ISTANBUL COMMERCE UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

A GENERAL APPROACH FOR MEETING SUMMARIZATION FROM SPEECH TO EXTRACTIVE SUMMARIZATION

Neslihan AKAR

Supervisor
Assist. Prof. Dr. Metin TURAN

MASTER'S THESIS
DEPARTMENT OF COMPUTER ENGINEERING
ISTANBUL - 2022

ACCEPTANCE AND APPROVAL PAGE

On 28.06.2022 **Neslihan AKAR** successfully defended the thesis, entitled "**A General Approach for Meeting Summarization From Speech to Extractive Summarization**", which she prepared after fulfilling the requirements specified in the associated legislations, before the jury members whose signatures are listed below. This thesis is accepted as a **MASTER'S THESIS** by Istanbul Commerce University, Graduate School of Natural and Applied Sciences **Department of Computer Engineering**.

Approved by:

Supervisor	**Assist. Prof. Dr. Metin TURAN** Istabul Commerce University	.....
Jury Member	**Assist. Prof. Dr. Feyza Merve HAFIZOĞLU** Istabul Commerce University	.....
Jury Member	**Assist. Prof. Dr. Murat ORHUN** Istanbul Bilgi University	.....

Approval Date: 08.07.2022

Istanbul Commerce University, Graduate School of Natural and Applied Sciences, accordance with the 2nd article of the Board of Directors Decision dated 08.07.2022 and numbered 2022/351, "Neslihan AKAR" who has determined to fulfill the course load and thesis obligation was unanimously decided to graduated.

Prof. Dr. Necip ŞİMŞEK
Head of Graduate School of Natural and Applied Sciences

DECLARATION OF CONFORMITY
IN ACADEMIC AND ETHIC CODES

Istanbul Commerce University, Institute of Science, thesis writining accordance with the rules of this thesis study,

- I have obtained all the information and documents in the thesis under the academic rules,
- I give all information and results in a visual, auditory and written manner in accordance with scientific code of ethics,
- in the case that the works of others are used, I have found the works in accordance with the scientific norms,
- I show all of the works I have cited as a source,
- I have not tampered with the data used,
- And that did not present any part of this thesis as a dissertation study at this university or at another university

I declare.

08.07.2022

Neslihan AKAR

CONTENTS

Page

ABSTRACT

M.Sc. Thesis

A GENERAL APPROACH FOR MEETING SUMMARIZATION: FROM SPEECH TO EXTRACTIVE SUMMARIZATION

Neslihan AKAR

Istanbul Commerce University
Graduate School of Applied and Natural Sciences
Department of Computer Engineering

Supervisor: Assist. Prof. Dr. Metin TURAN

2022, 56 pages

Developing technologies and techniques have increased amount of information and easier access to information resources. However, due to the ever-growing amount of information sources, it has become difficult to access the needed information in a limited time. As a result of this situation, the need for summary information has become important. In this research, it is focused on creating inferential written summaries of communications that occur in oral environments such as meetings, lectures, conferences. However, since this type of problem requires conversion from audio to text, it also includes many issues such as the human factor, sound recording environments, and language-specific problems. In this study, it is aimed to take the audio recordings of the meetings, especially the IT sector, to process and summarize. Spontaneous conversations were converted into audio recordings and the obtained texts were summarized using extractive summarization techniques. The motivation of the study is to catch the important points that may escape the attention of the individuals at the meeting and to summarize the main agenda items for the personnel who could not attend the meeting. The experimentally generated dataset (converted from audio recordings to text) was summarized by 3 different person and compared with the summaries obtained from the developed inferential summative model. The results obtained are remarkable and it is seen that approximately 71% success has been achieved.

Keywords : Automatic extractive summarization, meeting summarization, speech recognition, speech summarization, speech to text, tf-idf.

ÖZET

Yüksek Lisans Tezi

TOPLANTI ÖZETLEMEYE GENEL YAKLAŞIM:
KONUŞMADAN ÇIKARIMSAL ÖZETLEME

Neslihan AKAR

İstanbul Ticaret Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr. Öğr. Üyesi Metin TURAN

2022, 56 sayfa

Gelişen teknolojiler ve teknikler bilgi kaynaklarının artmasını ve bilgi kaynaklarına erişimin kolaylaşmasını sağlamıştır. Bununla birlikte artan bilgi kaynaklarının sayısından ötürü, ihtiyaç duyulan bilgiye sınırlı zamanda erişim zorlaşmıştır. Gelinen bu durumun sonucu olarak, özet bilgiye duyulan ihtiyaç önemli bir hale gelmiştir. Bu çalışmada toplantı, ders anlatımları, konferanslar gibi sözlü ortamlarda oluşan iletişimlerin çıkarımsal yazılı özetlerinin oluşturulması üzerine durulmuştur. Fakat bu problem türü özellikle sesten metine dönüşüm gerektirdiği için, insan faktörü, ses kayıt ortamları, konuşulan dile özgü sorunlar gibi birçok konuyu da probleme dahil etmektedir. Bu çalışmada IT sektörü başta olmak üzere yapılan toplantıların ses kayıtlarının alınarak işlenip özetlenmesi amaçlanmıştır. Anlık konuşmalar ses kaydına dönüştürülmüş ve elde edilen metinler çıkarımsal özetleme teknikleri kullanılarak özetlenmiştir. Çalışmanın motivasyonunu, toplantıda bulunan bireylerin dikkatlerinden kaçabilecek önemli noktaları yakalamak ve toplantıya katılamayan personele ana gündem maddelerini özetlemek oluşturmaktadır. Deneysel oluşturulan veri seti (ses kayıtlarından metine dönüştürülmüş) 3 farklı kişi tarafından özetlenmiş, geliştirilen çıkarımsal özetleyici modelden elde edilen özetlerle kıyaslanmıştır. Elde edilen sonuçlar kayda değerdir ve yaklaşık %71 oranında başarının yakalandığı görülmektedir.

Anahtar Kelimeler: Konuşmadan metine, konuşma tanıma, konuşma özetleme, otomatik çıkarımsal özetleme, tf-idf, toplantı özetleme.

ACKNOWLEDGEMENTS

First of all, I would like to thank to my supervisor, Assist. Prof. Dr. Metin TURAN for helping me to overcome the difficulties I faced during my thesis study and also for his guidance and advices throughout this research.

I would like to thank my beloved spouse, Samet AKAR and my precious mother, Zöhre ARSLAN for moral and material support and for guiding me the most stressful times. And I would like to thank the sweetest of the world, my daughter, Esma AKAR for her patience during the dissertation phase.

Lastly , I would like to thank my all colleagues, especially Büşra ULUDAG, who helped me summarize the meeting texts of my thesis.

Neslihan AKAR
ISTANBUL, 2022

LIST OF FIGURES

Page

LIST OF TABLES

Page

SYMBOLS AND ABBREVIATIONS LIST

AI	Artificial Intelligence
API	Application Programming Interface
ASR	Automatic Speech Recognition
E2E	End-to-End
HMM	Hidden Markov Models
ICSI	International Computer Science Institute
IDF	Inverse Document Frequency
IE	Information Extraction
IR	Information Retrieval
IT	Information Technology
L	Length
LSTM	Long Short Term Memory
ML	Machine Learning
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NN	Noun (singular)
NNP	Proper Noun
NNS	Noun (plural)
POS	Part of Speech
QA	Question Answering
SA-ASR	Speaker Attributed – Automatic Speech Recognition
SDM	Single Distant Microphone
SEE	Summary Evaluation Environment
SOT	Serialized Output Training
TF	Term Frequency
UPI	United Press International
VB	Verb (base form)
VBD	Verb (past tense)
VBG	Verb (gerund)
VBN	Verb (past participle)
VBP	Verb (sing. present, non-3d)
VBZ	Verb (3rd person sing. present)
W	Word Ratio
WER	Word Error Rate

1. INTRODUCTION

Information is the resolution of uncertainty as mentioned in Shannon's information theory (1940). Human wants to remove or decrease uncertainty because uncertainty leads to misconception. Humanity has been producing information since its first existence because it is a necessary phenomenon for the continuation of vitality.

Today, increasing communication resources provide faster and more effective spreading of information. However, the increase in the speed of information spreading, by the help of developing technological tools, it becomes uncontrollable. Increasing information sources have caused information pollution and exposure to unnecessary information. Always scanning huge archives to reach the necessary information is a waste of time and energy. The inner motivation to use energy in the most efficient way has increased the importance of well-summarized text and even the ability to summarize. With the development of natural language processing techniques and tools, the studies in this field have been increasing lately. Summarizing written texts, oral sources and videos are now possible with current technologies (Borji et al., 2022).

Managing information has become important in all areas of life. Especially in professional life, it has great importance. Even in small organizations, the technical information, know how and meeting notes accumulated over time, can turn into an unmanageable, jumbled set of information. If there are problems to be solved or for a process that needs to be managed well, managers want to know what is at the core of the business and want offer point solutions. This situation is experienced commonly, especially in companies where meetings are held continuously and included in the decision-making process these meetings. Extended meeting hours and consecutive meetings actually consist of irregular sets of information, many consist of repetition. And real meeting has contain complex and overlapped voices, interrupted or abandoned utterences (Ang et al., 2004). The ability to automatically summarize meetings and to extract important

(key) information can greatly increase the efficiency of many areas of business (Huebner et al., 2021). The motivation of this study is to take advantage of the developing technical possibilities for speech summarization.

In most organizaions, staff and managers spend many hours each week in meetings. Developing techniques and technological tools have allowed meeting data be recorded and stored routinely. As a result, automated tools to transcribe meeting will greatly increase the productivity of both meeting attendees and non-attendees. The meeting domain has a various subdomains include conferences, seminars, symposia, board meeting and etc. All these types of meetings can greatly get benefit from the development of automatic speech recognition (ASR), comprehension and information extraction technologies.

Depending on the type of the text to be summarized, the approaches will be very different. If a method to summarize scientific article is used while summarization an interview, it will not be efficient enough. The benefit to be obtained from the summary of an article is not the same as the benefit to be obtained from the summary of an interview. For this reason, customizations will always be made according to the application area. In this thesis, the meeting in the IT sector were targeted, so the process of the collecting the voices to be summarized and determining the procedures to be applied led to privatization. The dictionary (Roesler, 2020) has used to make sure that the sentences to be chosen are the best candidates has greatly increased our success rate.

According to the a scoping review report pubslihed by Berkovssky et al. (2020) for automatic speech summarization, most studies focused on summarising broadcast news (43 studies) and meetings (41 studies) in the total 110 studies published between 2002 and 2018.

Meeting summarization is separeted from a normal one person speech summarization, since the conversations are carried out mutually, many people will have a voice. As a consequences of many effects such as acoustics and intonation of the voice, problems will be encountered while making speech to

text. In addition language specific issues, if the speaker is native or foreign, the perception of the language will be interrupted (Ang et al., 2004). Problems such as the sensitivity of the devices used to collect sound in meeting environments are among the problems that need to be solved.

The purpose of the meeting summary is to cover the important points and to create the main agenda items for a personnel that cannot attend the meeting. Also reviewing the meeting notes before the next meeting for successive meetings can provide quick recall for better planning and preparation. In general studies, repetitions and emphases have been given importance and extractive summary algorithms have been developed. On the other hand focusing only on repetitive information is not the right approach. As Shannon mentioned in the theory of information (1940), less seen information may be more important. So many experiments have been experimented by changing the weights of the words that occur less frequently. For example, Baxendale emphasized in a study that the first and last sentences are the most important (1969). This means increasing the weights of the first and last sentences of the all sentences in a text. One more example is Dejong's work. He obtained successful results by identifying keywords while filling out the template he used in his knowledge-based work. By increasing the weights of the keywords, it enabled the program to make more accurate choices (1979-1982). Since this study aims to summarize the meetings in the IT sector, the "A Computer Science Academic Vocabulary List" dictionary that was produced by Rosler as a result of his thesis output was used to. Experimentally generated meetings text dataset (converted from audio recordings to text) were summarized by 3 different person at the rates of 40% and 20%. The experiments showed that program extracted summaries have a sentence similarity success rate of approximately 71% compared to human summaries.

Since this thesis purpose is to summarize the meetings in the IT sector, the dictionary containing the words commonly used in computer engineering has been used. This dictionary was created as a result of a thesis study by scanning many academic sources in the field of computer engineering. If it is intended to summarize meetings in another topic, specific dictionary should be created

related to the topic. Keywords can be extracted by taking meeting audio recordings, all belonging to the same topic, and processing them. By using these keywords as a dictionary, it seems possible to extract summaries of other meeting audio recordings belonging to the same topic.

In short, although this thesis was aimed at the IT sector, it has brought a new approach to meeting summary with the help of the dictionary specific approach used and proposing sentence scoring accordingly. This approach can be extensible to any subject. The results obtained from experiments show us that it is an applicable model.

2. LITERATURE REVIEW

In literature review section, important summary studies and their details are given.

2.1. Text Summarization

The first text summarization study was started in the 1950s, and basic terms used today were published. The concept of "term frequency (TF)", which was introduced by Hans Peter Luhn in a study in 1957, is still the most important part of text summarization and is the most used weight measurement scheme (Bael et al., 2016)

In his study, Luhn proposed the idea that "the more a term or phrase is appeared in the text, then it is the more importance term (1957)" and developed the concept of "term frequency". Tf value is the ratio of the number of target terms in the document to the total number of terms in the document.

Another pioneering work on text summarization was published by Jones (1972). Jones introduced the concept of Inverse Document Frequency (IDF). It proposed how generate normalized frequency list in adding to Luhn's method. The IDF value is the logarithm of the ratio of the total number of documents to the number of documents in which the target term occurs. At this stage, it does not matter how many times the term occurs in the document. It is calculated how many documents a word in IDF value occurs in. The methods of Luhn and Jones were frequently used together and called as TFxIDF in text summarization.

There are other methods have been tried for text summarization. Now lets have a look at some of them.

- A study by Baxendale was published on the importance of the position of the sentence in the text (1958). It emphasizes that the first and last sentences of the text are more valuable.

- According to Edmundson, the structure of the text was taken into consideration and the title or the first sentences were emphasized to be valuable. He tried to process clue words as well as position and frequency (1969).
- A study by Dejong, it was developed within the scope of artificial intelligence (AI) studies (1979-1982). The study called FRUMP is the first knowledge-based study. The news on United Press International (UPI) was processed with approximately 60 pre-prepared templates called "Sketchy Script" and the sentences extracted from the news were placed in these pre-prepared templates, and very succesful result were obtained.
- Paice (1990) conducted a study to eliminate the balance and harmony deficiencies of the summaries extracted from the text. He was interested in anaphore and rhetorical connections.

All these mentioned studies are only perspectives developed on summarization.

There are two different approaches as "Extractive" and "Abstractive" in order to generate the summary of the text in any language. In the abstractive method, new sentences are created with the rules of language (Mani and Maybury, 1999). In this approach , the rules about the sentence structure of the language should be known and applied. This method is developed and supported by AI and machine learning (ML).

On the other hand, the extractive method, which is the subject of this thesis, is simpler to use. It finds the sentences that best summarize the whole of the text. In this way, a small summary of the large text is provided. This method is not suitable for novels, stories and similar writings, but it is quite functional for news texts that contain semantic integrity, meeting minutes and similar texts (Garcia-Hernandez et al., 2008). Therefore, in this thesis, extractive summarization approach was used and key sentences of the meeting were tried to be selected through the provide dictionary.

2.2. Meeting Summarization

On the other hand, meeting summarization can include text summarization or it can be completely independent. There are two approaches in order to summarize a meeting. In the first approach, two-stages work is carried out. In the first stage, the sounds are collected and translated into writing. In the second stage, the obtained texts are summarized. Many problems arise when switching from speech to text. Because spontaneous speech is formless and very different from written text. Spontaneous speech often includes unnecessary information such as fluency, fillings, repetitions, corrections, and word fragments (Ang J. et al., 2004). Any researcher using real meeting datum will largely have trouble transitioning from speech to text. And there will be mistranslations that will cause loss of meaning. Because many people attend to the meeting and there exist affecting factors such as interference, talking at the same time, malfunctions caused by voice recorders, and noise. According to the report published in the research with the ICSI corpus, 19 utterances other than 9 were observed to be missing due to interruption by another speaker, and high speaker overlap was observed in many regions (Ang J. et al., 2004). ICSI corpus has over 180,000 handannotated dialog act tags and accompanying for roughly 72 hours of speech from 75 naturally-occurring meetings. Their aim is provide a brief summary of the annotation systems, labeling procedure, inter-annotater reliability statistics, overall distributional statistics.

2.3. Speech Summarization

Many text-based and speech-based features have been proposed in speech summarization systems to summarize different speech datum. Hori and Furui studied automatic speech summarization based on word significance and linguistic likelihood. Their aim was to summarize Japanese broadcast news. They focused on a method of finding important words using the number of characters relative to the target compression ratio (Hori and Furui, 2000). For speech summarization, it is possible to benefit not only from the lexical features of words, but also from their acoustic features. Because speech recordings contain

intonation unlike written language. It offers us the opportunity to extract the most important sections according to this intonation (Zhang and Yuan, 2013). For example, Zhang et al. conducted a study comparing two types of speech, Mandarin Broadcast News and Cantonese Parliament Speech, using acoustic/prosodic, linguistic and structural features for speech summarization (Zhang and Yuan, 2014). Having tried many approaches to improve query-based meeting summarization performance, Huebner et al. have expanded their approach to locate-then-summarize approach of QMSum (Awadallah et al., 2021). And they also investigated the effectiveness of pre-training the model with a large news summary dataset using HMNet (Zhu et al., 2020). Finally, they validated their approach by comparing the performance of their model with BART, a state-of-the-art language model that is efficient for summarizing. They observed improved performance by using clustering methods to extract key information (Huebner et al., 2021).

In the second approach, speech summary can be obtained from the direct speech. The techniques applied in speech-to-speech automatic summarization are different. Original spoken sentences are analyzed separately as words and filler units (Furui et al., 2004). It is aimed to extract important information by processing the audio recordings taken instantaneously and to make them more meaningful by reducing the repetitive sentences and concepts to one in extended meetings. In a study, Baykal et al. benefited from the repetitive features of sounds (2008). Repetitive structures like chorus in musical audio can be observed in meetings. Accurate detection of repetitions can improve performance in automatic speech summarization. In the presented method, it transforms the 1-D time-domain speech signal into a 2-D image representation, a (dis)similarity matrix. Detects possible repetitions in the matrix using appropriate computer vision techniques. According to the results obtained, speech signals catch the key-concept (or topic) and computational analysis performs well. Also, the method does not convert the speech signal into words, phrases or sentences. Therefore, it can be generalized as a speech-to-speech summarization method, in which summarizing results are presented by speech rather than text. Also, it does not need any prior knowledge of language and grammar. In addition, it is possible to

benefit not only from the spoken, but also from the video features. For example, Li et al. developed an abstractive meeting summarizer from both video and audios of meeting records. They proposed a multi-modal hierarchical attention mechanism across three levels: topic segment, utterance and word. They took advantage of speaker interaction and participant feedback to discover salient utterances (Li et al., 2019).

2.4. Commercial Tools

In addition to all these academic studies, commercially developed automated tools are available. Thanks to the developed automatic tools, meeting datum can be recorded and stored. They help to increase, the productivity of both meeting attendees and non-participants. Moreover, they allow evaluation of all meeting content by an independent third eye. One of them is the meeting recognition and learning assistant, called Calo, which was developed for this purpose. The audio stream from each meeting participant is transcribed to the text using two separate recognition systems. The first recognition is achieved when creating live copies with a real-time recognizer. The other recognition is applied when the meeting ends, another transcript is generated, it is offline recognition system. The results are obtained by using the prosodic and contextual information of the obtained texts (Tur et al., 2008). The other example is MeetingVis, which summarizes not only what is spoken, but also the materials (presentation slides) used, has added as another dimension to this issue. MeetingVis introduced a visual narrative approach to summarization (Bhamidipati et al., 2018).

For another example, Watson speech to text commercial application developed by Watson IBM offers many advantages to its customers with adapted speech models in the field of customer service. With language and acoustic training options, personalized scenarios and detailed search features, it increases the accuracy in speech recognition. It offers models optimized for low latency in real-time speech applications. It analyzes and corrects weak audio signals. In addition, its smart formatting feature allows conversion of dates, times, numbers, currency values, e-mail and website addresses into traditional formats. It understands who

is saying what in a voice exchange with multiple participants. It is optimized for two-way call center chats but can detect 6 different speakers. It can filter out inappropriate content or specific words using keyword detection and profanity/corrupt language filtering features (Watson IBM, 2022).

Speech recognition can be quite challenging in environments with multiple conversations. As an example is given, if overlapping sounds are translated directly to text, data loss will occur. That's why the correct collection of voices is one of the most important process of the meeting summarization. Developing technological innovations enable us to correct this voices. For example in a study by Kanda et al. (2021) focused on counting the speakers in the meeting, speech recognition, and speaker identification. In this study, a speaker deduplication mechanism was proposed to reduce speaker identification errors in highly overlapping regions. Experimentally, the results obtained on the LibriSpeechMix dataset showed that the transformer-based architecture was especially good at counting speakers, and proposed model reduced the speaker-attributed word error rate by 47% over the LSTM-based baseline.

In another study, a new approach was presented for neural speaker diarization. Its called Transcribe-to-Diarize. It uses an end-to-end (E2E) speaker-attributed automatic speech recognition (SA-ASR). The proposed model counts speaker, multi-talker speech recognition, and identification of speaker from monaural audio that contains overlapping speech. Also this model does not predict any time-related information, the start and end times of each word can been predicted with sufficient accuracy (Chen et al., 2022).

Using only a single distant microphone (SDM) to collect sounds, transcribing meetings containing overlapping speech has been one of the most challenging problems for automatic speech recognition (ASR). In the two-step approach developed by Chen et al., (2021) serialized output training (SOT) was first investigated using large-scale simulation data. In the second step, they fine-tuned a small amount of actual meeting. Experiments were conducted using 75,000 hours of internal single speaker recording to simulate a total of 900,000 hours of

multi-speaker audio segments for supervised pre-training. With 70 hours of the AMI-SDM training data , proposed SOT ASR model achieved a word error rate (WER) of 21.2% for evaluation of AMI-SDM while speakers counted automatically in each test segment.

Involvement hot spot concept is the one of the most useful concept and have been studied on over 15 years for meeting analysis. As judged by human annotators, these are regions of meetings that are marked by high participant involvement. Dimitriadi et al., (2020) focused on this concept and explored the extent to which various acoustic, linguistic and pragmatic aspects can assist in hot spot detection. In this context, they used the openSMILE toolkit to extract features based on acoustic-prosodic cues. They used BERT word embeddings to encode lexical content and various statistics based on speech activity to describe them. They conducted experiments on the ICSI meeting corpus. They found that the lexical model was most informative with contributions from the acoustic-prosodic model components.

3. METHODOLOGY

In this section, detailed information about meeting summary studies, types, definitions and processes is given. Useful information about speech to text conversation, text summarization, detailed information about NLP's structure and work, extractive summarization, evaluation metrics used in this thesis, and application details are given.

It is an important requirement to examine and understand these concepts at the beginning of the meeting summary process.

3.1. Speech to Text

Speech recognition is an interdisciplinary subfield of computer science and computational linguistics that develops technologies in order to enable computer recognition and translation of spoken language into text. Automatic speech recognition (ASR) is also known as computer speech recognition or speech to text (STT). It combines knowledge and research in computer science, linguistics and computer engineering.

There are two main types as “speaker-independent” and “speaker dependent”. In speaker dependent systems require “training” (alias “enrollment”) where text or isolated vocabulary is read by an individual speakers into the system. In this way, a person’s specific voice analyzed by the system is used to fine-tune and results in increasing accuracy. And speaker-independent (Fifthgen, 2013) systems do not use training.

Speech to text software is a type of software that effectively takes audio content and converts it into written words in a word processor or other display target. This type of speech recognition software is invaluable and very useful for anyone who needs to create a lot of written content without a lot of handwriting. Speech to text applications are used to produce text messages from voice (exp. for the visually impaired person or to speed up typing), unlocking the key of the

smartphone, chatbox applications, bank account login, smart assistant, etc. to create written texts.

Speech recognition has a long history with major innovations. Recently, it has benefited greatly from advances in deep learning and big data. The developments have become more popular not only with the increase of academic papers published in this field, but also with the use of deep learning methods in designing and deploying speech recognition systems of the worldwide industry.

Key concepts for speech recognition systems growth were: vocabulary size, speaker independence, and processing speed.

Let's take a brief look at the evolution of speech recognition and its historical milestones. In 1952, a number recognition system named “Audrey” was built for single speaker number recognition by Balashek et al (Juang and Rabiner, 2017). The speaking ability of the 16-word "Shoebox" machine produced by IBM in 1962 was introduced to the whole world (Pinola, 2011). In the late1960s, sustained speech was adopted by some scholars and found suitable for study (Reddy, 1960-1969). Previous systems required users to pause after each word.

In the mid-1980s, IBM's Fred Jelinek's team created a voice-activated typewriter called Tangora that could handle a vocabulary of 20,000 words. The N-gram language model is introduced. The acceleration of the progress in this field was increasing in parallel with the increasing capabilities of computers. Because the processors of the computers used to process data at that time were very slow and insufficient. In 1992, Voice Recognition Call Processing service was started to be used to route AT&T phone calls without using a human operator (Juang and Rabiner, 2017).

The work progressed and took its present form. Recordings taken through GOOG-41, a phone-based directory service, produced data that helped Google improve its recognition systems. Google voice search is now supported in over 30 languages.

In the United States, the National Security Agency has been using speech recognition systems for keyword detection since at least 2006 (Froomkin, 2005). Thanks to this technology, a large number of recorded conversations can be searched. It can run database queries using keywords and distinguish interesting conversations from others. Some government research programs have focused on speech recognition technology in intelligence applications. DARPA's EARS program and IARPA's Babel program are examples of these applications.

In the early 2000s, speech recognition was still dominated by traditional approaches such as Hidden Markov Models combined with feedforward artifical neural networks (Bourlard and Morgan, 1994). However, today it is maintained by a deep learning method called Long short-term memory (LSTM), a recurrent neural network published by Sepp Hochreiter & Jürgen Schmidhuber (1997).

Microsoft has pioneered developments in this field and achieved significant success in converting telephone conversations to text in its historical turning point, Switchboard (2017). It used multiple deep learning models to optimize speech recognition accuracy. The success rate of this model is remarkable and has inspired studies on this subject.

In this thesis, the audio files of the meetings were tried to be converted into text by using the Speech Recognition API offered by the Python. But it seemed insufficient the transcription process was completed and we decided using the Speech to Text Cloud API offered by Google.

3.1.1. Speech to text tools

Speech to text tools are speech recognition software that enables spoken language to be recognized and translated into text through computational linguistics. It is also known as speech recognition or computer speech recognition. There are many paid and free tools available in this domain. These tools can copy audio streams in real time or offline to scan and generate text. For

examples: Speech to Text Converter by Amberscript, Watson Speech to Text by IBM, Transcribe by Amazon, Automatic Speech Recognition by Google Cloud are its predecessors.

There are applications for only voice recognition, but since the processing of voice recordings taken from a meeting environment in this thesis study, these applications will be insufficient for detecting conversations, preventing noise, etc. We used applications with AI support to minimize the problems. Now, the products supported in the thesis work will be mentioned.

3.1.1.1. Speech recognition api

The Speech Recognition Library, offered free of charge by the Python, converts a sound file or the sound it receives from the microphone into text. This speech recognition library cannot transcribe real-time words. It waits for the sentence to end in order to detect a speech, and the recorded audio file is sent to the free Google servers and a response is received.

In order to use the library, it must be installed and imported. Microphone and recognizer must be defined in the included package. The threshold value is defined to prevent noises in the system, which is optimized for background sounds. It is possible to listen continuously, as well as to wait for a certain time by assigning a treshold value. Although this library is seen as an attractive option because it is free and easy to implement, it contains many problems. It is not possible to make speech to text using this library in environments where many people speak. Even when we tried it for two people, the threshold value given according to the results we received, it perceived the thin sound as noise and ignored it. Its accuracy is open to debate, as the purpose is to collect the voices in the meeting room. For this reason, a more advanced tool was preferred and the speech-to-text tool offered by Google Cloud was used.

3.1.1.2. Google cloud speech to text api

Cloud speech-to-text convert speech into text using API powered by Google's AI technologies. It is a paid tool, which is advantageous in many ways. It can handle both live streaming and pre-recorded audio, as well as handle audio in noisy environments. For sounds longer than one minute, the uri space in Google Cloud Storage should be used; It is charged at $0.02 per GB per month (Choi et al., 2018). However, it offers the possibility of converting audio files to text, with a daily limit 60 minutes for 3 months free of charge. Google Cloud Speech API has the best recognition rate in standard language (Lee and Roh, 2017).

This commercial product encourages its customers to provide a better user experience for their products through voice commands. It provides usage areas such as converting content to text with subtitles and correct customer interactions for companies that want to improve service quality. Cloud speech-to-text features:

- It is a global application with support of more than 125 languages.
- It can process real-time voice or voices from a prerecorded audio file. It can store audio files via Cloud Storage or inline.
- Speech recognition is customized, copying domain-specific terms and rare words and enhancing transcription of phrases.
- Automatically converts spoken numbers to addresses, years, currency and more through the classes it provides.
- Allows you to have control over your data within the company, thanks to the speech recognition technology it uses.
- With the multi-channel recognition feature, it can recognize different channels, keeping the order and annotating the transcripts.
- It can handle the noisy sound coming from the environment without any additional adjustment.
- Voice control optimized for specific quality requirements and a choice of trained models for phone call and video transcription.

- It offers the ability to filter texts to detect inappropriate or unprofessional content with its content filtering feature.
- You can make Transcribe reviews. By uploading your own audio file and its text, you can evaluate its quality by repeating your own configuration.
- Automatically inserts punctuation marks (beta).
- Provides the ability to find out who said what by getting automatic predictions about the phrases spoken by speakers in a conversation (beta) (Google Cloud Speech-to-Text, 2021).

With the video transcription model, it is possible to transcribe video and/or multi-speaker content or to add subtitles. The use case diagram of this model is given in Figure 3.1. It uses ML technology similar to the video subtitle on YouTube.

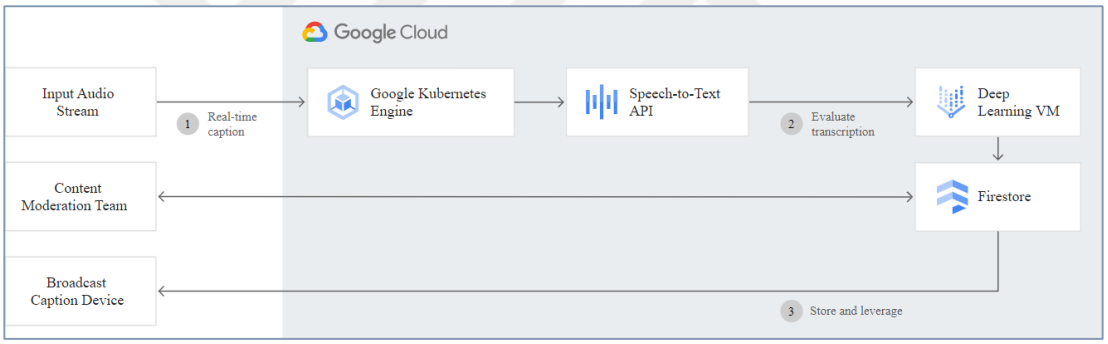


Figure 3.1. Google cloud speech to text use case diagram (Google cloud speech-to-text, 2021)

In this thesis, the Google Cloud Speech to text tool was used, which has a very high performance. The code was generated using Visual Studio Code, and the audio files given as input , produced as output the text files with the ".txt" extention.

3.2. Text Summarization

Definition of summary is in Cambridge dictionary "a short, clear description that gives the main facts or ideas about something (Cambrigde Dictionary, 2022)". This is the most general definition.

Today, data is produced and consumed very quickly. As the work in the field of big data increases, it is an important issue to divide the data into easy-to-manage small pieces for easy storage. In terms of data mining, the definition of “Summary” has different variations, such as compression, tagging, association, etc. All of the methods can be evaluated in summary (Mirkin, 2011).

For text summarization studies, the summary is defined more clearly: “Summary is to transform the source text into a shorter one while preserving the information content and general meaning (Gupta & Lehal, 2010)”. There are also some rules in the definition of "summary" for text summarization studies. For example, for a text to be accepted as a summary, it should not exceed half the length of the source text to which it belongs to (Hovy et al., 2002).

3.2.1. Automatic text summarization

In its most general definition, text summarization is the process of converting large documents into shorter and more precise paragraphs or sentences. The most useful information is obtained without losing the meaning of the text, while the irrelevant or less important ones are eliminated. Written texts such as web pages, news articles, blogs, etc. in the form of digital documents are rapidly turning into large amounts of data. There is a great need to reduce most of this text data into shorter, focused summaries that capture salient details so that they can both be reviewed more effectively and checked to see if they contain the information sought in large documents. It is not possible to create all text summaries manually; With the development of technology, automatic methods have been applied and the need for automatic summarizers has increased. Let's list the benefits of automatic summarization:

- Summaries shorten reading time.
- For someone searching documents, summaries make the selection process easier.
- Automatic summarization improves indexing efficiency.

- Automatic summarization algorithms are less biased than human summarizers.
- Personalized summaries are useful in the question answering system, thanks to the specific information they provide.
- It increases the number of texts that commercial summary services can process (Manuel and Moreno, 2014).

When a written document is to be summarized, it is easy for a person to understand what is important and what is not. In order for them to create summary texts just like humans, appropriate coding and programs must be added to the machines. The process of summarizing text with ML or AI programs and the others are known as automatic summarization.

However, automatic text summarization has its challenges. The first problem is to select the appropriate information from the main document. After that, the summarizer should present the final summary to the reader in a friendly manner. Automated text summarizers aim to overcome these challenges and optimize topic coverage and readability.

3.2.2. Text summarization architecture

Text summarization studies have the same architectural structure, although different methods are used. There are many examples in daily life. In his book "Advances in Automatic Text Summarization" (1999), the list of everyday uses for summarizing text is quite extensive:

- headlines (from around the world)
- outlines (notes for students)
- minutes (of a meeting)
- previews (of movies)
- synopses (soap opera listings)
- reviews (of a book, CD, movie, etc.)
- digests (TV guide)

- biography (resumes, obituaries)
- abridgments (Shakespeare for children)
- bulletins (weather forecasts/stock market reports)
- sound bites (politicians on a current issue)
- histories (chronologies of salient events) etc.

Different criteria and processes come into play to teach the machine the evaluation filter, which varies even from person to person in text summarization.

Document(s): The source text to be summarized can be a single document or related or unrelated documents.
Compression: It is the ratio of the summary to the source text to be extracted. It can be longer for single documents (about 30%), while for multiple documents it may need much more compression (about 5%).

Function: The purpose of the summary is the most important fact. Since the purpose of summarizing a news text and a scientific publication is different, an informative summary should contain all the lines of the subject, while a descriptive summary can only contain the main lines.

Fluency: It is about the sequence of sentences and paragraphs in the summary. Compliance is expected during a stream in an appropriate summary.

3.2.3. Types (categories) of text summary

Text summarization studies can be done for different purposes as well as in different ways. When categorized according to the methods used and the desired outputs, it can be divided into the following headings in general terms:

By source type:

- Single Document: It is aimed to summarize a single text.
- Multi Document: It is aimed to summarize more than one document with or without a semantic unity.

According to the method: Summaries based on the method used:

- Extractive: Summing up by selecting the sentences that best represent the whole text from the source text.
- Abstractive: After examining the subject of the source text, the summary is made with new sentences created by the system. It is similar to the human abstract.

By purpose: Depending on what purpose the produced summary will be used for:

- Indicative: Summaries created by taking the important parts of the subject from the source text. All information in the source may not be included in the summary.
- Informative: It is intended to convey all the information in the source text. It is preferred for summaries of scientific studies.

According to the content of the conclusion: It is created depending on what the summaries to be drawn will contain:

- Domain: Result summaries for a domain.
- Genre Specific: Summarizing studies for only one subject.
- Independent: Summarizing studies for the text regardless of the field or subject.

By query: Depending on how the generated summaries will be queried, displayed:

- Query Based: Creating summaries that match the query made. It is a summary format generally used by search engines.
- Generalized: Displaying generalized summaries without a query (Özkan, 2019).

In this thesis study, the extractive summarization method was used. In this section, information about the details of extractive summarization will be given.

3.2.4. Extraction summarization

Extractive summarization takes sentences directly from the document based on the scoring system to create a compatible summary. This method works by identifying the important parts of the text and cutting it and combining the sections to get the summary text.

They allow extracting sentences from the original text. It can take a single document or multiple documents as input. The flow of this summarization system is typically as follows:

1. Creates an intermediate representation of the intro text to find featured content. Calculates TF metrics for each sentence in the given matrix.
2. Calculates the total scores of the sentences and assigns each sentence a value indicating the probability of being summarized.
3. Sorts all sentences, starting with the most important (Gonçalves, 2020).
4. It finds the number of sentences to be summarized according to the percentage value given and writes that many sentences.

In this section, we needs to give some information basic concepts in extractive summarization. In order to explain extractive summarization in detail, we should tell on the concept called TF – IDF. TF stands for term frequency – IDF stands for statistical values known as inverse document frequecy. They are numerical values that are intended to reflect how important a word is in a document (Rajaraman and Ullman, 2011). This value is a weighting factor often used in information retrieval, text mining, and user modeling searches. The Tf-Idf value increases proportionally to the number of times a word appears in the document and is balanced by the number of documents containing the word. It helps fix the fact that words appear more often in general. It is one of the most popular term weighting schemes. A 2015 study showed that 83% of text-based recommendation systems in digital libraries use tf-idf (Breitinger et al., 2015).

The usage areas are very many. It can be used successfully to filter stop words in many areas, especially in text summarization and classification. For example, search engines use a central tool in scoring and ranking user queries for the relevance of the given document (tf-idf wikipedia, 2022).

3.2.4.1. Term frequency

We have the keywords we want to search and the texts to search. If we want to find which of these written texts are more related to this keyword, the first thing we will do is to eliminate the documents that do not contain these keywords. But we may still have many texts left. As a second filtering method, we can use the text with the most keyword is most relevant approach. There are many ways to determine this. We will propose the TF-IDF method as a solution to this problem and explain how it is applied.

Term frequency is a formula intended to identify the importance of a keyword or phrase in a document or web page. Term frequency, $tf(t,d)$, is the relative frequency of term t within document d,

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{3.1}$$

where $f_{t,d}$ is the raw count of a term in a document, the number of times that term t occurs in document d. And you can find the exact form in Equation (3.1).

3.2.4.2. Inverse document frequency

The inverse document frequency metric is a measure of how much information the selected word provides. It looks for common or rare in all documents. It is obtained by dividing the total number of documents by the number of documents containing the term and taking the logarithm of that section. In short, it is the logarithmically scaled inverse fraction of documents containing the word. It is expressed as in the Equation (3.2).

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \quad (3.2)$$

N: total number of documents in the corpus $N = |D|$

$|\{d \in D : t \in d\}|$: number of documents where the term t appears.

If the term does not exist in the corpus, adding 1 to the denominator does not allow the value in the denominator to be 0.

The tf-idf value is obtained by multiplying the found term frequency and the reverse document frequencies (Equation 3.3).

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \quad (3.3)$$

3.3. Meeting Summarization

3.3.1. What is meeting?

A meeting define as two or more people coming together to achieve a common goal. The porposes of meeting are sharing information, reaching agreement etc (Wikipedia Meeting, 2022). Developing technical opportunities offer online or face-to-face meeting opportunities. Meetings can be held through tools such as Skype call, phone call, zoom. In addition, it is the most preferred way for the participants to have a meeting directly by sharing the same environment.

As the meetings are held for different purposes, the form of the meeting also varies. Meetings where managerial decisions are made or in-team meetings or meetings between management and staff, etc. The characteristics of each meeting are different. There is a manager or someone in a higher role running the meeting. Other meeting members take turns taking the floor and have a say in the meeting.

Since the meetings held in the field of Information Technologies are aimed in this thesis, it would be appropriate to examine its specific features. IT employees determine meeting times and intervals according to the way they work. For example, a team working in agile methodology does a daily scrum standing up at the same time every morning. In Scrum meetings, team members make a brief

statement about what I did yesterday, what I am doing now, what I will do today, any problems. And after all team members have finished this process, what should be done in order to solve these problems is reviewed and the daily scrum is completed in a way that does not exceed 15 minutes. An example of another meeting type, the product requirements meeting with the customer can be given. In this type of meeting, the customer tells the requirements of the product he wants to be developed. This shows how the basic requirement dataset is prepared. The language that customers generally use is the requests. It can often use the imperative mood by saying "must". In addition, the verb "want" is used commonly when expressing wishes.

It is possible to work on these meetings by using the dictionaries or templates prepared by emphasizing the characteristic features of the speeches. This area contains so many factors that it is a thesis topic on its own. . The study called FRUMP by Dejong on this subject is one of the best exemplary studies (1979). He obtained informative document by filling in the drafts he used with the help of AI.

Meeting summarization has many problems, unlike a normal speech summarization. For example, even though everyone is given the right to speak in the meetings in turn, problems such as noisy voices coming from the back or one person entering the conversation before the end of the conversation cannot be prevented. It is very difficult to record sound in such an environment. In addition, the sound collection devices used should be increased according to the size of the meeting room and the number of participants. One microphone may not always be enough. Or, problems such as a member's low voice and whether the language he/she uses is his/her mother tongue can make it difficult to perceive his/her voice. The solution of all these problems is big enough to be a separate dissertation study.

3.3.2. Automatic meeting summarization

After the meeting, which is held with a purpose, some new decisions can be taken, or a new working method can be determined, and solutions can be produced for

the problems raised. As a result, what was spoken during the entire meeting hour should be reported as a summary, otherwise all the resources (human, space, time etc.) spent will be wasted.

Our motivation for automatic summarization is that we want to find fast and accurate solutions to these problems, by taking advantage of the developing technical possibilities. In this thesis, it was studied by targeting the meetings held in the field of Information Technologies on automatic meeting summarization. With the automatic summarization study, it is aimed to obtain the main agenda items for both meeting participants and non-participants, to catch the issues that may be important even if they are less repeated, and to prepare the meeting minutes on a regular basis. In this way, time will be saved and automatic documents will be obtained without missing the focus of the meeting.

3.4. Natural Language Processing

3.4.1. What is natural language processing?

Natural language processing (NLP) is a sub-science of AI that deals with how to program the interaction between computers and human language to process and analyze natural language data (Wikipedia, 2022). The aim is to teach the structure of the languages used by humans to the computer, and to obtain computers that make sense of documents such as text and audio belonging to that language and understand their content.

NLP is a collection of computational techniques for automatic analysis and representation of human languages (Chowdhary, 2020). NLP started with the intersection of AI and linguistics. The methods that call NLP are algorithm-based science that emerged by imitating the human language of computers. The roots of NLP date back to the 1950s. In the article titled "Computer machines and Intelligence" published by Alan Turing, he proposed the criteria known as the Turing test (Chowdhary, 2020). The proposed test includes rules that enable

automatic interpretation and generation of natural language. The first example of NLP is demonstrated in this application.

Developments related to NLP were symbolic until 1990. The practice outlined by John Searle's China Room experiment works like this. Given a dataset with questions and matching answers, the computer emulates them. It returns the answer according to the question it encounters. The chatbot with Racter and Jabberwacky, are the one of works of this period. Developments in the field of NLP slowed down until 1990, but a work that led to a statistical return from the symbolic era was an important milestone. With the introduction of ML algorithms, the development of NLP accelerated (Guida and Mauri, 1986).

The first studies in this field were generally automatic translation studies. The studies carried out depended on specially developed corpora. With the growth of the web in the 2000s, studies started to produce more accurate results with the increase of the language content which can be used (Wikipedia, 2022).
By the 2010s, representational learning and deep neural network-style ML methods became widespread in NLP. NLP is used in many fields today. For example; It is gaining more and more importance in medicine and healthcare, where it analyzes notes and texts in electronic health records (Turchin et al., 2021).

3.4.2. Application of nlp

The speed and importance of information in the increase of internet resources has increased the need for NLP. Because www is a huge resource containing at least 20 billion pages and important information can be accessed through NLP. Some of the uses of the data are:

- Indexing and searching large texts,
- Information retrieval (IR),
- Classification of text into categories,
- Information extraction (IE),

- Automatic language translation,
- Automatic summarization of texts,
- Question-answering (QA),
- Knowledge acquisition,
- Text generations/dialogues (Chowdhary, 2020).

Automatic summarization of texts is the second part of this thesis. The audio recordings taken at the meetings were translated into text and then the text was summarized. Various models based on ML have been proposed for text summarization. These approaches model a classification problem that outputs whether a sentence should be included in the summary.

3.4.3. Tools for nlp

Many tools have been developed for use in scientific research and speech processing as open source or closed source. These tools can be written using scripts, It has functions to perform a series of tasks by calling them directly. For example, for NLP, they can tokenize the given text, splitting root and POS (speech tagging), find the word frequencies in the given documents. The most widely used of these is NLTK. In following section, we will give general information about NLTK used in this dissertation.

3.4.3.1. Python and the natural language toolkit (nltk)

NLTK (Natural Language Toolkit) is a collection of Python libraries and programs for symbolic and statistical NLP. It is suitable for those who learn NLP, research in areas close to NLP, such as AI, ML. It has also been used as a teaching tool, as a prototyping platform for researchs and industry users. Python includes standard libraries with tools for speech processing, NLP, numerical processing, and graphics programming. Most importantly, it is an open source, free and community-oriented project (Chowdhary, 2020).

NLTK is a leading platform for building programs to work with human language data. It offers easy-to-use interfaces to many corporate and lexical resources, as well as performing classification, tokenization, resource creation, labeling, parsing; WordNet over 50+ resources (NLTK, 2022).

3.5. Evaluation Metrics

3.5.1. Text summarization success rate

In text summarization studies, evaluating the quality of the summary text is important. It is the stage in which it is determined how well the applied summarization methodology works. It develops in parallel with the studies in this field together with the text summarization studies. There is no single solution to this task. It has many parameters that needs to be considered in evaluation. The main lines of how this process should work have been clarified. There are some hurdles to be overcome in order to measure summary achievements (Jones and Galliers, 1996):

- Summary outputs are created by systems and they are in natural language flow. In some cases, although the summary obtained is the answer to the question sought, it may need to be better expressed.
- Since the summaries created will be evaluated by human beings, the developed methods may not be resource friendly and may cause waste of time and resources.
- Summarizing is related to compression. Evaluating summaries at different compression ratios can increase the complexity and size of the process.
- Since summaries are created for different purposes, criteria for what the purpose is must be included in the measurement process in order to determine its success, this evaluation can lead to complexity of process design.

Methods for evaluating text summarization studies and NLP studies can be grouped under two main headings: “Intrinsic” and “Extrinsic” evaluation methods in Figure 3.2.

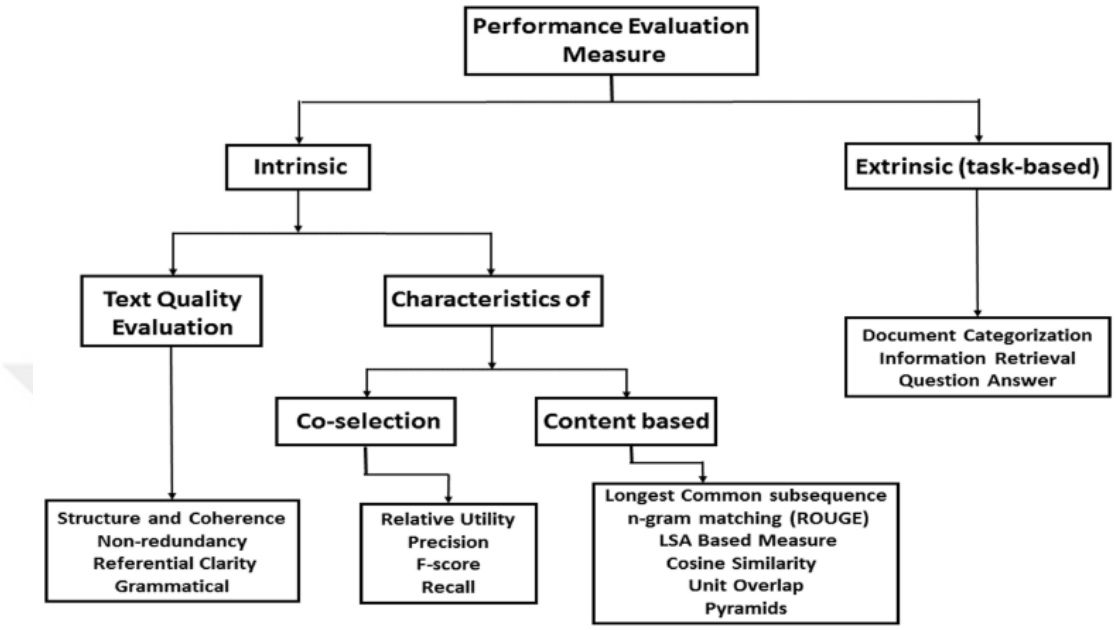


Figure 3.2. Classification of performance evaluation measure (Bharti and Babu, 2017)

Intrinsic Evaluation Methods: These methods provide the evaluation of the summaries obtained as outputs within themselves.

- Text Quality Evaluation: It is checked whether the information in the summary is useful and sufficient.
- Content Evaluation: The harmony of the sentences that make up the summary is checked.

Extrinsic Evaluation Methods: On the other hand, these methods try to measure the success of the summaries with their contribution to other tasks and processes they are related to. Is the classification correct? Is the information transfer sufficient? Does it respond to the request?.

Human summarizers are very important in evaluating the text summary outputs. In intrinsic evaluation, success is measured by golden summaries created by the

people. On the other hand, during the extrinsic evaluation, how successfully the summary completes its task can be evaluated with the modules of the system it is a part of. At the same time, "real" people can be included in this process.

In addition, it is difficult to obtain the same output even two different people summarize the same text. Because the important points may be different according to the person. Moreover, the quality of the language used by the person making the summary will differ in abstract summaries. When the evaluation is made by comparing the human summaries with the machine summaries, perhaps the summaries produced by the machine can be more beautiful and meaningful. Determining this is a difficult task.

Many tools and methods have been developed for measuring success. In this thesis, the outputs from 3 different human summarizers were compared with the machine summary outputs. A general conclusion was reached by taking the averages of these similarity ratios.

3.5.2. Success rate evaluation tools/methods

In the process of measuring the success of text summaries, it is very useful to create a detailed and repeatable benchmark procedure and automate some of these processes. Many tools/methods have been developed for this purpose:

Summary Evaluation Environment (SEE): Developed by C. Lin, the application measures success by comparing two different texts entered by the user side by side. One of the text summaries entered is the reference summary, while the other is the peer/candidate summary. The application pre-processes the texts and divides them at sentence level and measures according to the criteria selected by the user.

SimHash: In computer science, SimHash is a technique for quickly estimating how similar two sets are. The algorithm is used by google crawler to find duplicate pages.

We can explain the working logic with a small example;
(1001101101) xor (1001101011) = 00000000110
1 – (2 / 10) = 0.8 similar

All parts that are divided into meaningful parts are processed with the xor gate. Values are calculated as “0” for the same values and as “1” for the different ones. If we calculate the number of "1" values and divide by the total length, we can calculate how different they are. If we subtract this value from 1, we find how similar they are.

In this thesis, summaries from humans and machine summaries were compared using simhash. The averages of the obtained results are presented through visuals.

3.6. Proposed Model

A two-stages approach is adopted in the proposed model. The audio files collected in the first stage were translated into written documents using the speech-to-text tool. In this transcription step, the speech to text api offered by Google Cloud was used. This API is implemented in Visual Studio Code ide using python language. It was ensured that written documents with “.txt” extension were obtained from the “.wav” extension audio files given as input.

The resulting texts were summarized as 40% and 20% of the whole text. Example, if the total number of sentences is 100, the most important 40 sentences were selected for the 40% summary. For the 20% summary, the top 20 most important sentences were selected. The golden summaries produced independently by 3 different human summarizers. 3 engineers working in the IT department were selected for this job. Finally, the summaries from our program, which was written by applying the extractive summarization technique, were obtained.

For a meeting text, we compared the outputs of the program with the golden summaries and it was evaluated in sentence based similar selection to determine how much it is similar to the human summary. And the similarity averages of these 3 comparison gave the final success. Hundreds of such experiments have been carried out. Significant results were recorded and the results were tried to be improved. The proposed solution and work steps are clearly shown in Figure 3.3.

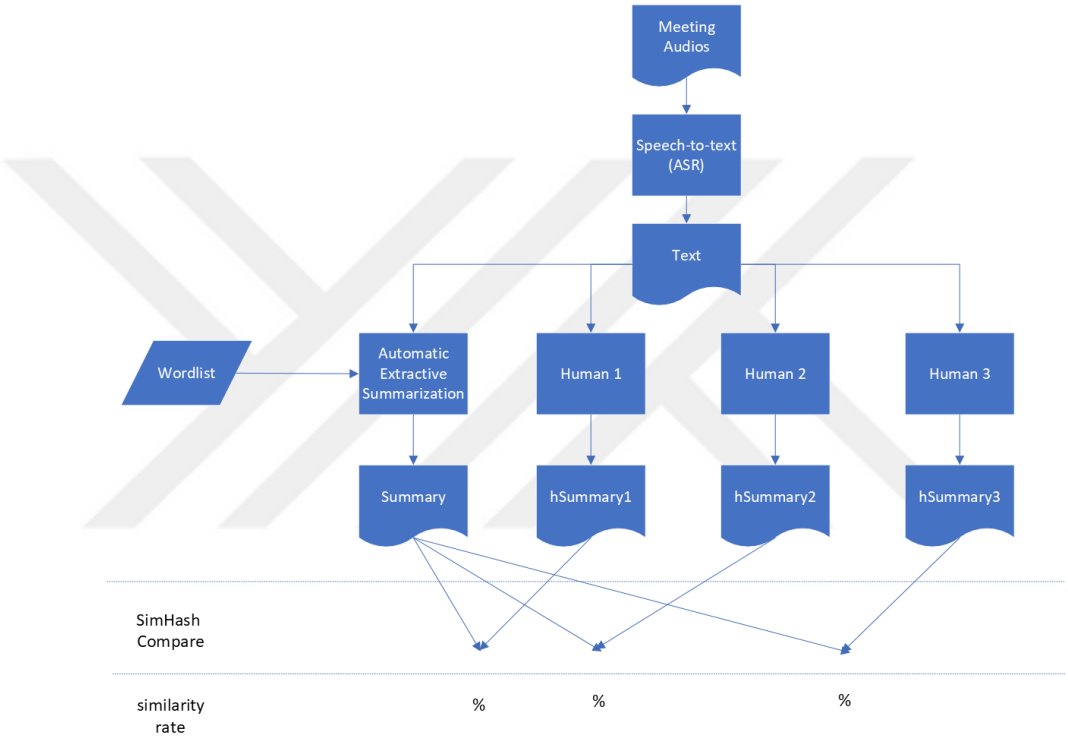


Figure 3.3. Summarization process map

3.6.1. Speech to text

Speech to text conversion is a methodology in which it provides the textual equivalant of the speech given as input to the computer program. The number of speakers, the style of speech, the purpose of the speech, the medium to which the speech is addressed, the audience, etc. depending on the elements, speech forms have developed. For example; a telephone conversation is made between two people through mutual devices, a news broadcast is formed by the unilateral transmission of the news text read by a single speaker. In environments such as

lecture presentations, interviews and meetings, if there is a need to transcribe the speech to text, it will have very different requirements.

For example, if you want to convert a voice recording of a news broadcast into text, a single voice recorder will suffice. In addition, since the announcers who present the news would have a proper pronunciation and way of addressing and it is a voice that belongs to only one person, there would not be many problems. However, if it is desired to record audio in a meeting room where 2 or more people are speaking and sharing their ideas, more effort would be required. When there are voices belonging to two or more people, then these sounds will have individual intonations and frequencies. In addition to the noise of the environment, non-meeting conversations can take place in whispers in small groups. Repetitive words, corrections, gaps, non-native speakers are all units that can create different problems where audio-to-text conversion is required. Although multiple speakers environment has matters to be cope with it, recording audio for crowded meeting room which is the real case for most of the meeting, is the main focus of this research .

Online tools were used to provide transcription of these audio recordings. There are many paid and free tools developed in this domain. Free Python library “SpeechRecognition” was the first attempt. Limited text conversion is applicable using the services of this library. Although it was sufficient to take and convert a single short-term sound recording, it was not efficient enough in an environment of mutual conversations. The low frequency sound was ignored as noise. In this way, it was not possible to process the audio recording we had, causing information loss.

The speech to text api offered by Google cloud was one of the most ambitious tools for voice processing. It was seen through experiments that the service provides a lot of flexibility and it has an advanced infrastructure. It offers the ability to accurately convert speech to text using API powered by Google's AI technologies. It is explained in more detail under the title of 3.1.1.2 Google cloud speech to text. The API enabled with Python coding enabled the

conversion of all audio recordings that were intended to be transcribed (Google Cloud Speech to Text, 2022).

3.6.2. Extractive summarization

Speech summarization is the process of obtaining information that will be useful in presenting the most important points of the speech to the end users in the shortest way. The speech summarization process includes speech recognition technology and summarization techniques. Written texts are summarized through NLP algorithms (NithyaKalyani and Jothilakshmi, 2019). There are many approaches produced according to the purpose of the summary. The two most basic approaches are extractive and abstractive summarization techniques (Ferreira et al., 2013). The extractive summarization technique is based on the selection of sentences that best express the text. Sentences are extracted directly to the summary. In abstract summarization, it is obtained by restating all the information in the text (Manjari et al., 2020). In the proposed model, the TF-IDF algorithm is used to select the original parts that best express the text (extractive summarization).

TF-IDF is the product of two statistics, term frequency and reverse document frequency. Term Frequency (TF) – Inverse Document Frequency (IDF) is a technique for measuring words in a series of documents. Its purpose is to define the importance of keywords or phrases in a document (Rajaman & Ullman, 2011).

The applied text summarization steps are shown in Figure 3.4. The details of the preprocessing and summarization of the TF-IDF algorithm are explained step by step from (a) to (f) below.

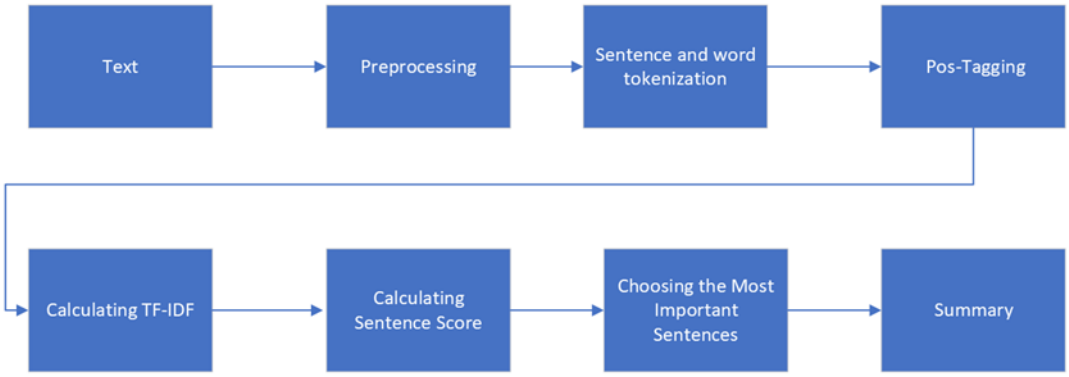


Figure 3.4. Summarization process steps

a) Preprocessing

Preprocessing is the first step required in any field working with data. It is important step to make data ready to be processed. It provides cleaning and standardization of the text before summarization. It is indispensable for NLP projects. The data should be well analyzed and visualized in order to decide how deep cleaning is required. The preprocessing steps vary depending on the content of the text and the problem definition. The processing steps used in this study are given below:

- Remove punctuation characters
- Lowercase conversion
- Remove numerals
- Spelling corrections
- Singularization
- Converting all words into base form
- Stop-words removal (Verma and Patel, 2017).

b) Tokenization

Tokenization is to divide large blocks of text into smaller units called tokens. For example, sentences in a paragraph and words (sometimes called terms) in sentences are extracted through tokenization. Tokens are assigned weights. These weights can depend on many parameters such as frequency, uniqueness.

In this study, sentence-based tokenization was used. At the end of this step, a weight was assigned to each sentence and a matrix containing them was obtained. This matrix was used to calculate the significance of the sentence.

c) Pos-Tagging

POS Tagging (Part of Speech Tagging) is the process of marking words. Marks words in text according to definition and context for a particular part of speech. Each word is grammatically labeled according to its type (POS tags, 2018). The accepted tags in the proposed model are: NN - noun (singular), NNS – noun (plural), NNP – proper noun (singular), VB – verb (base form), VBD – verb (past tense), VBG – verb (gerund) / present participle), VBN – verb (past participle), VBP – verb (sing. present, non-3d) and VBZ – verb (3rd person sing. present) tags were only evaluated in the model as nature of the problem, the nouns and verbs are considered as valuable. The other words were not included in.

d) Calculating TF-IDF

According to the symbols used in the TF-IDF formula, the lowercase letter d represents a document in the D document set, where t represents the word (noun or verb) selected in document d. In order of reduced frequency of the listed words, only the best percentage of the most frequently used words (20% and 50% word ratios were the best) were selected for further evaluation. This pattern is a technique used when we need to eliminate unnecessary words. It enables us to obtain faster and more efficient results by focusing on the most important words of the long word list produced.

e) Scoring the sentences

The score of each sentence is calculated using Equation (3.4), where the TF-IDF values of the words belonging to this sentence is the summed up.

$$Sentence\ Score = (\sum tfidf(t, d, D)) \quad (3.4)$$

The context of the problem is important when evaluating the words effect in the score evaluation. The special dictionary includes the specific words in the context is the one of the technique used in such problems in order to evaluate scores better or in other terms normalizing. The terms that occurs in the context dictionary (Roesler, 2020) were used with a learning alpha coefficient as an additional weight for sentences score. As a result, the ranking of sentences in importance worked better than the first classical approach. Final form of formula for scoring sentences is given in Equation (3.5).

The score of each sentence was calculated by adding the TF-IDF values of the words in this sentence. The equation expressed with the sum symbol is shown in Equation (3.4).

The context of the problem is important when evaluating the effect of words in scoring. Each problem contains characters, words and phrases specific to its context. The special dictionary includes the specific words in the context is the one of the technique used in such problems in order to evaluate scores better or in other terms normalizing. The terms in the dictionary used (Roesler, 2020) are included in the total with a special coefficient when calculating the sentence score. As a result, it has been observed that the order of importance of the sentences gives better results than the standard approach. The final form of the sentence scoring formula is given in Equation (3.5).

$$Sentence\ Score = (\sum tfidf(t,d,D)) + \alpha * \sum_{for\ all\ t \in d} \begin{Bmatrix} 1\ if\ t \exists\ dictionary \\ 0\ if\ t \nexists\ dictionary \end{Bmatrix} \qquad (3.5)$$

All the processing steps can be explained by an example. Assume “I am testing the current project.” sentence is being processed in meeting. All the process is explained stepwise below.

1. First of all preprocessing steps are applied one by one as given in Table 3.1.

Table 3.1. Sample sentence preprocessing steps

preprocessing step	output
Remove punctuation characters:	I am testing the current project
Lowercase conversion:	i am testing the current project
Remove numerals:	no
Spelling corrections:	no
Singularization:	no
Converting all words into base form:	i am test the current project
Stop-words removal:	test current project

2. Next tokenization is applied. Tokens are obtained as a list of words "test", "current" and "project".

3. Pos Tagging is applied to the tokens additionally in order to select only "Noun" and "Verb" tokens as expressed in Table 3.2.

Table 3.2. Pos tagging in sample sentence

word	pos tagging
"test"	VB – verb (base form)
"current"	not tagged, because it is not a verb or a noun
"project"	NN - noun (singular)

4. Evaluation requires counts, so that $tfidf$ values are calculated for tagged words.

$tfidf$ value of "test" is 0.065

$tfidf$ value of "project" is 0.12

5. Finally, sentence scoring is obtained using the formula given in (3.5) in Table 3.3.

Table 3.3. Results of sentence score with/out dictionary

approaches	test	project	sentence score
Standard Tf-Idf approach	0.065	0.12	0.185
Added Dictionary	10.2 x 0.065	10.2 x 0.12	1.887

As a result of repeated experiments, the coefficient of words in the special dictionary was determined as 10.2. Smaller numbers tried to determine this were

insufficient as not significantly affect the mean. On the other hand, coefficients larger than the selected coefficient, ignored the effect of words out of the dictionary, increased the average too much. In fact, it was observed that the coefficient, which was determined as 10, was added as 0.2 as the margin of error, and it gave optimum results as 10.2. Although the impact of the proposed model is shared in the Results and Evaluation section, it is a pleasure to share the 0.48 to 0.75 increase in similarity for the dictionary model with human summarizers (for 20% inferential summary). It is a strong evidence that the proposed model works correctly and makes a significant improvement.

f) Generating the Summary

Finally, the summary is obtained. After calculating the importance scores of the sentences in the whole text, a ranking is made from the most important to the least important. The extractive summary is obtained by placing sentences into the summary following the importance order decreasing order until it reaches the summary size.

4. DATA

Data is the most important building material of a scientific study. Data collection has become easier which leads to increase technical possibilities. Thanks to the collected data by researchers, scientific studies have increased. It is necessary to provide a experimental environment in order to prove the theories.

It was very difficult to access the actual meeting recordings (not only technical, also open the technical details public prevent such a real recording considering security issues). The voices in the dataset to be studied would belong to the meetings held in the field of information technologies. As a result of the research on the web, the desired recordings set was not found. That's why 10 meeting texts with real topics were prepared. Since these meetings were experimental, they were about 30 to 90 sentences related to the information technology. The characteristics of the audio recordings of each meeting are given in Table 4.1 below; the number of sentences and the length of the audio recording.

Table 4.1. Information of prepared meetings

#meeting	#sentences	#minutes
meeting 1	58	3’ 33’’
meeting 2	57	3’ 19’’
meeting 3	43	2’ 55’’
meeting 4	53	2’ 58’’
meeting 5	68	4’ 50’’
meeting 6	50	3’ 04’’
meeting 7	81	4’ 16’’
meeting 8	53	3’ 07’’
meeting 9	59	4’ 06’’
meeting 10	37	3’ 03’’

In the process of speech-to-text transcription, many obstacles arose. Problems occured the speaker's accent, ambient noise or inadequacy of voice recorders through speech-to-text process. Language support of automatic speech recognition tools was quite inadequate for non-native speakers. The output of these tools failed to produce satisfactory text output. Sometimes it wrote the closest homonym for the word it could not perceive, sometimes it wrote the

closest similar word. It also had difficulties due to the misunderstood speaker's accent. To solve these problems, support was obtained from the online tool Freetts (Freetts, 2021). Audio files were obtained by using the voices of people whose native language was English in the prepared meeting texts.

Each of the summary texts is named with a special naming style in order to avoid confusion. It will be mentioned how this nomenclature was created in case anyone wants to examine the data of the thesis work. Texts whose filename starts with “hSum” refer to summary texts obtained from golden summaries produced by human summarizer. The numbers in the hSum1, hSum2, hSum3 notation indicate that the text is summaries of 1st, 2nd, and 3rd summarizer. Filenames beginning with “sum” are used to express the summary outputs automatically obtained from the extractive summarization program. The number immediately following is given to show which meeting the text belongs to; given from 1 to 10. The capital letter “L” stands for the digest size followed by the percentage to represent the first letter of the word length. The next capital letter “W” is the first letter of the word ratio and means the percentage of words coming in. If it does not contain the letters "L" and "W", it means that it is not used.

In Table 4.2, the results in which the summaries of the meeting number 10 are compared with the summary texts of the 3rd person. According to this table, summaries with a length of 40% were compared. The data of the 10th meeting contains 37 sentences. In the first naming, it is understood that the summary text obtained by using only the TF-IDF algorithm is 0.53 similar to the summary obtained from the third person. The “word ratio” in the second line is the unused dictionary form. The third line uses the dictionary form with a “word ratio” of 50%. The last one uses the dictionary form with a “word ratio” of 20%.

Table 4.2. Sample data

#37 total sentences	hSum310-L40%
sum10-L40%-tf-idf	0,53
sum10-L40%	0,64
sum10-L40%-W50%	0,65
sum10-L40%-W20%	0,73

Human summaries obtained from our experimentally created dataset were compared for similarity. In Figure 4.1, 40% gives a comparison of summaries; Figure 4.2 presents a comparison of 20% summaries. Similarity charts show that the data are not contradictory and the distribution is consistent. These mean similarities can also be considered as benchmarks for the success of model summaries.

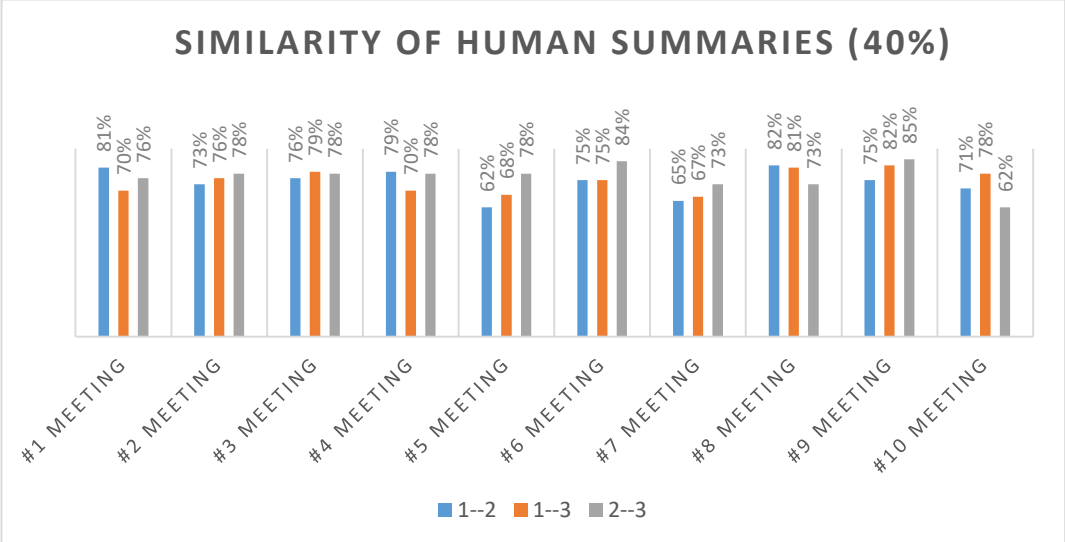


Figure 4.1. Similarity rates of human summaries of 40% length

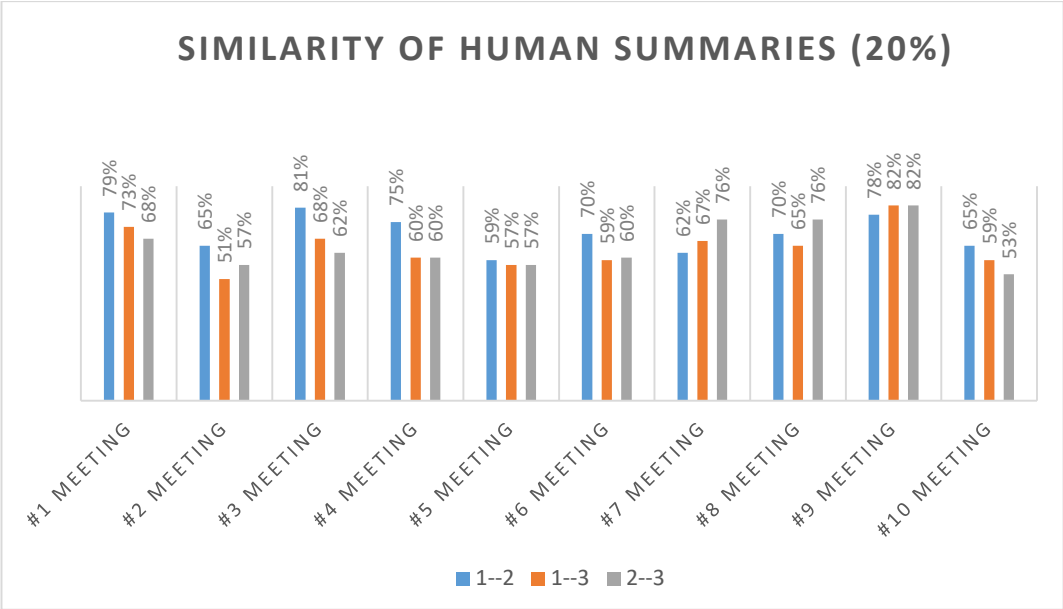


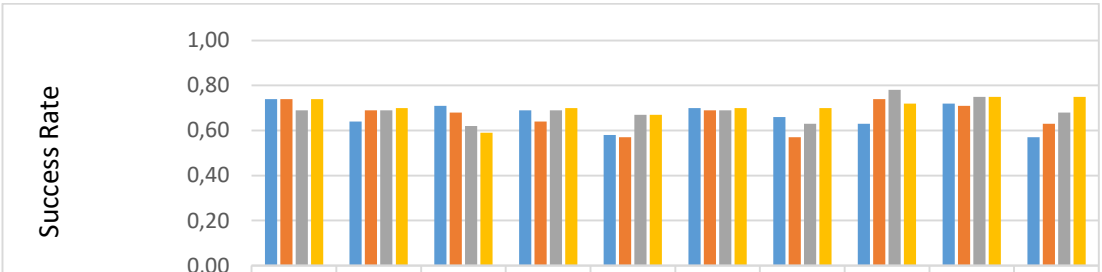
Figure 4.2. Similarity rates of human summaries of 20% length

As a result, similarity rates of human summaries range from between 60% and 86% for length 40%. Its average is 75%. In the data of the summary length of 20%, the similarity rates vary between 50% and 83%. When took the average, a value of 67% was found. It is important to talk about the nomenclature related to the graphic in order to clarify the question marks in the minds. The area indicated by the blue line represents the similarity ratio of human 1st and 2nd human summaries; the orange line represents the similarity ratios of the 1st and 3rd human summaries; The gray line represents the similarity ratios of the 2nd and 3rd human summaries.

5. RESULTS AND EVALUATION

The results obtained during the whole thesis study were recorded and the success rates were examined. As a result of the studies, the texts with a summary length of 40% and 20% provided sufficient evidence to show the accuracy of our algorithm. It has reached the expected level of success. When the term frequencies of the words were examined, it was observed that 50% was the optimum value. By taking the average of all similarity rates, it was observed that approximately 71% of success was achieved. The optimum results of the study are given in Figure 5.1 and Figure 5.3.

Some explanations about Figure 5.1 are as follows. In this table, the results of the similarity of human and machine outputs with a summary length of 40% are given. The blue line represents the points where only the texts obtained by applying the TF-IDF algorithm were compared. The orange line represents the dictionary form that does not use "word ratio". The gray line represents adding a dictionary with 50% “word ratio”, the yellow line representing using a dictionary with 20% “word ratio”.



	mean1	mean2	mean3	mean4	mean5	mean6	mean7	mean8	mean9	mean10
sumX-L40%-tf-idf	0,74	0,64	0,71	0,69	0,58	0,70	0,66	0,63	0,72	0,57
sumX-L40%	0,74	0,69	0,68	0,64	0,57	0,69	0,57	0,74	0,71	0,63
sumX-L40%-W50%	0,69	0,69	0,62	0,69	0,67	0,69	0,63	0,78	0,75	0,68
sumX-L40%-W20%	0,74	0,70	0,59	0,70	0,67	0,70	0,70	0,72	0,75	0,75

Figure 5.1. Similarity chart human & extractive summaries of 40% length

In order to show the increase in the success rate, the average of all similarity rates was taken at each applied step and it was observed that the success rate increased at each point. When reflected on the graph, an increasing graph was encountered

as in Figure 5.2. The highest similarity was found where the “word ratio” was 20%, with almost 71% success.

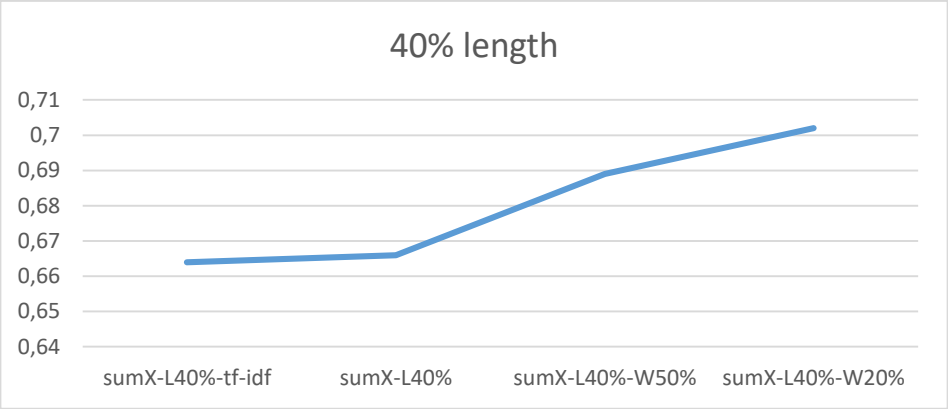
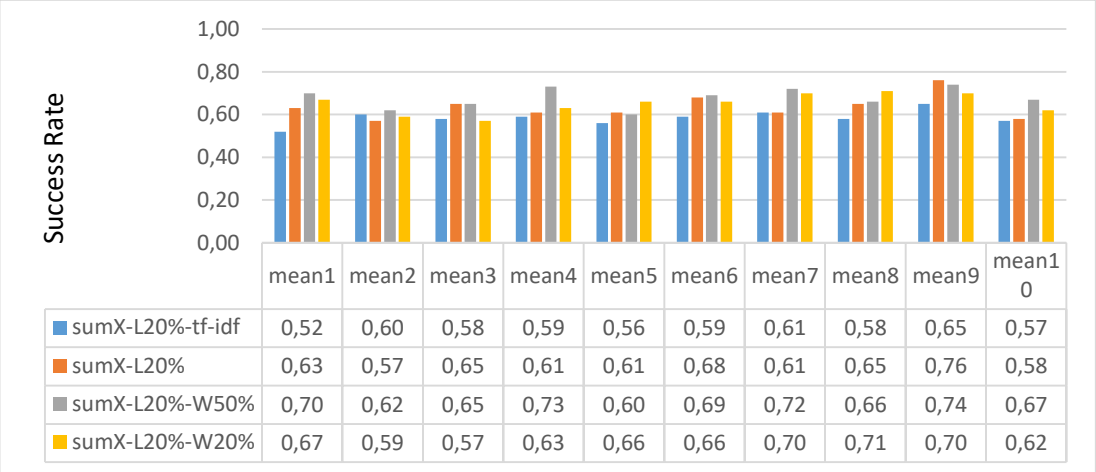


Figure 5.2. Mean of all averages of the similarity ratios 40% length

According to the results obtained by evaluating the same metrics, the similarity results of the abstracts with a text length of 20% are shown in Figure 5.3.



	mean1	mean2	mean3	mean4	mean5	mean6	mean7	mean8	mean9	mean10
sumX-L20%-tf-idf	0,52	0,60	0,58	0,59	0,56	0,59	0,61	0,58	0,65	0,57
sumX-L20%	0,63	0,57	0,65	0,61	0,61	0,68	0,61	0,65	0,76	0,58
sumX-L20%-W50%	0,70	0,62	0,65	0,73	0,60	0,69	0,72	0,66	0,74	0,67
sumX-L20%-W20%	0,67	0,59	0,57	0,63	0,66	0,66	0,70	0,71	0,70	0,62

Figure 5.3. Similarity chart human & extractive summaries of 20% length

By taking the arithmatic average of the similarity ratios in each application step and plotting these points on the graph, obtained the graph in Figure 5.4 (for summaries with a text length of 20%).

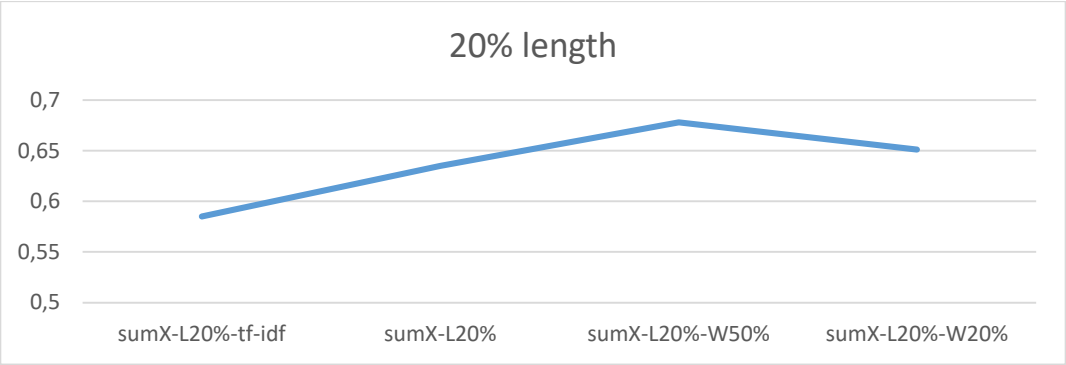


Figure 5.4. Mean of all averages of the similarity ratios 20% length

If a briefing is made by looking at all the results; Figure 5.2 and Figure 5.4 will show that the success rate increases with each approach. The rate of increase in our approach including the dictionary is remarkable. The blue lines in Figure 5.1 and Figure 5.3 show the success rate without dictionary, and the orange lines show the success rate including the dictionary. It is seen that if the selective words are determined well (the best ratio is 50% of the wordlist produced from tf-idf), more suitable sentences can be selected for the summary.

6. CONCLUSION AND IMPLICATIONS

Finally, need to make a few comments in this section. In addition, some future working ideas and working methods that may be good to realize for a next step that comes to our mind while doing this study will also be mentioned.

While evaluating the summaries taken from our extractive summarization model, references are the golden summaries produced by independent human summarizers whose are expert in the related field. 3 people who are actively developing software in the IT sector were selected. They were asked to summarize the texts of the meetings at the rates of 20% and 40%. The golden summaries were also compared for similarity ratio among themselves. The similarity of golden summaries shows us how they are similar or differ from others. This can be accepted as a benchmark metric to compare, even a low success ratio of automatically created summary. The results are summarized in Figure 4.1 and Figure 4.2.

It has been observed that the similarity rates of the summaries prepared by 3 different experts are the highest 82% and the lowest 51% in Figure 4.1 and Figure 4.2. In the figures, which are represented by lines in 3 different colors, the first line gives the 1st and 2nd human summaries, the second line gives the 1st and 3rd human summaries, and the third and last line gives the similarity ratios of the 2nd and 3rd human summaries. The average value obtained by comparing the summaries produced by the program with the summaries of humans suggests that the summary obtained from the program produces more accurate results compared to humans, even if it is created differently.

The graphs in which the averages taken are reflected are Figure 5.2 and Figure 5.4. These graphs are shown by averaging the rows using the same method. It is observed that the success rate increases with each step. Especially since the use of the dictionary causes the selection of more appropriate sentences, the graphs are in the increasing direction. In addition, the summaries produced using the first 20% and 50% of the words belonging to the word list produced using tf-idf

show that they are more similar to human summaries. Because long sentences are more advantageous when the standard approach is used. The important thing is not the length of the sentence, but its effectiveness in the dictionary approach. Thus, more appropriate sentences were extracted using keywords. On average, approximately 71% similarity was achieved.

When these results are compared with the average similarity of the summaries produced by the summators, it is seen that the proposed model produces almost the same similarity (71%). This result encourages us to further work to achieve better than human.

A few suggestions and approaches for future work will be shared. All spoken language rules can be included that help improve speech summarization performance. Also, this model can be used in many fields as it automatically converts and summarizes speech to text; lectures, conferences, telephone conversations, etc. It can help both employees and students take notes in places such as training, seminars, symposiums. Multi-language support can be developed (especially non-native English speakers) and can be used with languages other than English. In addition, by using speech recognition algorithms, personal labels can be made to keep the information of who the summary sentence belongs to.

This study was carried out in English, but its development in Turkish and its support with AI algorithms will contribute very well to our language and will be a study that will inspire new generation researchers.

REFERENCES

Awadallah, A., H., Celikyilmaz, A., Jha, R., Liu, Y., Mutuma, M., Radev, D., Yin, D., Yu, T., Zaidi, A., Zhong, M., Qiu, X., 2021. Qmsum. https://github.com/Yale-LILY/QMSum/tree/main/data/ALL

Baxendale, P., 1958. Machine-made index for technical literature - an experiment. IBM Journal of Research Development, 2(4), 354-361.

Baykal, B., Sert, M., Yazici, A., 2008. A Combining structural analysis and computer vision techniques for automatic speech summarization, in Proceedings - 10th IEEE International Symposium on Multimedia, ISM 2008, 515–520.

Bhamidipati, S., Shi, Y., Bryan, C., Zhao, Y., Zhang, Y., Ma, K., 2018, June 1. MeetingVis: Visual narratives to assist in recalling meeting context and content, in IEEE Transactions on Visualization and Computer Graphics, 24(6), 1918-1929.

Bharti, S., K., Babu, K., S., 2017. Automatic Keyword Extraction for Text Summarization: A Survey, Journal of Computing Research Repository, arXiv:1704.03242.

Breitinger, C., Gipp, B., Langer, S., 2015, July 26. Research-paper recommender systems: a literature survey, in International Journal on Digital Libraries. 17 (4), 305–338.

Cambridge Dictionary, Summary, Date Accessed: 22.04.2022. https://dictionary.cambridge.org/dictionary/english/summary

Choi, J., Gill, H., Ou., S., Song, Y., Lee, J., 2018. Design of Voice to Text Conversion and Management Program Based on Google Cloud Speech API, in 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 1452-1453.

Chowdhary, K., R., 2020. Fundamentals of Artificial Intelligence. Springer India, ISBN: 978-81-322-3970-3.

DeJong, G., Kolodner, J., Schank, R., 1980. Conceptual Information Retrieval, Yale University, Department Of Computer Science.

Edmundson, H., P., 1969. New methods in automatic extracting, Journal of the ACM, 16(2), 264-285.

Elfeki, M., Wang, L., Borji, A., 2022. Multi-stream dynamic video Summarization, in IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 185-195.

Fernandez, S., Graves, A., Schmidhuber, J., 2007. An application of recurrent neural networks to discriminative keyword spotting, in Proceedings of ICANN, (2), 220–229.

Ferreira, R., Cabral, L., S., Lins, R., D., Silva, G., P., Freitas, F., Cavalcanti, G., D., C., Lima, R., Simske, S., J., Favaro, L., 2013. Assessing sentence scoring techniques for extractive text summarization, Expert Systems with Applications, 40(14), 5755-5764, ISSN 0957-4174.

Fifth Generation Computer Corporation, 2022. Speaker Independent Connected Speech Recognition. Date Accessed: 11.03.2022. http://www.fifthgen.com/speaker-independent-connected-s-r.htm

Freetts, Text to Speech, 2021, September 25. [Online]. Available: https://freetts.com/ads

Froomkin, D., 2015, May 5. The Computers Are Listening, The Intercept. Archived from the original on 27 June 2015. Date Accessed: 20 June 2015.

Furui, S., Kikuchi, T., Shinnaka, Y., Hori, C., 2004, July. Speech-to-text and speech-to-speech summarization of spontaneous speech, in IEEE Transactions on Speech and Audio Processing, 12(4), 401–408.

Garcia - Hernandez, R., A., Gelbukh, A., Montiel, R., Ledeneva, Y., Rafael, C., Rendon, E., 2008. Text Summarization by Sentence Extraction Using Unsupervised Learning, in MICAI 2008: Advances in Artificial Intelligence, vol. 5317, 133-143.

Gonçalves, L., 2020. Automatic Text Summarization with Machine Learning — An overview**,** Date Accessed: 28.04.2022. https://medium.com/luisfredgs/automatic-text-summarization-with-machine-learning-an-overview-68ded5717a25

Google Cloud Speech-to-Text, 2022. Date Accessed: 01.10.2022. https://cloud.google.com/speech-to-text#all-features

Guida, G., Mauri, G., 1986, July. Evaluation of natural language processing systems: Issues and approaches, in Proceedings of the IEEE, 74(7), 1026-1035.

Gupta, V., Lehal, G., S., 2010. A Survey of Text Summarization Extractive Techniques, Journal of Emerging Technologies in Web Intelligence, 2(3), 258-268.

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory, Neural Computation, 9(8), 1735-1780.

Hori, C., Furui, S., 2000. Automatic speech summarization based on word significance and linguistic likelihood, in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings (Institute of Electrical and Electronics Engineers Inc., 2000), vol. 3, 1579–1582.

Hovy, E., McKeown, K., Radev, R., D., 2002. Introduction to the Special Issue on Summarization, Computational Linguistics, 28(4), 399-408.

Huebner, A., Ji, W., Xiao, X, 2021, November 16. Meeting Summarization with Pre-training and Clustering Methods.

Jones, K., S., 1972. A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, 28(1), 11-21.

Jones, K., S., Galliers, J., 1996. Evaluating Natural Language Precessing Systems: An Analysis and Review, in Springer, Lecture Notes in Artificial Intelligence 1083.

Juang, B. H.; Rabiner, L., R, 2017. Automatic speech recognition–a brief history of the technology development, Date Accessed: 17.08.2014.

Kanda, N., Xiao, X., Gaur, Y., Wang, X., Wang, X., Meng, Z., Chen, Z., Yoshioka, T., 2022, May. Transcribe-to-Diarize: Neural Speaker Diarization for Unlimited Number of Speakers using End-to-End Speaker-Attributed ASR, in ICASSP 2022.

Kanda, N., Ye, G., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Yoshioka, T., 2021, August. End-to-End Speaker-Attributed ASR with Transformer, in Interspeech 2021.

Kanda, N., Ye, G., Wu, Y., Gaur, Y., Wang, X., Meng, Z., Chen, Z., Yoshiokai T., 2021, August. Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone. Interspeech 2021.

Li, M., Zhang, L., Ji, H., Radke, R., J., 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarizationi, in ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2190–2196.

Luhn, H., P., 1957. A statistical approach to mechanized encoding and searching of literary information, IBM Journal of Research and Development, 1(4), 309-317.

Makhervaks, D., Hinthorn, W., Dimitriadis, D., Stolcke, A, 2020. Combining Acoustics, Content and Interaction Features to Find Hot Spots in Meetings, in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8054-8058.

Mani, I., Maybury, M., T., 1999. Advances in Automatic Text Summarization, MIT Press, Cambridge.

Manjari, K., U., Rousha, S., Sumanth, D., Sirisha Devi, J., 2020. Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm, in 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), 648-652.

Manjari, K., U., Rousha, S., Sumanth, D.,Sirisha Devi, J., 2020. Extractive Text Summarization from Web pages using Selenium and TF-IDF algorithm, in 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), 648-652.

Mirkin, B., 2011. Core Concepts in Data Analysis: Summarization, Correlation and Visualization, Springer.

Munot, N., Govilkar, S., S., 2014. Comparative Study of Text Summarization Methods, International Journal of Computer Applications, 102(12), 33-37.

Nadkarni, P., M., Machado, L., O., Chapman, W., W., 2011, September. Natural language processing: an introduction, Journal of the American Medical Informatics Association, 18(5), 544-551.

NithyaKalyani A., Jothilakshmi, S., 2019. Intelligent Speech Signal Processing, editor(s): Nilanjan Dey, Academic Press, 113-138.

NLTK, Date Accessed: 27.04.2022. https://www.nltk.org/

Özkan, C., 2019. İnternet Tabanlı Türkçe Metinler için Otomatik Özetleme Tekniği, Maltepe University, Graduate School of Natural and Applied Sciences, M.Sc. Thesis, 62p, Istanbul.

Pinola, M., 2011, November 2. Speech Recognition Through the Decades: How We Ended Up With Siri, Date Accessed: 28.07.2017, PC World.

POS tags, Sketch Engine, Lexical Computing, 2018, March 27. Date Accessed: 01.12.2021.

Rajaraman, A., Ullman, J., D., 2011. Data Mining, (PDF), Mining of Massive Datasets, 1–17.

Robertson, S., 1972. Understanding inverse document frequency: on theoretical arguments for IDF, Journal of Documentation, 6(5).

Roesler, D., 2020. A Computer Science Academic Vocabulary List, Dissertations and Theses, Dep. Applied Linguistics, Portland State Univ., Paper 5540.

Roh., H., Lee, K., 2017, December. A Basic Performance Evaluation of the Speech Recognition APP ofStandard Language and Dialect using Google, Naver, and DaumKAKAO APIs, Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology, 7(12), 819-829.

Shannon, C., E., 1940. An Algebra for Theoretical Genetics, Dept. of Mathematics, Massachusetts Institute of Technology, Ph. D. Thesis, 74p., Cambridge.

Shriberg, E., Dhillon, R., Bhagat, S., Ang J., Carvey, H., 2004. The ICSI meeting recorder dialog act (MRDA) corpus, in Proceedings of the SIGDIAL 2004 Workshop - 5th Annual Meeting of the Special Interest Group on Discourse and Dialogue (ACL), 97–100.

Thompson, H., S., 1991. Natural Language Processing: An Overview, 354-362.

Torres-Moreno, M., J., (Ed.), 2014, November 10. Automatic Text Summarization. Wiley-ISTE, 376p.

Tur, G., Stolcke, A., Voss, L., Dowding, J., Favre, B., Fernandez, R., 2008. The CALO meeting speech recognition and understanding system, in 2008 IEEE Spoken Language Technology Workshop, 69-72.

Turchin, A., Florez B., Luisa F. 2021, March 19. Using Natural Language Processing to Measure and Improve Quality of Diabetes Care: A Systematic Review, Journal of Diabetes Science and Technology, 15 (3), 553–560.

Verma, J., P., Patel, A., 2017. Evaluation of Unsupervised Learning based Extractive Text Summarization Technique for Large Scale Review and Feedback Data, Indian J. Sci. Technol, vol.10, 1–6.

Watson IBM corporation, 1911. Watson speech to text. Date Accessed: 16.04.2022. https://www.ibm.com/tr-tr/cloud/watson-speech-to-text

Wikipedia, Meeting, Date Accessed: 25.04.2022. https://en.wikipedia.org/wiki/Meeting

Wikipedia, Natural Language Procecessing, Date Accessed: 26.04.22. https://en.wikipedia.org/wiki/Natural_language_processing

Wikipedia, Tf-idf, 2022. Date Accessed: 29.04.2022. https://en.wikipedia.org/wiki/Tf%E2%80%93idf

Zhang, J., Yuan, H., 2013. Speech Summarization without Lexical Features for Mandarin Presentation Speech, in Proceedings - 2013 International Conference on Asian Language Processing, IALP 2013, 147–150.

Zhang, J.,Yuan, H, 2014. A comparative study on extractive speech summarization of broadcast news and parliamentary meeting speech, in Proceedings of the International Conference on Asian Language Processing 2014, IALP 2014 (Institute of Electrical and Electronics Engineers Inc., 2014), 111–114.

Zhu, C., Xu, R., Zeng, M., Huang, X., 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining, in Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020, 194-203.

BIBLIOGRAPHY

Name and Surname : Neslihan AKAR

Educational Status

High School : Cengizhan Anatolian High School, 2011

Bachelor's Degree : Marmara University
Faculty of Engineering
Computer Engineering, 2016

Master's Degree : Istanbul Commerce University
Graduate School of Natural and Applied Sciences
Computer Engineering, 2022

Publications

Akar, N., Turan, M., 2022, June, 24. A General Approach For Meeting Summarization: From Speech To Extractive Summarization, Journal of Conscientia Beam, Review of Computer Engineering Research, 9(2), 83-95.