T.C.

MARMARA UNIVERSITY

FACULTY OF ENGINEERING

COMPUTER ENGINEERING DEPARTMENT

CSE4078 Introduction to NLP

Dataset Report

April 5th, 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	Row Number
1	TR-News	TR-News	Hugging Face	TR-News Description	307,562
2	lr-sum	lr-sum	Hugging Face	lr-sum Description	28,698
3	wikipedia-tr-summarization	wikipedia-tr-summarization	Hugging Face	wikipedia-tr-summarization Description	125,379

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.

abstract string · lengths	author string · lengths	content string · lengths	date string · lengths	source string · lengths	tags string · lengths	title string · lengths	topic string · lengths	url string · lengths
1⋅134 38.9%	null 82.3%	1⋅24.8k 99.9%	18⋅20 84.4%	7⋅12 41.4%	2⋅60 75.4%	47⋅69 25.7%	6⋅9 64.4%	99⋅117 23.2%
Şarkıcı Tuğba Özerk, annesine canlı yayında yönelttiği soru için Deniz Akkaya'ya 100 bin TL'lik manevi tazminat davası açtı.	null	Tuğba Özerk, "Duymayan Kalmasın" programında, annesi Günay Kapani'ye "Kızınızla aranız bozukmuş. Gerçekten onun sevgilisiyle birlikte oldunuz mu?" diye soran Deniz Akkaya'ya dava açtı. Ünlü şarkıcı, konuyla ilgili şöyle dedi: "Canlı yayında anneme akıl almaz bir soru yöneltti. Neye uğradığını şaşırdı kadın. Telefonu kapattıktan sonra rahatsızlandı. Deniz Akkaya hakkında suç duyurusunda bulundum, 100 bin liralık manevi tazminat davası açtım. Tazminatın tamamını şehit ailelerine bağışlayacağım."	13.01.2017 - 12:38	haberturk	[]	Tuğba Özerk'ten Deniz Akkaya'ya 100 bin TL'lik dava	Fiskos	https://www.haberturk.com/magazin/fiskos/haber/1348886-tugba-ozerkten-deniz-akkayaya-100-bin-tllik-dava
MHP Lideri Bahçeli, cumhurbaşkanlığı hükümet…	null	MHP Genel Başkanı Bahçeli, Manisa'nın Selendi ilçesinde temel atma ve topl…	23.11.2017 - 15:56	Anadolu Ajansı	['Siyaset', 'Türkiye', 'Manisa', 'Devlet…	Bahçeli: Cepheleşme keskinleşirse MHP buna…	Türkiye	https://www.ntv.com.tr/turkiye/bahceli-cephelesme-keskinlesirse-mhp-buna-tepkisiz-…
Pentagon, ABD askerlerinin Suriye'de Kürt birliklere…	Dış Haberler Servisi	Rusya'nın Suriye'ye havadan müdahaleye başlamasının ardından,…	29 Ekim 2015 Perşembe, 03:48	cumhuriyet	null	YPG ve ABD omuz omuza	dunya	http://www.cumhuriyet.com.tr/haber/dunya/398143/YPG_ve_ABD_omuz_omuza.html
Galatasaray'da forvet transferinde halen somut…	Cumhur Önder Arslan	Galatasaray'da forvet transferinde halen somut adımların atılmaması…	13 Ocak 2019 Pazar, 03:30	cumhuriyet	null	Nezedesin forvet?	futbol	http://www.cumhuriyet.com.tr/haber/futbol/1199763/Nezedesin_forvet_.html
Fenerbahçe Teknik Direktörü Christoph Daum, İspanyol…	null	İspanya'nın Marca Gazetesi'nde yayımlanan demecinde Daum, Güiza'nın…	26.04.2010 - 11:29	null	[]	Daum: Guiza mutsuz	Spor	https://www.ntv.com.tr/spor/daum-guiza-mutsuz,ZDGzN_PBakiY2GPKNTwCbw
Cumhurbaşkanlığı himayesinde bu yıl ilki…	null	Cumhurbaşkanlığı himayesinde, Kültür ve Turizm Bakanlığı Sinema Genel…	04.12.2018 - 13:20	Anadolu Ajansı	['Film', 'sinema', 'Sanat']	Uluslararası Dostluk Kısa Film Festivali başlıyor…	Sanat	https://www.ntv.com.tr/sanat/uluslararasi-dostluk-kisa-film-festivali-basliyor-76ulkeden-789-…
Galatasaray antrenörü Johan Neeskens, 'Ancak şunu…	null	Galatasaray Antrenörü Johan Neeskens, bu sezon istedikleri sonuçlara…	06.06.2010 - 15:48	null	[]	'Sonuna kadar mücadele ettik mi acaba ?'	Spor	https://www.ntv.com.tr/spor/sonuna-kadar-mucadele-ettik-mi-acaba,U1_rtQBICE6UFxB4Yocn7w
Yargıtay Cumhuriyet Başsavcılığı, Isparta'nın…	HABERTURK.COM	Yalvaç Ağır Ceza Mahkemesi, Isparta'nın Yalvaç ilçesinde…	16.01.2019 - 09:01	haberturk	['nevin yıldırım', 'yargıtay', 'son…	'Kesik baş' cinayetinde Nevin Yıldırım için…	Gündem	https://www.haberturk.com/kesik-bas-cinayetinde-nevin-yildirim-icin-muebbete-onama-istendi-2293249

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 TR 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.