LrSum.json X
C: > Users > admin > Desktop > datasets > CSE4078S24_Grp8_AlpaStyle_LrSum.json
1 {
2   {
3     "instruction": "Aşağıdaki metni özetle",
4     "input": "İSTANBUL – Cumhurbaşkanı Recep Tayyip Erdoğan, Suriye konusunu görüşmek üzere İran ve Rusya liderleriyle bir araya gelecek. Bu Erdoğan'la İran Cumhur
5     "output": "Cumhurbaşkanı Erdoğan, Haziran ayında İran'ı bölgedeki nüfuz alanını arttırmakla suçlamış ve bunun mücadele edilmesi gereken bir tehdit olduğunu söy
6   },
7   {
8     "instruction": "Aşağıdaki metni özetle",
9     "input": "HOUSTON – Dünyanın bir bölgesinde düzenlenen bir saldırıyı televizyonlarımdan izleyip görmek mümkün. Ancak öyle saldırılar var ki, bunların yaşandığı
10    "output": "Savunmasız bilgisayar sistemlerini hedef alan siber korsanlar büyük felaketlere yol açabiliyor. En büyük kaygı enerji sistemlerinin hedef alınması.
11  },
12  {
13    "instruction": "Aşağıdaki metni özetle",
14    "input": "ANKARA – Zamlar geri alınmasını talebiyle Ankara Ulus Atatürk Heykeli önünde bir araya gelen DİSK, KESK, TMMOB ve ASMMMO üyeleri zamları protesto etti
15    "output": "DİSK, KESK, TMMOB ve ASMMMO üyeleri son dönemlerde çeşitli hizmet ve ürünlere getirilen zamları protesto etti, ülkeyi mali krize sürükleyen in, iktid
16  },
17 }
```

Statistics

Datasets	Number of Instructions	Average of Instructions	Standard Deviation	Input Length	Output Length
ML-Sum	249277	249277	1775.60355	2042.84167	145.806829
XL-Sum	27176	27176	2324.761663	3094.11407	178.055158
Wikipedia-TR-Summarization	119110	119110	1081.028574	2232.84568	272.254579
LR-Sum	28672	28672	2037.292967	3112.25024	174.514753
TR-News	277573	277573	1982.429520	1881.97642	165.533365