Web Mining Final Homework Report

Turkish News Summarizer

Prepared by:

Oğuzhan Şahin 160709027

Nida Nur Kapmaz 160709014

Contents

1 Motivation

- Text summarization condenses a longer document into a short version while retaining core information. When this is done through computer, we call it Automatic Text Summarization.

- With the development of natural language processing, text summarization task has become more important.

- There are bunch of application to solve this task in the literature, however most of them are English.

- Our aim is to generate text summarization pipeline for Turkish.

2 Project Explanation

- Text summarization can be defined as “is the task of producing a concise and fluent summary while preserving key information content and overall meaning”.

- There are bunch of studies about this task. Most of them are in English.

- We aimed to create a Turkish abstractive text summarization pipeline from scratch (Crawler to deployment).

3 Used Technologies

- First of all we use Python.

- Scrapy crawler for Webtekno.com.

- Reques and bs4 librarieis for scraping.

- Fine-tuned BERT model for model.

- HTML, CSS and Flask for Web Site.

- As a future work, Heroku or Streamlit will be used.

4 Work Plan

Action	Oguzhan Sahin	Nida Kapmaz
Crawler	x	x
Scraping		x
Data Preprocessing	x	x
Labelling		x
Encoder-Decoder Model	x	
Flask	x	x
Deployment	x	

Figure 1: Contribution schedule of project members.

5 Pipeline

- Our pipeline consist of 6 components. These are: crawling, scraping, labelling, model, flask and deployment.

Figure 2 : Pipeline of a Turkish News Summarization project.

5.1 Crawler

- Web crawling is a powerful technique to collect data from the web by finding all the URLs for one or multiple domains.

- Python has several popular web crawling libraries and frameworks.

- A web crawler starts with a list of URLs to visit, called the seed. For each URL, the crawler finds links in the HTML, filters those links based on some criteria and adds the new links to a queue. All the HTML or some specific information is extracted to be processed by a different pipeline.

- We first build a crawler by using just basic libraries like requests and bs4.

- The code was very simple, however there are many performance and usability issues.

- The crawler was slow and supports no parallelism. It tooke about one second to crawl each URL.

- The crawler does not identify itself and ignores the robots.txt file.

- Then, we decided to use Scrapy library which is the most popular web scraping and crawling Python framework.

- One of the advantages of Scrapy is that request are scheduled and handled asynchronously. This means that Scrapy can send another request before the previous one is completed.

- We build a Scrapy crawler for Webtekno.

5.2 Scraping

- Web scraping is used for extracting data from websites.

- Collected about 18k URL from crawler script.

- By using collected URLs, scraped news data for every single URL.

- Turned into a .csv file, (columns: Url, Metin)

- Requests and bs4 libraries are used for this task.

5.3 Labelling

- Since automatic text summarization is mostly supervised task, the news needed to be labelled (summarized).

- For this task, we just used TFxIDF to summarize news. This methods returns importance sentences as a summary.

- Turned into a csv file (columns: Url, Metin, Özet)

5.4 Model

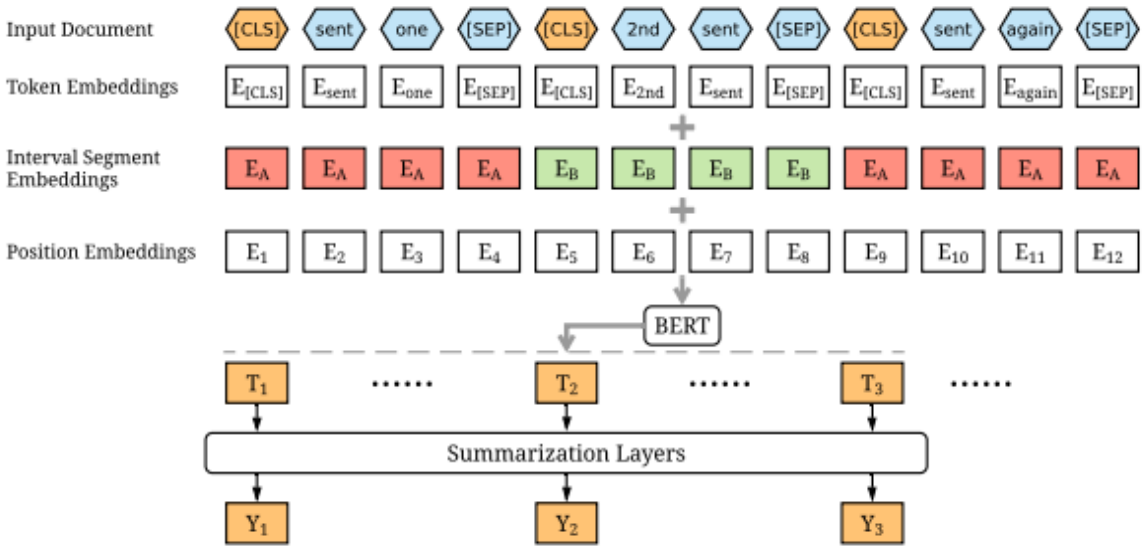

Figure 3: Working principle of the model.

- With the development of natural language processing, pre-trained language models (BERT, GPT, Electra) achieve state-of-art performance.
- We fine-tuned BERT for this task.
- Basically, the output of BERT was feed into summarization layers for summarization.
- In paper, author tested numbers of summarization layer's structure (RNN bases, Transformer based).

5.5 Flask

- Flask is a web framework. This means flask provides you with tools, libraries and technologies that allow you to build a web application.
- This web application can be some web pages, a blog, a wiki or go as big as a web-based calendar application or a commercial website.
- We used Flask to create summarization web application.

6 How Turkish News Summarizer Run

- If you do not have data, you can run crawler first. In scrapy_crawlers/spiders/, run below script:

  scrapy crawl webtekno --logfile webktekno.log -o webtekno.json -t jsonlines

- Once you run this script, you will have 2 files (webtekno.log, webtekno.json). In webtekno.json, you will have urls. If you change the urls, you need to adjust webtekno.py.
- In scrapy_crawler/spiders/, there is parse_json.py for parsing json files and gives an .csv files as an output.
- To get new text, scraping.py takes an input urls csv, and return urls and text csv file.
- tf_idf.py here will be used for labelling news text.
- Once you prepare your data for fine-tuning, you can run fine-tune.ipynb noteboook.

7 Results

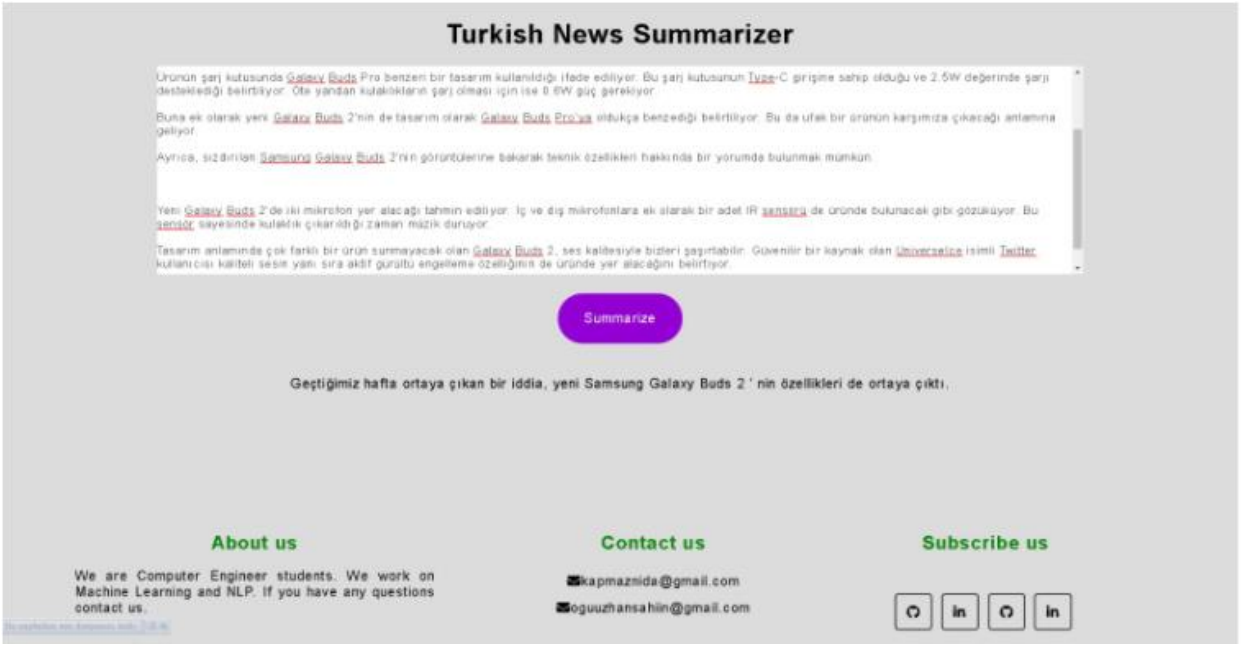

Figure 4: Use of the website with text.

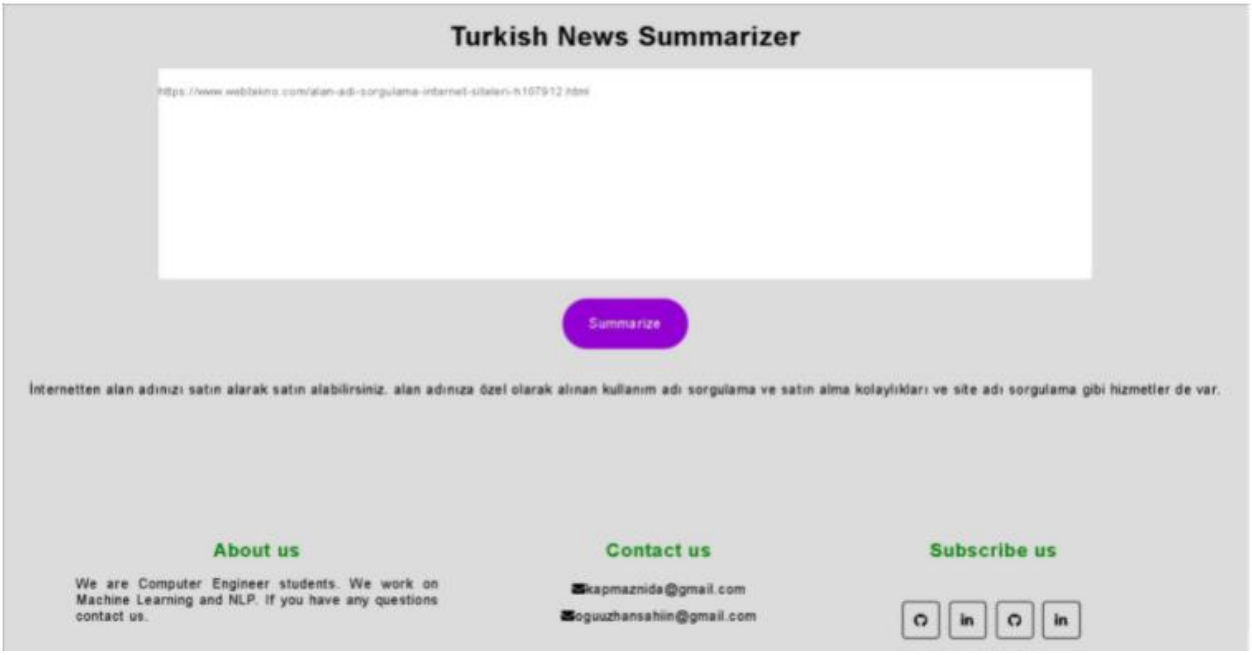

Figure 5: Use of the website with URL.

8 Future Works

- Since BERT max input size 512, news that more than 512 tokens will be truncated to the 512. That's why, we can try different models.

- We created Flask application, but we did not deploy it. As a feature work, we'll deploy it in Heroku or Streamlit.

Oğuzhan Şahin 160709027

Nida Nur Kapmaz 160709014