School of Information Technology and Engineering at the ADA University

School of Engineering and Applied Science
at the George Washington University

Text Summarization for Azerbaijani Documents Using Hybrid Neural Networks

A Thesis
Presented to the Graduate Program of Computer Science and Data Analytics
of the School of Information Technology and Engineering
ADA University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computer Science and Data Analytics
ADA University

By
Mir Amir Pashayev

April, 2025

THESIS ACCEPTANCE

This Thesis by: Mir Amir Pashayev
Entitled: *Text Summarization for Azerbaijani Documents Using Hybrid Neural Networks*

has been approved as meeting the requirement for the Degree of Master of Science in Computer Science and Data Analytics of the School of Information Technology and Engineering, ADA University.

Approved:

Dr. Samir Rustamov	Rustamov S.	28.04.2025
(Adviser)		(Date)
Dr. Abzatdin Adamov		28.04.2025
(Program Director)		(Date)
Dr. Abzatdin Adamov		28.04.2025
(Dean)		(Date)

ABSTRACT

This thesis investigates efficient text summarization techniques for Azerbaijani documents through hybrid neural approaches, combining extractive and abstractive methods. Due to the low-resource nature of the Azerbaijani language, significant challenges arise in developing reliable summarization systems. To address this, an Azerbaijani-specific dataset was prepared, consisting of documents paired with human-written summaries, and both extractive and abstractive summarization models were developed and evaluated.

In the extractive summarization part, sentence embeddings were used to construct similarity matrices, followed by a TextRank-based algorithm to rank and select key sentences. Evaluation using ROUGE metrics demonstrated strong results, achieving ROUGE-1 recall, precision, and F1 scores of approximately 0.47, 0.52, and 0.49 respectively, ROUGE-2 scores around 0.44, 0.47, and 0.45, and ROUGE-L scores comparable to ROUGE-1. These results indicated a strong alignment between the extracted summaries and human references.

For the abstractive summarization task, the multilingual pre-trained mT5-base model was fine-tuned on the Azerbaijani dataset. Fine-tuning significantly improved performance over the baseline. The baseline (zero-shot) mT5 model achieved ROUGE-1, ROUGE-2, and ROUGE-L scores of approximately 45%, 25%, and 40%, respectively, with BLEU and METEOR scores around 35% and 42%. After fine-tuning, the model achieved ROUGE-1, ROUGE-2, and ROUGE-L F1 scores of approximately 64%, 47%, and 57%, with BLEU and METEOR scores improving to about 32% and 50%, respectively.

Visualizations of evaluation metrics, dataset length distributions, and comparative analysis were provided to better interpret model performance. Both extractive and abstractive systems showed significant promise for Azerbaijani text summarization, overcoming challenges related to data scarcity and linguistic complexity.

This work demonstrates that adapting multilingual pre-trained models and combining them with classical graph-based extractive methods can yield highly effective summarization systems for low-resource languages. Future research directions include expanding the dataset, exploring reinforcement learning techniques, and further optimizing model architectures for improved generalization across diverse Azerbaijani text domains.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Explanation
NLP	Natural Language Processing
NLU	Natural Language Understanding
mT5	Multilingual Text-to-Text Transfer Transformer
BERT	Bidirectional Encoder Representations from Transformers
BART	Bidirectional and Auto-Regressive Transformers
TF-IDF	Term Frequency-Inverse Document Frequency
METEOR	Metric for Evaluation of Translation with Explicit ORdering
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
BLEU	Bilingual Evaluation Understudy

1 INTRODUCTION

1.1 Definition of the Problem

Automatic text summarization is the task of summarizing a document or set of documents into a summary that preserves the main ideas and core information of the original content. Every day, people and companies receive overwhelming amounts of textual information in this very age of digitalization. Manually reading through all this data to gain useful insights has now become progressively harder. Summarization systems help tackle this challenge as it offers short summaries that can save a lot of time, increases the ability to access information, and helps the reader get the main content of the documents without needing to read them. Consequently, the emergence of effective automatic summarization has turned into a significant goal in natural language processing (NLP), owing to the burgeoning quantity of digital data in domains like journalism, scientific articles, and governmental communication.

There are two main methods for automatic text summarization: extractive and abstractive. Extractive summarization methods operate by selecting out the most relevant or significant sentences, restating the summary directly. The approaches rely on the fact that if certain parts of the original text can be on their own to represent the whole; a segment of text can be statically represented using certain specific characteristics or by using a graph-based algorithm to rank segments of text based on its relevance. Abstractive summarization models generate new content that conveys the main ideas of the source text. Unlike previous technique which utilizes exact sections of the content, an abstractive system paraphrases and compresses the source text and may use vocabulary and phrasing not present in the source. Function of abstractive summarization systems are closer to what human summaries do, but they are also more difficult to do because they need natural language generation capabilities and deep understanding of the material. Both approaches preserve the most salient information of the input, however they do so in a fundamentally different manner: one through selection, the other through generation.

For high-resource (e.g., English, Chinese) languages with sufficient training data and resources, automatic summarization has made great progress [2]. Researchers developed advanced models and techniques for these languages that produce fluent and informative summaries. Languages like Azerbaijani, on the other hand, are still relatively neglected. The Azerbaijani language is a clear example of a summarization gap: It is a major language with millions of speakers and a substantial digital presence, yet there are no good summarization tools available for it, unlike languages like English. One of the reasons for this discrepancy is the lack of sufficient large, annotated datasets and pretrained models for Azerbaijani. Most state-of-the-art summarization approaches use supervised learning on large corpora of example summaries to learn from – resources not available for Azerbaijani until recently (there is till some lack present). Hence, an state-of-the-art English summarization system cannot be reused to get the same performance on Azerbaijani, because they either have never been exposed to that language or they are unable to pick up language-specific features during training.

Azerbaijani not only faces issues related to scarce resources; it also presents linguistic challenges that shape the problem space for summarization. The Azerbaijani language is a Turkic language

spoken by more than 30 million people around the world, which has a rich morphology and a complex agglutination structure. Azerbaijani has a lot of suffixations and a syntax that differ greatly from the Indo-European languages. This implies that lexy/semantic phenomena, such as extensive inflection, vowel harmony, and free word order, need to be tackled by NLP approaches (such as summarization algorithms). Models not specifically trained on Azerbaijani should be challenged by these characteristics, which may result in errors, such as identifying essential parts of input and summarizing properly. In addition, the absence of NLP tools (such as high-quality tokenizers, part-of-speech taggers, semantic analyzers, etc.) in Azerbaijani further complicates the task of analyzing and summarizing its text. Hence, the main problem discussed in this thesis can be formalized as how to develop a lightweight system for automatic text summarization of Azerbaijani documents taking into account its low-resource language status and richness of linguistic characteristics? This is the main challenge to be addressed by the current research, which would be to investigate and propose methods that fill the gap between what is offered for high-resource languages and what is available for Azerbaijani.

1.2 Objective of the Study

This research aims to create a reliable and effective text summarization methods for Azerbaijani based on the aforementioned challenges with the use of modern hybrid neural network approaches and then qualify the results. Specifically, the thesis explores both extractive and abstractive summarization methods and explores them in detail. One of the main objectives is as such:

Build an Extractive Summarization System: Build an extractive summarization pipeline on the Azerbaijani language. The system will utilize neural network-based language representations, as well as, statistical features to interpret the input texts and choose the most important sentences. In particular, the system first leverages BERT embeddings (Bidirectional Encoder Representations from Transformers) to represent the semantic content of the sentences, and subsequently applies the TextRank algorithm (a graph-based ranking algorithm) to find and extract the top scoring sentences to form the summary. Further improvements, like rule-based post-processing may be applied to guarantee that the produced summaries are logical and not redundant.

Train an Abstractive Summarization model: Create a summarization model which can generate new summary sentences in Azerbaijani. The model is fine-tuned on a corpus of Azerbaijani documents and uses sequence-to-sequence transformer model. For this purpose, mT5 (Multilingual T5) is implemented - a transformer model pre-trained across multiple languages, including Azerbaijani. A pre-trained language model, mT5 is adapted to generate concise, fluent outputs that rephrase or synthesize the information from the source text, rather than simply copying sentences, through fine-tuning on a large, bespoke dataset of Azerbaijani text-summary pairs.

Find and preprocess a large-scale dataset: At the beginning of this work, there was not any big size dataset about Azerbaijani summarization, so one of the main goals was to find or create a useful dataset. After some research, the suitable dataset was found that was constructed on available sources. The dataset is a collection of a considerable amount of Azerbaijani articles, with their

corresponding human-written summary (more than 115000 document-summary pairs). The dataset has been thoroughly cleaned and preprocessed and is split into training, validation and test sets.
Evaluate Performance and Interpret Results: Rigorously evaluate the extractive and abstractive summarization system with the use of standard automatic evaluation metrics and analyze the performance. These metrics include ROUGE (Recall-Oriented Understudy for Gisting Evaluation) to allow for measuring overlap between the system-generated summary and a reference summary, along with complementary metrics like BLEU (Bilingual Evaluation Understudy) and METEOR (Metric for Evaluation of Translation with Explicit ORdering) to evaluate the quality of the generated text. Through comparison of the outcome of these two approaches, this study seeks to underline their advantages and disadvantages for the Azerbaijani context. This analysis will show what each type of summary captures in terms of relevant information, the linguistic quality of the summaries produced, and also when they fall short (e.g. if the abstractive model tends to miss details, or if the extractive model produces summaries that read badly/aren't contiguous).
By accomplishing these objectives, this study aims to show that advanced NLP models can be successfully utilized for the task of Azerbaijani text summarization. The goals of these tasks will lead to a pair of summary systems (extractive / abstractive) that embody the best-known methodologies for this low-resource language. Furthermore, by presenting both techniques side by side, the research can reflect upon whether a full combination, or "hybrid" approach could improve summarization performance even more. The objectives focus on developing summarizers but also cover insights from the summary pipelines that work best for Azerbaijani, thus paving the case for future objective measurements in the area.

1.3 Significance of the Problem

The implications of this study should be profound, on many levels: To begin with, automatic text summarization for Azerbaijani is of practical value for millions of Azerbaijani readers and speakers. In a world where most information is kept in form of digital text, being able to quickly obtain a summary in one's native language is a huge win in terms of information access and efficiency. Journalists, analysts, or average readers grappling with news articles about Azerbaijan, could use a summarization system to provide a rapid digesting of the key points of each article instead of reading each one in its entirety. This may also help save time and increase productivity in large workplaces or universities where high volumes of Azerbaijani Documents may need to be examined. Furthermore, this technology would enable better information sharing in Azerbaijani — important news or research results written in a long form could automatically be distilled in short form that content creators can use to share key updates, and also the audience can read the highlights to help them make up their mind on whether to read in full or skip reading. A suitable summarization tool will not only ease information overload, but will also reduce information latency, which should be directly beneficial for Azerbaijani end-users.
Beyond the immediate user-facing benefits, this work carries significance for the field of natural language processing, especially in the context of low-resource languages. Despite being a rich and complex language, Azerbaijani is underrepresented in Natural Language Processing (NLP) research