TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Henry Härm 191935IAPM

ABSTRACTIVE SUMMARIZATION OF NEWS BROADCASTS FOR LOW RESOURCE LANGUAGES

Master's Thesis

Supervisor: Tanel Alumäe

PhD

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Henry Härm 191935IAPM

RAADIOUUDISTELE ABSTRAHHEERIVATE KOKKUVÕTETE GENEREERIMINE VÄHESTE TREENINGANDMETE JUURES

Magistritöö

Juhendaja: Tanel Alumäe

PhD

Author's declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Henry Härm

08.05.2021

Abstract

Abstractive summarization models can generate much shorter summaries and avoid redundancy as compared to extractive methods [1]. However, training these models typically requires large datasets, which is not practical in most industry use-cases. This paper describes an approach for generating abstract summaries for Estonian news broadcasts in a low-resource setting. Given a news broadcast episode recording, the goal is to create a summary that captures the essential information in a short format. The approach consists of two steps: Automatically generating the transcript, applying a state-of-the-art text summarization system. Several methods are proposed with the best performing leveraging large English BART model pre-trained on CNN/DailyMail dataset and fine-tuned on task data machine translated from the target language. The method outperformed our baseline by 5.78 ROUGE-L and improved on the baseline in human evaluation. The applicability of the proposed solution needs to be considered in languages where machine translation systems are not mature, where national BERT and multilingual RoBERTa models should be considered. The experimentations showed the former improving on the baseline by 0.73 and later by 0.29 ROUGE-L.

This thesis is written in English and is 35 pages long, including seven chapters, ten Figures and sixteen Tables.

Annotatsioon

Raadiouudistele abstrahheerivate kokkuvõtete genereerimine väheste treeningandmete juures

Internetiplatvormide nagu YouTube ning multimeedia failide kasvuga on tekkinud vajadus tagada nendele ressurssidele lihtsat ligipääsu [2]. Kasutajatel kulub palju aega informatsiooni leidmiseks ning rohkete otsingutulemuste läbi vaatamine pole realistlik [1]. Lühikokkuvõtted teevad multimeedia otsimise lihtsamaks, kuid nende käsitsi loomine on kulukas. *Transformers* mudelite baasil ehitatud automaatsed abstraheerivad kokkuvõtte süsteemid on saavutanud häid tulemusi, kuid vajavad treenimiseks suurel hulgal annoteeritud andmeid.

Käesolevas töös kirjeldatakse lähenemisviisi abstraktsete kokkuvõtete loomiseks ERR (Eesti Rahvusringhääling) uudistesaadete jaoks madala ressurssidega tingimustes. Saades uudistesaadete episoodi, on eesmärk luua kokkuvõte, mis esitab kõige olulisemat teavet võimalikult lühidalt. Lähenemisviis koosneb kahest etapist: transkriptsiooni automaatne genereerimine ning tipptasemel teksti automaatkokkuvõtte süsteemi rakendamine. Esimese sammuna koostatakse ERR uudiste andmestik koos autorite poolt käsitsi kirjutatud kokkuvõtetega. Protsessi lihtsustamiseks luuakse automaatne andmete kogumise ja töötlemisse tööriist. Parima lahenduse leidmiseks võrreldakse kõige olulisemaid avalikult jagatavaid *transformer* mudeleid genereeritud kokkuvõtte kvaliteedi alusel. Piiratud andmestiku probleemide ületamiseks uuritakse eeltreenitud mudeleid, mitmekeelseid mudeleid ning masintõlkimise rakendamist.

Töös välja pakutud meetoditest parima tulemuse saavutas suur ingliskeelne BART mudel, mis on eeltreenitud CNN/DailyMail andmekoguga mida omakorda treeniti sihtkeelest masintõlgitud ülesande andmestikul. Meetod ületas meie referentssüsteemi 5,78 ROUGE-L (*Recall-Oriented Understudy for Gisting Evaluation*) võrra ning ületas inimhinnangute tulemused. Keeltes, kus masintõlkesüsteemid pole saavutanud vajalikku täpsust tuleb kaaluda rahvuslike BERT (*Bidirectional Encoder Representations from*

Transformers) ja mitmekeelseid RoBERTa mudeleid. Katsed näitasid, et muldeid parandasid referentssüsteemi tulemust 0,73 ja 0,29 ROUGE-L võrra vastavalt.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 35 leheküljel, kuus peatükki, kümmet joonist, kuusteist tabelit.

List of abbreviations and terms

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
ATS	Automatic Text Summarization
BERT	Bidirectional Encoder Representations from Transformers
CNN	Cable News Network
DAPT	Domain-Adaptive Pre-Training
ELMo	Embeddings from Language Models
ERR	Estonian Public Broadcasting
GLUE	General Language Understanding Evaluation
LRL	Low-Resource Languages
NLP	Natural Language Processing
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SDPT	Source Domain Pre-Training
Seq2Seq	Sequence-to-Sequence
SQuAD	Stanford Question Answering Dataset
TAPT	Task-Adaptive Pre-Training
TREC	Text Retrieval Conference
QDAS	Quality-Diversity Automatic Summarization

Table of contents

List of Figures

List of Tables

1 Introduction

1.1 Problem Context

With the growth of internet platforms such as YouTube and the ubiquity of online multimedia, there is a significant need to provide easy access to these resources [2]. This media can contain documents such as talks, presentations, lectures and news [3]. The user spends a lot of time finding the information they are searching for and cannot realistically look through all search results [1]. Although speech is the most natural and effective method of communication between human beings, it is not easy to quickly review, retrieve and reuse speech documents if they are simply recorded as an audio signal [4], [5]. Manual summarization is expensive and consumes a lot of time, making it practically impossible to use massive datasets [1]. Therefore, processing speech automatically into transcripts and summaries is a crucial area of research.

Automatic summarization of a document produces a concise and fluent summary while preserving critical information from the original text. It is considered a challenging task because humans, to highlight a documents main points, work through the source entirely to develop their understanding. The person, in many cases, needs to have background information on the topic and understand what readers regard as common knowledge. Since computers lack human knowledge and language capability, it makes such summarization an exceedingly non-trivial and challenging task [6]. There are two major approaches for automatic text summarization: extractive and abstractive. The extractive summarization approach produces summaries by choosing a subset of sentences from the original text. Abstractive text summarization aims to shorten the text highlighting the essential information. In addition, analytical summarization techniques such as LexRank compute the relative importance of words and sentences to produce the summary [7].

As podcasts primarily contain spoken-word content, summarization can also be performed in the text domain on the transcript of an episode. One such example is the PodSumm method, which works by first transcribing the spoken content of a podcast,

then identifying essential sentences in the transcript, and finally, stitching together the respective audio segments [8]. The deep learning-based neural attention model performs well when applied to abstract text summarization [9] compared to standard analytical learning-based approaches. In general, neural summarization is solved using an encoder-decoder architecture with recurrent neural networks or self-attention [10]. However, there is an inherent limitation to natural language processing tasks such as text summarization for resource-poor and morphologically complex languages owing to a shortage of quality linguistic data available [11]. For the state-of-the-art neural models, annotated English datasets of hundreds of thousands or millions of data points are used. At the same time, such quality and quantity are not feasible for most languages, including Estonian. One such popular dataset for text summarization is the CNN (Cable News Network) and Daily Mail dataset with around 300 000 news articles with their handwritten summaries.

1.2 Objective

Recent neural abstractive summarization models have proven to have excellent performance but are generally trained on vast datasets. This work focuses on improving news broadcast summarization quality under low resource conditions by utilizing recent research insight from this field. A dataset of news broadcasts and handwritten summaries is created from the ERR archive in the Estonian language as a first step. To facilitate this, an automatic data collection and processing tool is developed. The most significant transformer-based state-of-the-art models are compared to the low-resource abstract summarization task based on their performance. To overcome the problems of limited available data for training, transfer learning methods on pre-trained models, multilingual models and machine translation are explored and included in the summarization pipeline. The hypothesis is that utilizing one or a combination of these techniques will significantly improve the performance achieved in the state-of-the-art.

1.3 Outline

This work is divided into seven chapters, with the first introducing the topic, goals and outlining the problem. The second Chapter explains vital theoretical ideas and related information on the subject. The third chapter reviews the existing research and findings of authors in the problem area. The fourth explains the methodology that has been chosen

for the study. The fifth chapter outlines the conducted experiments and the observed results. The sixth chapter presents the key developments and conclusions of the work. The last Chapter discusses future work and areas of research.

2 Background

2.1 Automatic text summarization

Automatic text summarization (ATS) is a challenging task in NLP (Natural Language Processing) and AI (Artificial Intelligence) with open research areas such as summarizing single long document such as a book, summarizing multi-documents, evaluating a generated summary without a handmade reference summary, generation of an abstractive summary [1]. According to [1], researchers are still working to improve these systems to cover all the main topics in the text and is cohesive to read.

There are three main approaches for generating summaries: extractive, hybrid and abstractive Figure 1). The underlying summarization methods that the approaches utilized are, for example, deep-learning, graph and structure-based.

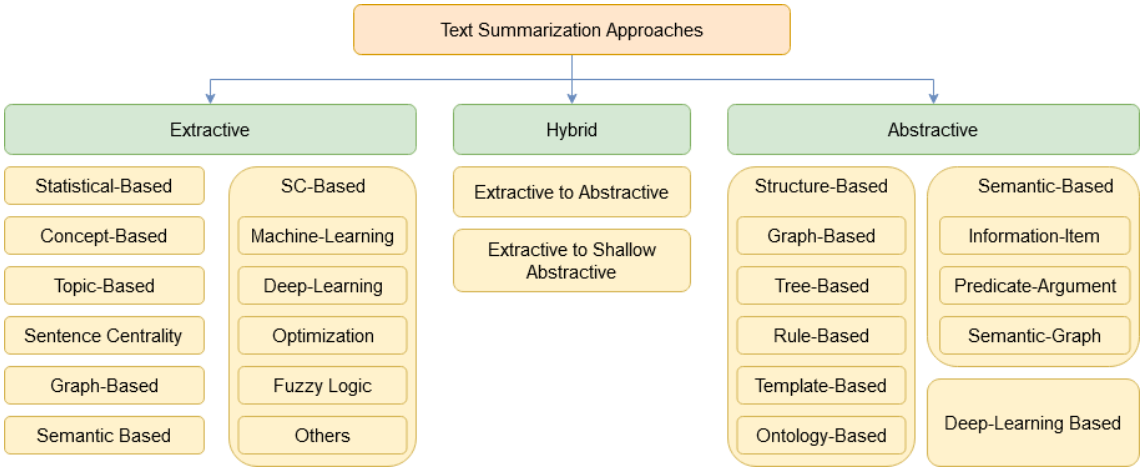

Figure 1. Automatic text summarization approaches [1].

2.1.1 Extractive approach

An extractive text summarization system (Figure 2) consists of pre-processing, processing, and post-processing. The processing consists of creating a representation of the input text and scoring the sentences according to a ranking system. High score

sentences are extracted, preserving the original order of the text with a determined cut-off length. The extractive approach uses sentences directly from the document, giving higher accuracy and is more straightforward than the abstractive approach. However, the method can lead to redundancy, a lack of cohesion and temporal conflicts in sentences [1].

The field has significantly benefited from the introduction of robust statistical techniques. For example, a stochastic graph-based method for computing the relative importance of textual units for text summarization has been proposed called LexRank [7]. The technique works by calculating sentence importance from the eigenvector centrality in a sentence graph representation. A connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation. LexRank outperforms other systems and centroid-based methods and in several tasks [7].

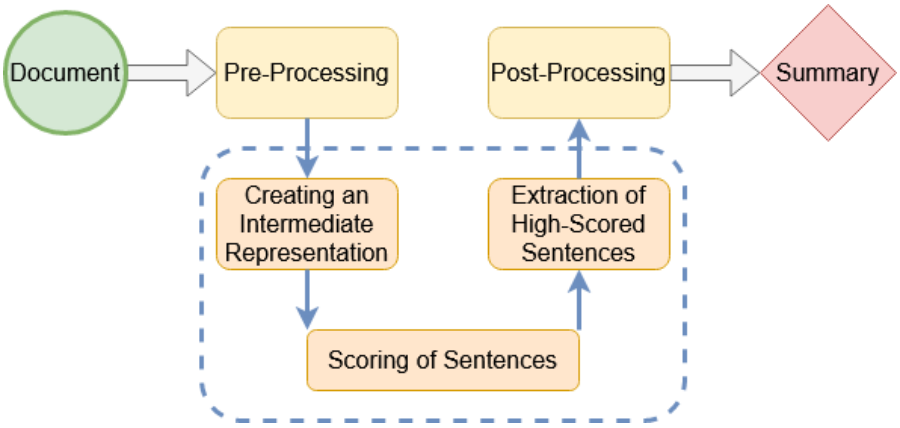

Figure 2. Extractive summarization system architecture [1].

2.1.2 Abstractive approach

The abstractive system (Figure 3) consists of pre-processing, post-processing and processing parts. The processing task includes creating an internal representation of the document and generating a summary. The abstractive system understands the main concepts and paraphrases the text for a clear and concise summary. The system can generate much shorter summaries and avoids redundancy [1]. Recent advances in deep neural networks have made this method viable; however, high accuracy results are hard to achieve in practice. The system often suffers from generating repeating words and struggles with out-of-vocabulary words [12]; the system is limited by creating an

intermediate representation, which is possible only in certain domains. General-purpose solutions are currently not viable [13]. To achieve the best results, it is recommended to combine other methods in addition to deep learning techniques such as template-based, semantic-based, graph-based, tree-based, rule-based and ontology-based methods [1].

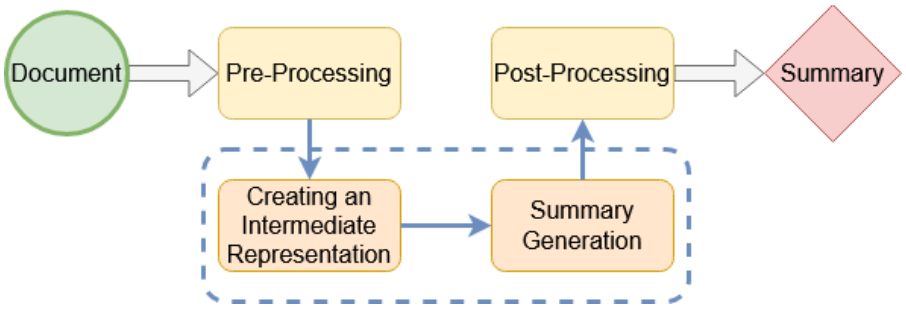

Figure 3. Abstractive summarization system architecture [1].

2.1.3 Hybrid models

A hybrid system (Figure 4) consists of pre-processing, post-processing and processing tasks. The processing task generates an extractive summary based on sentence scoring and ranking. The extractive summary then is used to create an abstractive summary [14]. The system combines the advantages of both approaches, and overall performance is improved; however, it results in a lower quality summary than a purely abstract method [1].

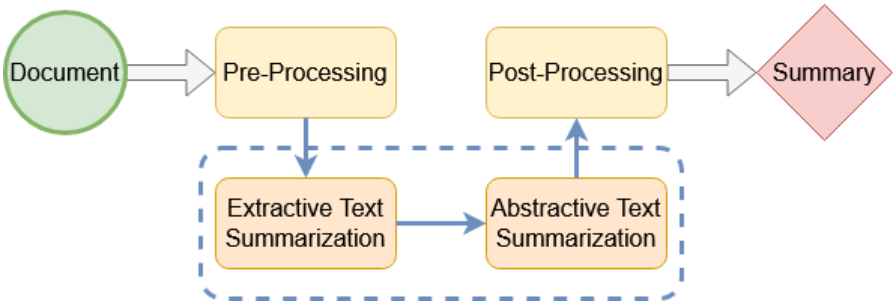

Figure 4. Hybrid summarization system architecture [1].

2.1.4 Evaluation methods

Evaluating an automatically generated summary is complex as there is no one ideal summary to a document. The correct result is mainly up to discussion, and human

summarizers have a low agreement for producing summaries. The lack of standardized metrics makes the task of automatic summarization more challenging [6]. The simplest but most expensive method of summary evaluation is to assess its quality manually. This can be done by an expert or scoring by a large group.

Automatic evaluation metrics have also been developed, with ROUGE being the most widely used. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It automatically determines the summary quality by comparing it to other human-made summaries by counting the number of overlapping units such as n-grams, word sequences and word pairs [15]. ROUGE-N compares n-grams based on recall as shown in equation (1).

$$ROUGE - N = \frac{\sum_{S \in (Reference\ Summaries)} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in (Reference\ Summaries)} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

The metric favours summaries that contain words shared by more references. This is intuitive as we usually prefer a summary that is close to the consensus among reference summaries. ROUGE-L counts in sequence unigram matches employing the concept of longest common subsequence between two texts [6]. It naturally captures sentence-level structure, but n-grams must be consecutive [15]. In [15], it was found that ROUGE-2 and ROUGE-L worked well in single document summarization tasks, whereas ROUGE-1 and ROUGE-2 worked reasonably well for multi-document summarization tasks when stop words were excluded.

2.2 Training for low resource languages

Most NLP research focuses on 20 of the 7000 languages of the world as they lack valuable training attributes such as supervised data, number of native speakers or experts [16]. Most deep learning methods require a large amount of expensive manually labelled data, which hinder their usefulness in domains with limited data. Languages with limited data are often referred to as low-resource languages (LRL). There are strong economic incentives to research LRLs as Africa and India host around 2000 LRLs and are home to more than 2.5 billion inhabitants [16]. Supporting languages with NLP tools can prevent their extinction and open knowledge of original works [17].

The deep learning-based neural attention model performs well when applied to abstract text summarization [18] compared to standard analytical learning-based approaches. In general, neural summarization is solved by using an encoder-decoder architecture with recurrent neural networks or self-attention [18]. However, there is an inherent limitation to natural language processing tasks such as text summarization for resource-poor and morphological complex languages owing to a shortage of quality linguistic data available [19]. Recent works show that synthetic data for machine translation tasks can significantly increase the quality of the result [20].

Models that can leverage linguistic information from unlabelled data can provide a valuable alternative to expensive and time-consuming manual data annotation [21]. Furthermore, learning meaningful representations in an unsupervised fashion can significantly improve performance. Several strategies exist from which in [21] it is shown that significant gains can be made in a wide range of NLP tasks by generative pre-training of a language model on a diverse corpus of unlabelled text and discriminatively fine-tuning on each task. The proposed approach uses task-aware input transformations during fine-tuning to achieve effective transfer while not requiring changes to the model architecture. The feature-based process utilized by ELMo uses task-specific architectures that use pre-trained representations as additional features [22].

2.2.1 BERT Model

In [22], the fine-tuning-based approaches are improved by proposing Bidirectional Encoder Representations from Transformers (BERT). The model uses a masked language model pre-training objective, masking some of the tokens from the input and predicting them from the context. The model is trained to produce bidirectional representations from the unlabelled text using bidirectional contexts on all layers. BERT can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks [23].

Significant effort has gone into developing multilingual BERT models to support other languages. Multilingual BERT models trained on several languages, including Estonian, can outperform all baseline models in Estonian for several NLP tasks [24]. In [25], it was found that XLM-RoBERTa achieved the highest results compared to other models. Several language-specific BERT models have been trained, such as Camem-BERT and FlauBERT, which have shown improvements over multilingual BERT models [25]. In

[24], an Estonian language-specific BERT model was proposed. The evaluation showed that the model outperforms multilingual BERT in most NLP task and proves the usefulness of language-specific models.

BERT was primarily developed for encoding text representations as an encoder only architecture. In [26], a Transformer-based sequence-to-sequence model allows to combine pre-trained BERT, GPT-2 and RoBERTa checkpoints. The model uses an encoder and decoder that both compose of Transformer layers. This results in new state-of-the-art results in tasks such as machine translation and text summarization [26]. Several combinations of model initializations can be used, such as BERT2BERT, a BERT-initialized encoder and decoder with randomly initialized encoder-decoder attention. RoBERTaShare has shared parameters between RoBERTa checkpoint initialized encoder and decoder, significantly reducing the model memory footprint [26].

2.2.2 BART model

Another self-supervised method is BART, a denoising autoencoder for pre-training sequence-to-sequence models [25]. BART is trained by firstly corrupting text with an arbitrary noising function and secondly learning a model to reconstruct the original text. According to [25], the model has good performance when fine-tuned for text generation and works well for comprehension tasks. The paper shows that the model matches the performance of RoBERTa on GLUE (General Language Understanding Evaluation) and SQuAD (Standford Question Answering Dataset) and achieves state-of-the-art results for abstractive dialogue, question answering and summarization tasks.

3 Related Work

3.1 Speech Summarization

Speech summarization requires directly processing audio streams and providing snippets to produce the summary. Spoken media primarily contains spoken-word content, but summarization can be performed in the text domain of the episode transcript. Automated speech summarization has many open research problems regarding multi-party speech, spontaneous speech and handling disfluencies [27]. Although high recognition accuracy can be easily obtained for speech, read from a text, such as anchor speakers' broadcast news utterances, the technological ability for recognizing spontaneous speech is much harder [5]. Spontaneous speech is often grammatically ill-formed and vastly different from written text. Spontaneous speech usually includes redundant information such as disfluencies, fillers, repetitions, repairs and word fragments. In addition, irrelevant information contained in a transcription caused by recognition errors is commonly inevitable. Therefore, an approach in which all words are transcribed is not an effective one for spontaneous speech. Instead, speech summarization that extracts essential information and removes redundant and incorrect information is necessary to recognize spontaneous speech. Efficient speech summarization saves time for reviewing speech documents and improves document retrieval [28].

To generate an automatic transcript for Estonian speech, a system for semi-spontaneous speech, such as broadcast conversations, lecture recordings, and interviews in diverse acoustic conditions, has been described in [29]. The system is trained multi-conditionally with various background noises. Robustness is increased by using a phoneme n-gram based decoding subgraph and FST-based phoneme-to-grapheme model to recover out-of-vocabulary words. The system achieves a word error rate of 8.1 [29] and performs punctuation recovery and speaker identification.

3.2 Podcast Summarization

In [27], the PodSumm method was described for obtaining podcast summaries by first transcribing the spoken content of a podcast, identifying meaningful sentences and finally stitching together the audio segments. The technique addresses speech summarization issues by posing it as a multi-modal data summarization problem with guidance from the text domain. In [30], abstractive podcast summarization was solved by taking the transcript, filtering the redundant sentences using a hierarchical attention model, and applying a BART fine-tuned system using a sequence-level reward function. BART and other pre-trained models use absolute positional embeddings, limiting the input sequence to 1024 tokens. The transcript of a podcast can be longer, which results in loss of data due to truncation. Experimentation in [30] showed that expanding the positional embeddings is inefficient and utilizing vanilla BART is recommended.

3.3 Abstractive Summarization Task

Transformers are large scale models pre-trained on massive text corpora with self-supervised objectives and fine-tuned on downstream tasks [31]. They have achieved state-of-the-art performance on a variety of summarization datasets [25], [32], [33] and are promising candidates for zero-shot and low-resource summarization [34]. Transformers with self-supervised objectives on large text corpora have shown outstanding performance when fine-tuned on text summarization and other NLP tasks. However, pre-training objectives for abstractive text summarization are less explored. In [34], a method for pre-training large Transformer-based encoder-decoder models on massive text corpora with a new self-supervised objective called PEGASUS. Important sentences are masked from an input document and are generated together as one output sequence. Evaluation has shown state-of-the-art performance on all datasets measured and surprising performance on low-resource summarization, surpassing previous results with only 1000 examples [34].

Despite the advances in the domain, several research challenges need to be pursued. Most Seq2Seq-based (Sequence-to-Sequence-based) summarization models rely on beam-search to generate summaries, but sampling-based approaches prove helpful in recent works. Since they increase the diversity of the generated texts, they have achieved success in open-ended language generation [35]. Summarization models are widely trained and

evaluated on news corpora [36], promoting journalistic writing style. The models tend to favour extraction rather than abstraction with leading paragraphs of most news articles as summaries [36], [37]. Several datasets from other domains have been introduced to alleviate this issue; likely, more will be released in the future to build better abstractive summarization systems [31], [37]. Most automatic evaluation protocols, including ROUGE, do not sufficiently evaluate the overall summary quality, ignoring features like factual correctness, fluency and relevance [38], [39]. Attempts have been made for generic text generation that better agree with human evaluation [40], [41].

3.4 Long-document Summarization

Neural summarization models are generally trained on short documents such as news articles. The solutions do not scale with extended media such as podcasts (often several thousand tokens long). A solution proposed by [42] extends the BART model with state-of-the-art performance on summarization tasks such as with the CNN/DailyMail dataset [43] and fine-tunes it on the TREC (Text Retrieval Conference) podcast transcript dataset [44]. To consider the entire input document, the attention layers are replaced with the attention mechanism used by the Longformer [45], with its linearly scalable global and local windowed attentions. The method effectively captures the entire long transcript and outperforms the descriptions made by the podcast creators when evaluated [42].

Working with long documents in a challenging, low-resource setting such as legal briefs [43] proposes an extract-then-abstract pipeline. The pipeline compresses long documents by identifying salient sentences with an algorithm that operates in the low resource regime, using a GPT2 perplexity and BERT classifier. The compressed documents are fed to a BART model, after which a 6.0 ROUGE-L improvement was observed, beating several competitive salience detection baselines and agreeing with independent domain expert labelling.

3.5 Low Resource Summarization

Training abstractive summarization models typically requires large datasets, which is not practical in most industry use-cases. Literature shows that when conducting pre-training for generative models, the effectiveness of pre-training is correlated with the similarity of the target domain and pre-training data known as the domain shifting problem [46], [47].

When encountering novel tasks, overfitting is a possibility when lacking high quality labelled examples. In [47], two knowledge-rich sources are proposed which can be utilized to tackle this problem, which are large pre-trained models and diverse existing corpora. Incorporating various corpora into training can help discover standard syntactic or semantic information and improve generalization ability. The author outlines that the approach achieves the state-of-the-art on six corpora in low-resource scenarios, with only 0.7% of trainable parameters compared to other works [47].

Three different methods for second phase pre-training were studied by [46], which are Source Domain Pre-Training (SDPT), Domain-Adaptive Pre-Training (DAPT) and Task-Adaptive Pre-Training (TAPT). Experiments showed that SDPT and TAPT could improve the performance of the fine-tuning method, whereas DAPT effectiveness depends on the similarity of the task and pre-training data. The author proposes the RecAdam method to alleviate the TAPT catastrophic forgetting issue and further boost its performance. In [48], explored to improve the performance of abstractive summarization task for small corpora of student reflections. The proposed approach consists of a novel template-based model to synthesize new data and domain transfer by tuning the model with newspaper data. Improved ROUGE performance was achieved compared to the word replacement synthesis baseline and the abstractive summarization model.

Starting with the BERT model, many new model variations have been developed in various languages [49]. The leading models have been increasing in size and have culminated in the 175-billion parameter GPT-3 model trained with around 45TB of data [50]. There is a rapidly growing amount of research on the use of multilingual models showing that they tend to be competitive with language-specific models, especially languages with smaller datasets that can benefit from the transfer effects from related languages with larger datasets [51]. A solution proposed by [52] is to transfer data to a more researched language such as English, enabling the use of large models to solve a specific task. The author shows that it is possible to reach better performance in specific tasks than language-specific models by translating data to English and using a large pre-trained English language model.

Improvements have been made to headline generation models for smaller datasets by enabling pre-training all model parameters and utilizing all available texts [53]. Neural

headline generation is a subtask of text summarization but much shorter and with a specific style [54]. The experiments were done on Estonian datasets showing that pre-training, in general, is beneficial, improving PPL by 29.6 – 32.4% and ROUGE by 0.85-2.84% [53].

3.5.1 Multilingual Tasks

The availability of datasets for multilingual text summarization is generally limited, and such datasets are expensive to construct. In [55], an abstract text summarizer for the German language using the "Transformer" model with an iterative data augmentation approach using synthetic data along with actual summarization data. The Common Crawl German dataset is used to generate synthetic data. Data augmentation is effective in low-resource scenarios where data is limited and where neural models require a large amount of training data. The system achieves an absolute improvement of 1.5 and 16.0 in ROUGE-1 F1 on the development and test sets compared to a system that does not utilize data augmentation [55]. A Quality-Diversity Automatic Summarization (QDAS) model enhanced by sentence2vec and applying transfer learning was proposed by [56] to tackle a multilingual low resource Wikipedia headline generation task. First summaries containing essential information are extracted from the article, after which a sequence labelling model using a pre-trained language model is applied for picking up key entry phrases [56].

4 Methodology

4.1 Summarization System Architecture

The method proposed by this work comprises of a sequence of steps, starting with the audio file and resulting in a summary as described in Figure 5. As the first step, the original audio is transcribed by automatic Speech recognition (ASR), which produces a transcript. A fine-tuned text summarization neural model is used to generate the final abstract summary. A data collection tool is used for building the corpora used for training the neural model.

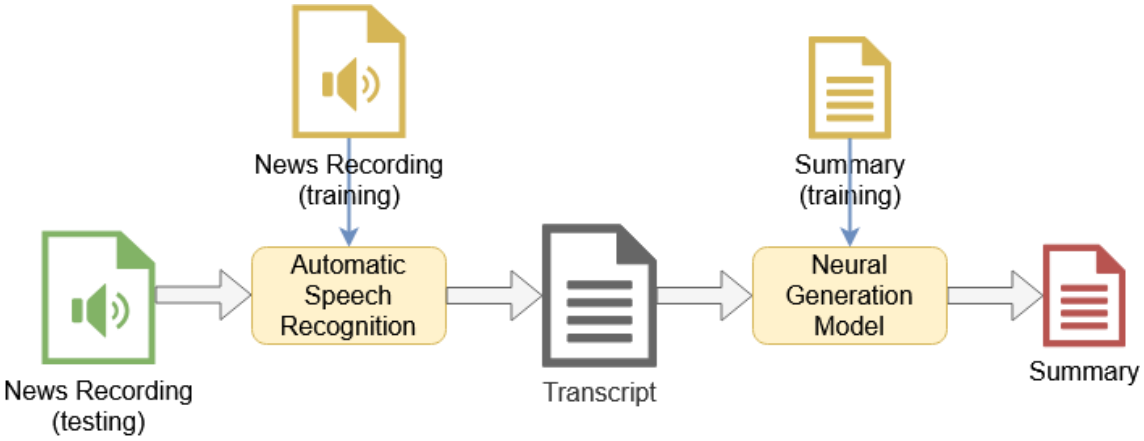

Figure 5. Summarization System Architecture.

4.1.1 Automatic Speech Recognition

The automatic speech recognition system performs audio transcription enabling the summarization to be solved in the text domain. Recent improvements to the Tallinn University of Technology Estonian speech transcription system described in Chapter 3.1 are leveraged to reduce errors and improve summary generation. The speaker identification feature of the system is not used for this work. The ASR system outputs a JSON file with the transcript split into speaker turn segments used downstream in the system.

4.1.2 Text Processing

The transcript file contains entries not used in the system, such as information about the timestamps of the words, confidence score, speakers' names, individual words and punctuation marks. In the text processing part, the speaker turns are assembled into one transcript. Transcripts that are shorter than three words are not good candidates for training and are filtered out.

4.1.3 Text summary generation

The system performs summarization using pre-trained BERT and BART models described in Chapter 2. Several models and configurations are evaluated in Chapter 5 to find the optimal model for the task. Firstly, experiments are conducted on BERT based models. The following pre-trained models are implemented and compared:

1. Multilingual BERT model trained with 104 languages in Encoder-Decoder configuration.
2. Estonian language BERT trained on the Estonian National Corpus [57]
3. XLM-RoBERTa trained on 2.5TB of CommonCrawl data.

Finally, the BART model with a language modelling head is implemented for summarization. The models are fine-tuned on the news dataset with handwritten summaries described in Chapter 4.2.

4.2 Dataset Creation

To train the neural model, large numbers of speech recordings must be available in the target language. As the training is supervised, every recording needs to be annotated with handwritten summaries. Nowadays, many platforms are available that distribute open-source multimedia. For speech summarization, the likely candidate sources are YouTube, broadcast news archives, podcasts archives and parliament session recordings. The ERR (Estonian Public Broadcasting) archive was found to be the most reliable resource for a large quantity of well-annotated news broadcasts for the Estonian language. The platform consists of audio and video archives, including various types of content.

The "Uudised" daily news show archive is used with around 9000 annotated episodes to date. This dataset was selected for its well-formed and short structure (around 2-minute

episodes). Each episode consists of the radio anchor reading one important news story of the day and usually includes at least one relevant interview. The show does not contain advertisements which simplifies the task, as these should be removed or ignored in the pipeline. The task of generating summaries for more free form shows with advertisements is an area that would benefit future research.

The following data points are extracted from the ERR website for each episode with an automated tool:

- The show in mp4 format.
- Metadata such as the broadcast date, recording length, guests and URL.
- Headline.
- Summary.
- Identification number.

The mp4 file is transcribed with the ASR system and added to the dataset. With visual dataset evaluation anomalies and problems with the summary, structures are found, and necessary filters were added to the pre-processing step. This is important to improve the system quality and normalize the data. For example, some summaries contain the broadcasting date and news anchor name or some characters that are not needed. In addition, missing punctuation marks are added, and unwanted characters such as line breaks are cleaned for well-formed sentences. The dataset consists of 5948 data points, each containing the following facts: episode id, generated transcript, summary and headline with a sample given in Figure 6. Full-length additional data points can be seen in Appendix 2. The dataset is split into training, testing, and evaluation corpora with 80%, 10%, and 10% ratios.

```
{"id": 5760,
"transcript":
      “Riigieelarve juures on teadagi oluline, millised on prioriteedid, mille
      peale raha kulutada ning millised ja kui suured on maksud, kust see raha
      saadakse.
      ...
      Nagu Kadri Simson juba ütles, eesmärgi saavutamine ei ole hoolimata east
      majanduskeskkonnast sugugi lihtne sest kõik koalitsioonierakonnad on aru
      saanud, et maksutõusust tuleb loobuda. Uued maksukavad, puudutagu need
      siis suhkrut või autosid esialgu unustada.”,
"summary":
      “Koalitsioonierakonnad valmistuvad riigieelarve strateegia aruteluks.
      Üksmeelsed ollakse selles, et miinuses riigieelarvet ei tohi järgmiseks
      aastaks teha.”,
"name":
      “Koalitsioonierakonnad järgmise aasta riigieelarvest.”}
```


Figure 6. ERR dataset example datapoint.

4.3 Transfer Learning

To increase multilingual models general understanding of the Estonian language, second phase pre-training experiments are conducted with the largest national dataset, the Estonian national corpus [57]. The corpora consist of Estonian articles, periodicals, blogs, Wikipedia and web pages. The dataset help to give a general understanding of the language. An example data point is shown in Figure 7. Estonian corpus sample data point and full-length example data points can be seen in Appendix 3. The additional training is only necessary for multilingual models, as the Estonian BERT is already pre-trained on the given corpus.

```
<doc id="1" src="Reference Corpus" filename="aja_EPL_2002_02_22.ma"
texttype_nc="periodicals" newspaperNumber="Eesti Päevaleht 22.02.2002"
heading="Majandus" title="Hotell Tallinna kõrval asuv auk saab
detailplaneeringu" texttype="Journals" texttype_src="source data">
      Tallinna linn algatab Paldiski maantee ääres Hotell Tallinna kõrval
      asuva suure vundamendiaugu ja tühermaa detailplaneeringu koostamise,
      ehitustööde alustamist takistavad aga ala segased omandisuhted.
      ...
      Seejärel tekkisid ehitajal kohtuvaidlused ühe kinnistu õigusjärgse
      omanikuga ja kogu ehitus jäi seisma. Vundamendiauk on krundil siiani
      alles.
</doc>
```

Figure 7. Estonian corpus sample data point.

Chapter 3.5 described the solution of machine translating task dataset into English to leverage larger models. A large English model can be used for a low resource language with an architecture described in Figure 8. We generated a corresponding English dataset (Example data point in Figure 9. Translated ERR corpus sample data point. and full-length examples in Appendix 4) for our news corpora, using the Neural Machine Translation Cloud API with its state-of-the-art performance. Dataset translation is to be executed before training. The pre-model generates the summaries, after which outputted text is machine translated back to the target language, giving the final results.

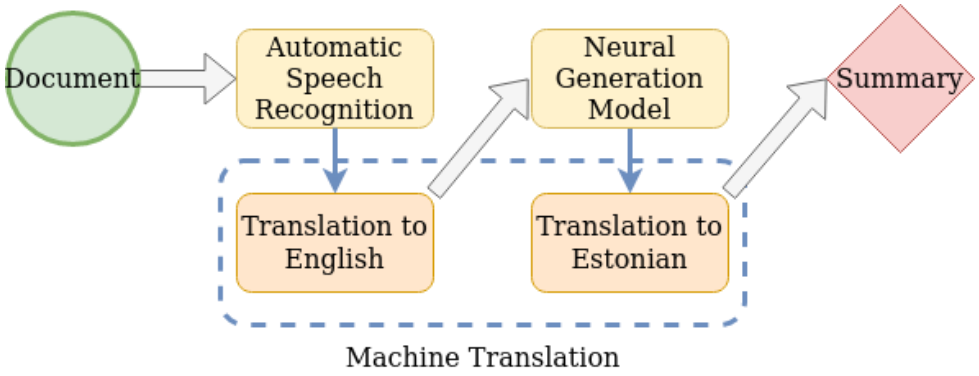

Figure 8. Machine translation summarization architecture.

{"id": 5760,
"transcript":
"Of course, in the case of the state budget, it is important what are the priorities, on which to spend the money and what and how big are the taxes, where does this money come from.
...
As Kadri Simson has already said, achieving the goal is not easy at all, regardless of the economic environment, because all coalition parties have understood that the tax increase must be abandoned. New tax schemes, whether for sugar or cars, will be forgotten for the time being.",
"summary":
"Coalition parties are preparing for the discussion of the state budget strategy. There is a consensus that the minus state budget should not be made for next year."}

Figure 9. Translated ERR corpus sample data point.

4.4 Data Augmentation

To improve the performance of the deep neural network, the amount of data available for training needs to be increased. As there are no large-scale datasets available for Estonian news broadcasts, considering that the ERR dataset consists of less than 6000 data points and gathering more data is not practical, other methods are explored. After transcribing the recordings, the summarization is done in the text domain, similar to article summarization tasks. Open-source datasets such as the large English CNN/Dailymail [58] can be utilized in combination with machine translation. To expand the limited data in the domain, a subset of CNN/DailyMail dataset [58] machine translated to Estonian containing about 9000 articles and summaries (sample in Figure 10 and full-length samples in Appendix 5). The exact translation method is used as in the last Chapter, and translation takes place before training.

{“id”: “0044e296ecfe3ba57a351ad2a36d034491e878ce”,
“article”:
“(CNN) Laulja ja laulukirjutaja David Crosby lõi pühapäeva õhtul autoga sörkjooksu, ütles pressiesindaja.
...
Esialgsete teadete põhjal tunneb ta kergendust, et härra vigastused ei olnud eluohtlikud," ütles Crosby esindaja Michael Jensen. "Ta soovib sörkjooksjale väga kiiret taastumist.”,
"highlights":
“Õnnetus juhtub Californias Santa Ynezis, kus asub Crosby. Sörkjooksja sai mitu luumurdu; tema vigastusi ei peeta eluohtlikeks.”}

Figure 10. Translated CNN/DailyMail sample data point.

4.5 Baseline Implementations

Two baseline summarization methods are implemented and evaluated. This includes a rule-based approach that takes the first sentence of every transcript as the summary. The method is based on the idea that the news show introduces the main points or outline. Secondly, the LexRank extractive method that is described in Chapter 2.1.1 is compared. Neural generation methods are benchmarked against the mentioned techniques to highlight the performance improvements.

4.6 Evaluation Metrics

Summary quality and accuracy are measured with the ROUGE metrics as described in Chapter 2.1.4. More specifically, ROUGE-1, 2 and L measures are used for comparisons. Additionally, for human evaluation, three samples for every method are generated and graded on a five-point scale as follows:

- 1 – Unintelligible.
- 2 – Confusing.
- 3 – Fair.
- 4 – Good.
- 5 – Very Good.

Grading was conducted as a public survey by 25 individuals, and results are averaged. During grading, the generated summaries for an episode were shown in a randomized order, and the related models were not disclosed. Experiments and improvements are made iteratively and compared with the performance of the baseline models.

5 Experimentation Results

To process the data and implement models PyTorch Python framework is selected. The texts vocabulary is built after which words are tokenized and vectorized for training neural networks. The dataset was split into training, evaluation and test datasets before starting training. To fairly select the optimum pre-trained models and training techniques for the summarization system, we select the most popular Hugging Face models (Table 1). To enhance the model's summarization performance, relevant datasets are chosen for the models according to methods selected in Chapter 0. For each considered model, the training and sampling parameters are not changed, making the main difference the architecture and training data being used. The generated outputs of each model are evaluated on the same ERR corpora subset, with the ROUGE scores and examples being captured for manual evaluation. The reference summaries from Table 2 are used as an example to visualize the model outputs.

Table 1. Models used in the experiments and their datasets.

Model name in Hugging Face	**Langage**	**Datasets Used**
xlm-roberta-large	Multi	ETNC, Trans. DailyMail, ERR
bert-base-multilingual-cased	Multi	ETNC, Trans. DailyMail, ERR
tartuNLP/EstBERT	Estonian	DailyMail, ERR
facebook/bart-large-cnn	English	Trans. ERR

Table 2. Example Summaries References.

Nr	Reference
1	Koalitsioonierakonnad valmistuvad riigieelarve strateegia aruteluks. Üksmeelsed ollakse selles, et miinuses riigieelarvet ei tohi järgmiseks aastaks teha.
2	Hiljem teatas Tartu Ülikool, et valitsuse pressikonverentsil jagati vananenud andmeid.
3	Alates 1. septembrist on Tartu Tasku keskuses suletud kinoketi Cinamoni kino, mis avati keskuses 2008. Cinamoni lahkumisega väheneb tartlaste jaoks võimalus kesklinnas linateoseid nautida, sest tõenäoliselt uut ja suurt kino Tartu südalinna lähiajal ei teki.
4	Eesti, Läti ja Leedu valitsusjuhid allkirjastasid Tallinnas Rail Balticu kokkuleppe. Kokkuleppega pannakse muu hulgas paika raudtee rajamise tähtajad, selle kulgemine ja mitmed tehnilised detailid.
5	Briti rahandusminister Rishi Sunak esitleb täna lisaeelarvet, mille eesmärk on taaselustada koroonaviiruse tõttu kannatanud majandust. Plaan peaks toetama rohkem kui 100 000 adat tuhandet töökohta rohemajanduses, aga on saanud ka kriitikat.

As the pre-trained models have been trained with fixed size embedding, the encoder dimensions have been selected to be 512 and decoder dimension to be 128, with an exception for the BART model that supports larger embedding sizes. Top-k sampling is used and configured to top 50 results. For fine-tuning, the learning rate is set to 2e-5, weight decay to 0.01 and training epochs range from 8 to 64. Seq2SeqTrainingArguments "predict_with_generate" flag is set to true to calculate generative metrics. The training checkpoints and fine-tuned models are saved for future use. For evaluation, a custom function for computing metrics is created and passed as an argument to the trainer. The function decodes predictions and labels, converts each sentence to a new line for rouge calculation and finally computes the metrics. The Seq2SeqTrainer from the Transformer's library is used to configure and conduct fine-tuning.

5.1 Baseline Implementations

At the start of the experimentation phase, baseline implementations described in Chapter 4.5 were tested. For the LexRank implementation, Estonian stopwords compiled by [59] are imported. The stopwords are needed such that words that do not add much meaning to the sentence, such as the, he, have are not taken into consideration while scoring. The sentences are split as the library takes a list of sentences for every text. The first sentence model extracts the first sentence of every text as the summary. To tokenize and split the

sentences, the EstNLTK [60] toolkit is used, and the computing of rouge scores are done with the Hugging Face datasets library.

The results in Table 3. Baseline systems ROUGE scores show that the most basic first sentence rule-based and LexRank had better performance with a ROUGE-L score of 10.14. The manually graded score is computed as described in Chapter 4.6. The manual evaluation finds the first sentence model summary quality to be good and LexRank to be confusing.

Table 3. Baseline systems ROUGE scores.

Model	**Rouge 1**	**Rouge 2**	**Rouge L**	**Human**	**Avg. len.**
First sentence	12.03	3.45	10.14	3.80	23
LexRank	10.88	2.86	9.00	2.40	23

As the system extracts whole sentences from the original text, the sentences are guaranteed to be coherent, as shown in the examples given in Table 4 and Table 5. On the other hand, summarization usually requires rephrasing, which means that the summaries with this system are not the most informative about the text's main points. Example number one has the same summary for both systems. The result is not general enough while acknowledging the construction of the power plant, focusing only on the negative aspects and not the impacts. The second and third example has in both cases a specific extract from the interviews during the show and not describing the event in general. From the samples (Table 5), the LexRank system does not always choose the first sentence of the transcript. The system selects the best number of sentences mathematically, but experimentation results mostly in summaries consisting of one sentence.

Table 4. First sentence example summaries.

Nr	First sentence output
1	Riigieelarve juures on teadagi oluline, millised on prioriteedid, mille peale raha kulutada ning millised ja kui suured on maksud, kust see raha saadakse.
2	Majandusminister Taavi Aas andis täna valitsuse pressikonverentsil hoiatuse Viljandi ja Haapsalu elanikele.
3	Cinamon Group põhjendab oma kodulehel Tartu kinokeskuse sulgemist lausetega, et kino on küll pidevalt täiustada tööd, aga pärast korduvalt nurjunud katseid leida koostööd Tasku kaubanduskeskusega ei jäänud muud võimalust, kui lõpetada taskus kino opereerimine ja sulgeda filmikeskus lõplikult alates septembrist.
4	Täna allkirjad saanud lepe reguleerib ehitatava taristu ning selle aluse maa omandiküsimusi, samuti rahastamise tingimusi.
5	Briti noore rahandusministri tänane avaldus mõjutab pea kõiki majapidamisi ja tema enda renomeed riigi majanduse kaitsmiseks koroona kriisi kõrgajal antud toetused, subsiidiumid ja laenud ei suutnud säilitada töökohti.

Table 5. Lexrank example summaries.

Nr	Lexrank output
1	Loomulikult 2019 aasta riigieelarvele mõeldes ja riigieelarvestrateegiale järgmisel neljal aastal me kindlasti peame seadma eesmärgi vähemalt tasakaalus eelarve.
2	Ja, ja see meetod on, on nagu hästi tundlik, nii et kui seal on juba mingi teatud hulk üle üle arvatava või üle määratu vampiiri, siis me juba seda viiruse RNA-d, näeme nüüd, mida see nüüd täpselt tähendab.
3	Teisisõnu kasvaks Apollo kino mõjuvõim Tartu turul veelgi.
4	Tänasel pressikonverentsil küsiti peaminister Jüri Ratasele, missugune on garantii, et projekti sellises mahus rahastamine Euroopa Liidu vahenditest jätkub ka järgmisel Euroopa Liidu eelarveperioodil.
5	Briti noore rahandusministri tänane avaldus mõjutab pea kõiki majapidamisi ja tema enda renomeed riigi majanduse kaitsmiseks koroona kriisi kõrgajal antud toetused, subsiidiumid ja laenud ei suutnud säilitada töökohti.

5.2 BERT based systems

The pre-trained BERT based models are loaded and trained using hugging face Transformers and PyTorch python libraries. BertTokenizer and EncoderDecoderModel from the Transformers library are used for loading and configuring models. The following models are used: bert-base-multilingual-cased, xlm-roberta-base, tartuNLP/EstBERT.

The models are configured with a maximum encoder length of 512 and a maximum decoder length of 128 tokens. The beginning of the string token is set to the BERT CLS token and the end of string token to the SEP token accordingly.

5.2.1 Fine-tuning models

The first experiment is to fine-tune the pre-trained models with the ERR dataset described in Chapter 4.3 for eight epochs. The best system is the RoBERTa model with a ROUGE-L score of 4.87 (Table 6). The multilingual BERT system achieves in comparison 40% and the estBERT 10% lower score (Table 9). The best baseline result, however, outperforms the RoBERTa system by 20%. The reference summaries for comparing the generated summaries are given in Table 2. Examining the mBERT generated examples from Table 7, it can be noted that some words are related to the topics, but the sentences do not have a coherent concept. The RoBERTa system generated much more coherent sentences (Table 8), with the main points being understandable. However, repetitions and odd wording can be found. The manual scoring yielded similar comparatively similar results, with the multilingual BERT summaries being unintelligible and others confusing.

Table 6. BERT Based Systems ROUGE Scores.

Model	**Rouge 1**	**Rouge 2**	**Rouge L**	**Human**	**Avg. len.**
bert-base-multilingual-cased	5.83	0.41	4.87	1.68	46
xlm-roberta-large	9.61	1.61	8.14	2.39	65
tartuNLP/EstBERT	8.44	1.36	7.28	2.09	26

Table 7. Multilingual BERT example generated summaries.

Nr	**mBERT output**
1	Valitsust järjest tulnud seaduse muutmist jõudis kokku 30.
2	Täna arutati koroonaviiruse leviku haigestumist Eestis. Eesti teraviiruse tõusu muutmise alusel ei ole uuringu, kuid see näitab Eesti kodanikest ja keskelt.
3	Pärnu linnaeelarve saab kinnitama kutsuval käispunktist. Tartu linnavalitsus teatas, et tänavatele on lahkub ka projekti.
4	Euroopa Liidu erakorralise liikmesriikidele on seotud lähiajalise arengu tõttu eelnõu, milles võib olla võimalik koroonaviirusega kaubanduskorralduse üle, kas valitsus ei soovenda, kuid seda ei ole, või mitte nii valitsusliid.
5	Riigikogus arutati kokkuleppele ettepaneku, et märgita Eestilt oma turgutestuse vähendanud tasvu ja tuleva aastaga tuvastada vähendanud energia.

Table 8. XLM-RoBERTa example generated summaries.

Nr	XLM-RoBERTa output
1	Riigikogu alustas koos Riigikogu arutamist järgmiseks aastaks, millest riigieelarve eelnõu saab järgmisel aastal eelarveks ligi 19 miljonit eurot.
2	TNS EÜS andmeil on koroonaviirus Eestis seni ära jäänud. Viimase viiruse kandmise juhtumi uurimisega tegeles rahandusministeerium.
3	Ka uues kinos Sõpruse kinomajas lõpetatakse sel nädalavahetusel kinokino laiendus. Keskuses on uksed juba üleval juba täna, kuigi omaniku kinnitusel sulgeb keskus alles järgmisel nädalal.
4	Euroopa Liidu riikide vaheliste erimeelsuste tõttu suletud Euroopa Komisjon teatas, et alustab Eesti-Vene ühistöid ka Eesti poolt, teatas Euroopa Liit. Euroopas on neid juba enam kui kümne aasta jooksul kokku võetud.
5	Briti rahandusministrid kogunesid järgmise nädala algul valitsusele arutama majanduse taaskäivitamise meetmeid. Rahandusminister Donald Trumpi sõnul on esimene samm aga paratamatult jõustumas, mida ta loodab ellu viia eelarvega võrreldes kogu Eestit lähitulevikus.

Table 9. EstBERT example generated summaries.

Nr	estBERT output
1	Valitsus otsustas teha lisaeelarvesse järgmise nelja aasta jooksul.
2	Praegu arvati, et Lihula on viirusest põhjustatud viirusepuhangu tõttu hukkunud kaheksa inimest. Inimeste sõnul aga on praegu veel lootus, et sügisel saaks Saaremaal viiruse kahtluse alla.
3	Tartu kesklinna uus kino hakkab avama sel nädalal kaks korda uut kinosaalide arvu täis läinud nädalal. Täna õhtul alustavad filmistuudio kinosnamataliku kinoga vaid kaks kino ning ühe kaubamüügiga saavad endiselt taas oma uksed avada. Altuse kinnitusel pole esialgu veel selge, kas kas ja kui palju uus kino tööle hakkab.
4	Euroopa Komisjoni analüüsi kohaselt tehakse Rail Balticu juures Rail Balticule veel üks tugi Rail Balticu jaoks. Rail Balticu ehitus läheb maksma miljon eurot. Eesti riigilt ja Läti kogemusest tulenevatele rööbastele loodab Railron jõuda Rail Balticu Euroopa Raudtee ehitamiseks.
5	UKA peaminister Donald Trump astus peaministri kohalt tagasi, lubas president pärast eilset eilset õhtul välja kuulutatud uue eelarve. Kaba on rekordväike, isegi kui praegu seda peetakse. Osa valitsust kui suure osa ettevõtete liikmeid on hädas sellega, et valitsuse rahanduspoliitika maksab ennast praegu ja ootab suuremat abi palgatõusuks. Paldrid hääletasid vastu ja hääletasid 40 %.

5.2.2 Data augmentation

The models from the previous chapter do not perform as well as expected, and improvements need to be made. The mBERT and RoBERTa models are fine-tuned firstly for two epochs on the article's corpus described in Chapter 4.3. The estBERT model is already pre-trained on the corpus, and as such, this step is skipped. Fine-tuning is conducted with a machine-translated CNN/Dailymail dataset for 64 epochs, and finally, the ERR dataset is used for a final fine-tuning stage for 64 epochs. In total, three fine-tuning stages are conducted with an exception for the estBERT model.

The estBERT based system is the best performing with a ROUGE-L score of 10.88 (Table 10), which is 6.8% better than the baseline. However, the human evaluation found the summaries to be slightly worse, being fair on average. Multilingual BERT has a score of 9.6% and RoBERTa 4.2% lower compared to the estBERT results. The human evaluation found the mBERT summaries to be on the average confusing and RoBERTa to be fair.

Table 10. BERT Systems Data Augmentation ROUGE Scores.

Model	**Rouge 1**	**Rouge 2**	**Rouge L**	**Human**	**Avg. len.**
bert-base-multilingual-cased	11.51	2.86	9.86	2.21	44
xlm-roberta-large	12.07	3.35	10.43	2.54	57
tartuNLP/EstBERT	11.72	3.13	10.88	3.46	27

Examining the example outputs (Table 11, Table 12 and Table 13), text generation confuses some words lowering the human score. For the estBERT system, the general cohesion of the sentences is good, and the essence of the sentences are understandable; however, some words are not appropriate - in example number two, a gibberish word was generated.

Table 11. Augmented multilingual BERT example generated summaries.

Nr	mBERT output
1	Valitsus tegi kokkuvõtteid riigikogu menetlusest järgmisel aastal.
2	Valitsus loodab peatada koroonaviiruse tõttu tehtud reoveeprobleeme seoses Hiinast algavateks biomeetmeteks.
3	Tartu kinokeskus suleb juunis suletuks oma teatrikeskuse laiendatud kinokeskusega. Kuigi seni kasvas Kihnu teatrikeskus kui unikaalset ühenda, läks ka kinokeskuse sulgemine lõpuks selgeks.
4	Kolme Balti riigi vahelise püsiühenduse rahastamise projekt tekitab küsimusi.
5	Briti rahandusminister Boris Johnson saadab taotluse töötuna riigikogus elektroonikatrvale panna. Suurema osa oma seisukohast tekitab see seadusemuudatus ja töötuse määr kahekordistab ootamatult inimesedovooli.

Table 12. Augmented RoBERTa example generated summaries.

Nr	XLM-RoBERT output
1	Valitsus kinnitas tuleva aasta riigieelarve ning esitab riigikogule järgmise aasta eelarve eelnõu.
2	Kuigi Eestis on koroonaviirusesse haigetumine madal, pole nähagi, et need on eraldiseisvad. Ühel juhul pole toimumiskohtlejaid sellist teha.
3	Läti kinode omanikfirma Megateater teatas, et lõpetab Tartu kino uue kaubanduskeskuse ja Näitleteti kino Küberkeskus, mis on seotud Apollo filmikeskusega.
4	Täna allkirjastati leping Euroopa Liidu transpordiministrite ühiskohtumisel.
5	Briti rahandusministeerium avaldas täna otsuse, et pensionireform toimub ühes riigis. Muuhulgas pakkus ministeerium välja uue meediaseaduse eelnõu, mis lubaks investeerida koolitussemadesse miljonites.

Table 13. Augmented estBERT example generated summaries.

Nr	estBERT output
1	Peaminister Jüri Ratas allkirjastas koalitsioonilepingu.
2	Mustgaasidoon on nüüdseks taastunud, kuid selle valmimine sõltub sellest, kas erakonnad testivad koroonaviirusest tingitud puhangu mõju linnaelanikele.
3	Vabariigi Teaduskeskus ja Tartu Vabamäe Kinnelnel kavatseb alustada ajutist renoveerimist, sest selle abil leitakse üha enam uusi kliente.
4	Ametisse kirjutati alla Rail Balticu keskkonnamõju hindamise leping.
5	Briti peaminister Boris Johnson viis riigi majandusministri ametisse. Ministrite sõnul on tema eesmärk taastada usaldus riigis endiselt toimuva vastu ja viia ellu riigi töökohtade loomine rekordiliselt kõrgele.

5.3 BART Based System

The BART model is pre-trained in the English language; therefore, machine translation is used to translate the dataset described in Chapter 0. As the model is pre-trained for the summarization task and can generate summaries out of the box, only a short one epoch fine-tuning run is used. For evaluation, the compute metrics function is modified such that the predicted sentences are translated back to Estonian. The original Estonian validation dataset is used as a reference. As the BART model is pre-trained on larger embeddings, the encoder max length is set to 1024, and the decoder max length is set to 512. The sampling parameters are the same as in previous experiments, but the generated text length is set to 512. Training parameters have not been changed from the earlier experiments.

The model's ROUGE-L score is 30% higher (Table 14) than the baseline and the best performing system from previous experiments. The generated examples (Table 15) are very coherent, with some punctuation errors and wrong wording. The fifth example seems to be most problematic, with the last sentences being challenging to understand, the object being missed or incorrect in sentence three and four. The mistakes are mainly caused due to errors in two-way machine translation. The human scoring averaged 3.93 or good, the best score so far. Results of experiments conducted so far are given in Table 16.

.

Table 14. BART System ROUGE Scores.

Model	Rouge 1	Rouge 2	Rouge L	Human	Avg. len.
facebook/bart-large-cnn	23.67	7.49	15.92	3.93	96

Table 15. BART System Example Summaries.

Nr	BART
1	Selle aasta eelarvet on selle puudujäägi osas korduvalt kritiseeritud. Sotsialist Jevgeni Ossinovski sõnul tuleks järgmine aasta koostada nominaalse või tegeliku ülejäägiga. Kõik koalitsioonierakonnad on aru saanud, et maksutõusust tuleb loobuda. Uued maksuskeemid, kas suhkru või autode jaoks, unustatakse esialgu.
2	Koroonaviiruse jälgi on leitud Viljandi ja Haapsalu reoveeproovidest. Majandusminister Taavi Aas tegi viljandlastele hoiatuse. Bioloogiaprofessor professor Irja Lutsar ütleb, et proovide tulemuste põhjal pole veel võimalik järeldusi teha.
3	Cinamon Group sulges Tartu kinokeskuse septembris. Tasku keskuse aktsiaseltsi tegevjuht Annika Oja ütleb, et kinoketi lahkumine oli kahepoolne otsus. Cinamoni lahkumine tähendab ka seda, et Tartus on filmi vaatamiseks neli kohta. Kahe Apollo kino juurde kuuluvad nii Eedeni kui Lõunakeskuse filmikeskused, Elektriteater ja Forum Cinemas.
4	Täna allkirjastatud leping reguleerib ehitatava infrastruktuuri ja selle aluseks oleva maa omandilist kuuluvust. Projekti maksumus on hinnanguliselt viis miljardit eurot, millest Euroopa Liit saab aastaks 2020 rahastada kuni 85 protsenti. Esialgse ajakava kohaselt peaks raudtee valmima aastatel 2024–2025. Leping tuleb parlamentidel ratifitseerida. kõigist kolmest riigist ja tagab valitsuse heakskiidul lepingu kehtivuse ka valitsuse vahetuse korral.
5	Rahandusministeeriumi paberid näitavad puudujääki 300 miljardit naela. On aeg tagajärjed kõrvaldada. Sunaksina pakub ta raha suunamiseks viis punkti. Lühiajalised käibemaksu vähendamised, nii et kaks kuni kolm protsenti võivad inimesi meelitada veidi ostlema. Plaanis on ka riiklik vautšer pühadeks. Ja viiendaks tuleb odavaid laene, mida laenab riik.

Table 16. Experimentation aggregated ROUGE scores.

Experiment	**Model**	**Rouge 1**	**Rouge 2**	**Rouge L**	**Human**	**Avg. len.**
BART	facebook/bart-large-cnn	23.67	7.49	15.92	3.93	96
Augmented BERT	bert-base-multilingual-cased	11.51	2.86	9.86	2.21	44
	xlm-roberta-large	12.07	3.35	10.43	2.54	57
	tartuNLP/EstBERT	11.72	3.13	10.88	3.46	27
BERT	bert-base-multilingual-cased	5.83	0.41	4.87	1.68	46
	xlm-roberta-large	9.61	1.61	8.14	2.39	65
	tartuNLP/EstBERT	8.44	1.36	7.28	2.09	26
Baseline	First sentence	12.03	3.45	10.14	3.80	23
	LexRank	10.88	2.86	9.00	2.40	23

6 Discussion and Conclusion

The results for the experimentation show that BART-Large-CNN with machine-translated data outperforms the native and multilingual models with augmented datasets. The system achieves a ROUGE-L score of 15.92, which is better than our baseline and shows improvements in human evaluation. The model does not need fine-tuning as it is pre-trained for downstream summarization tasks, significantly reducing the time and resources required. Other models needed extensive fine-tuning with larger datasets to produce comparative results, and as such, the best performing estBERT model achieved a 32% lower ROUGE-L score. An unexpected observation is that the simple first sentence model generally performed well for the use case. This can be attributed to the fact that news broadcasts start typically by giving a brief outline of the stories covered. With other media or broadcast types, the method might not be effective, and abstractive models are more relevant. With smaller datasets without any data augmentation, the BERT based models have relatively low performance, with XLM-RoBERTa achieving a ROUGE-L score of 8.14.

The proposed approach is dependent on the existence of a high-quality machine translation solution for the target language. The testing with the Estonian language shows that the translation is adequate; however, languages that are typologically vastly different to English and do not have suitable translation models need to consider if the solution is applicable. In such cases, a national BERT model can be considered as it was the best performing solution that does not require machine translation. The multilingual XLM-RoBERTa model should be considered where national language models are not available achieving comparable performance. In other words, the choice for different language depends on the quality of the translations and the availability of large national language models in the target language.

In this work, a system for the news broadcast abstract summarization task under low resource conditions consisting of automatic speech recognition that produces a transcript, and a neural summarization model was proposed. Three possible models with their

training methods were presented. The BART model, which is pre-trained on CNN/DailyMail data and combined with target data machine translation outperformed the baseline and alternative models and achieved the best human evaluation scores. In target languages where machine translation systems are not as mature, national BERT and multilingual RoBERTa models should be considered.

7 Future Work

The system can be improved in future works, as the broadcast audio contains information that is not utilized in the system, such as the tone and other audio cues. Further research is needed in order for the ASR system to handle these nuances better. The ERR news broadcasts are short, well-formed and do not contain advertisements, which is not the case for certain types of media. Other longer and more open-ended media types, such as talk shows, podcasts, and meetings, have much more complexity and need further research. This type of speech can be much less structured with disfluencies such as interruptions, overlapping speech and filler phrases with speakers losing their trail of thought. One possibility is to experiment with adding an extractive summarization method after the ASR step in order to filter out uninformative utterances. This may be used as the input for the neural summary generation step, making it possibly much more robust for less structured media. The proposed BART outperformed baseline systems; however, some disfluencies were noted during the experiments due to the Estonian machine translation models. Further improvements to translation models for languages that are not similar to English, such as Estonian, are required.

References

[1] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165. Elsevier Ltd, p. 113679, Mar. 01, 2021, doi: 10.1016/j.eswa.2020.113679.

[2] C.-E. González-Gallardo, R. Deveaud, E. SanJuan, and J.-M. Torres-Moreno, "Audio Summarization with Audio Features and Probability Distribution Divergence," Jan. 2020, Accessed: Jan. 30, 2021. [Online]. Available: https://arxiv.org/abs/2001.07098.

[3] S. Furui, K. Iwano, C. Hori, T. Shinozaki, Y. Saito, and S. Tamura, "Ubiquitous speech processing," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2001, vol. 1, pp. 13–16, doi: 10.1109/icassp.2001.940755.

[4] S. Furui, "Recent Advances in Spontaneous Speech Recognition and Understanding." Accessed: Feb. 11, 2021. [Online].

[5] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-Speech and Speech-to-Text Summarization." Accessed: May 08, 2021. [Online].

[6] M. Allahyari *et al.*, "Text Summarization Techniques: A Brief Survey," *arXiv*, Jul. 2017, Accessed: Feb. 06, 2021. [Online]. Available: http://arxiv.org/abs/1707.02268.

[7] G. Erkan and D. R. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457–479, Sep. 2011, doi: 10.1613/jair.1523.

[8] A. Vartakavi and A. Garg, "PodSumm -- Podcast Audio Summarization," *arXiv*, Sep. 2020, Accessed: Feb. 06, 2021. [Online]. Available: http://arxiv.org/abs/2009.10315.

[9] Ian Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[10] I. Sutskever, O. Vinyals, and Q. v. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, Sep. 2014, vol. 4, no. January, pp. 3104–3112, Accessed: Feb. 14, 2021. [Online]. Available: https://arxiv.org/abs/1409.3215v3.

[11] K. Kurniawan and S. Louvan, "IndoSum: A New Benchmark Dataset for Indonesian Text Summarization," in *Proceedings of the 2018 International Conference on Asian Language Processing, IALP 2018*, Jan. 2019, pp. 215–220, doi: 10.1109/IALP.2018.8629109.

[12] L. Hou, P. Hu, and C. Bei, "Abstractive Document Summarization via Neural Model with Joint Attention." Accessed: Feb. 18, 2021. [Online].

[13] V. Gupta and G. Singh Lehal, "A Survey of Text Summarization Extractive Techniques," 2010, doi: 10.4304/jetwi.2.3.258-268.

[14] I. K. Bhat, M. Mohd, and R. Hashmy, "SumItUp: A Hybrid Single-Document Text Summarizer," in *Advances in Intelligent Systems and Computing*, 2018, vol. 583, pp. 619–634, doi: 10.1007/978-981-10-5687-1_56.

[15] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," 2004. Accessed: Feb. 20, 2021. [Online]. Available: https://www.aclweb.org/anthology/W04-1013.

[16] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource Languages: A Review of Past Work and Future Challenges." Accessed: Feb. 21, 2021. [Online]. Available: https://www.mtsummit2019.com/.

[17] Y. Tsvetkov, "Opportunities and Challenges in Working with Low-Resource Languages," 2017. Accessed: Feb. 21, 2021. [Online].

[18] A. R. Openai, K. N. Openai, T. S. Openai, and I. S. Openai, "Improving Language Understanding by Generative Pre-Training." Accessed: Feb. 21, 2021. [Online]. Available: https://gluebenchmark.com/leaderboard.

[19] M. E. Peters, M. Neumann, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, “Deep contextualized word representations.” Accessed: Feb. 21, 2021. [Online]. Available: http://allennlp.org/elmo.

[20] M. Chinea-Ríos And´alvaroand´ And´alvaro Peris and F. Casacuberta, “Adapting Neural Machine Translation with Parallel Synthetic Data,” 2017. Accessed: Feb. 28, 2021. [Online].

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Feb. 14, 2021. [Online]. Available: http://arxiv.org/abs/1810.04805.

[22] C. Kittask, K. Milintsevich, and K. Sirts, “Evaluating multilingual BERT for Estonian,” 2021. Accessed: Feb. 22, 2021. [Online].

[23] D. Nozza, F. Bianchi, and D. Hovy, “What the [MASK]? Making Sense of Language-Specific BERT Models.” Accessed: Feb. 22, 2021. [Online]. Available: https://github.com/google-research/bert/blob/master/multilingual.md.

[24] H. Tanvir, C. Kittask, and K. Sirts, “EstBERT: A Pretrained Language-Specific BERT for Estonian,” *arXiv*, Nov. 2020, Accessed: Feb. 14, 2021. [Online]. Available: http://arxiv.org/abs/2011.04784.

[25] M. Lewis *et al.*, “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” *arXiv*, Oct. 2019, Accessed: Feb. 14, 2021. [Online]. Available: http://arxiv.org/abs/1910.13461.

[26] S. Rothe, G. Research, S. Narayan, and A. Severyn, “Leveraging Pre-trained Checkpoints for Sequence Generation Tasks,” 2020. Accessed: May 09, 2021. [Online]. Available: https://github.com/openai/gpt-2.

[27] A. Vartakavi and A. Garg, “PodSumm: Podcast Audio Summarization.” Accessed: Jan. 30, 2021. [Online]. Available: https://spacy.io/usage/linguistic-features#sbd.

[28] K. Mckeown, J. Hirschberg, M. Galley, and S. Maskey, “FROM TEXT TO SPEECH SUMMARIZATION.” Accessed: Feb. 11, 2021. [Online].

[29] T. Alumäe, O. Tilk, and Asadullah, “Advanced Rich Transcription System for Estonian Speech,” *Frontiers in Artificial Intelligence and Applications*, vol. 307, pp. 1–8, Jan. 2019, doi: 10.3233/978-1-61499-912-6-1.

[30] P. Manakul and M. Gales, “CUED_SPEECH AT TREC 2020 PODCAST SUMMARISATION TRACK.” Accessed: Feb. 06, 2021. [Online]. Available: https://huggingface.co/.

[31] T. Shi, Y. Keneshloo, N. Ramakrishnan, and C. K. Reddy, “Neural Abstractive Text Sum-marization with Sequence-to-Sequence Models,” *ACM/IMS Trans. Data Sci*, vol. 2, no. 1, 2020, doi: 10.1145/3419106.

[32] L. Dong *et al.*, “Unified Language Model Pre-training for Natural Language Understanding and Generation.” Accessed: May 09, 2021. [Online]. Available: https://github.com/microsoft/unilm.

[33] Y. Liu and M. Lapata, “Text Summarization with Pretrained Encoders.” Accessed: May 09, 2021. [Online]. Available: https://github.com/.

[34] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization,” 2020. Accessed: May 09, 2021. [Online].

[35] A. Holtzman, J. Buys, L. Du, M. Forbes, Y. Choi, and P. G. Allen, “THE CURIOUS CASE OF NEURAL TEXT DeGENERATION.” Accessed: May 09, 2021. [Online]. Available: https://github.com/ari-holtzman/degen.

[36] M. Grusky, M. Naaman, and Y. Artzi, “Newsroom: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, vol. 1, pp. 708–719, doi: 10.18653/v1/N18-1065.

[37] M. Koupaee and W. Y. Wang, “WikiHow: A Large Scale Text Summarization Dataset,” *arXiv*, Oct. 2018, Accessed: May 09, 2021. [Online]. Available: http://arxiv.org/abs/1810.09305.

[38] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On Faithfulness and Factuality in Abstractive Summarization,” *arXiv*, May 2020, Accessed: May 09, 2021. [Online]. Available: http://arxiv.org/abs/2005.00661.

[39] Y.-C. Chen and M. Bansal, “Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, vol. 1, pp. 675–686, doi: 10.18653/v1/P18-1063.

[40] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, “Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance,” in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2020, pp. 563–578, doi: 10.18653/v1/d19-1053.

[41] T. Sellam, D. Das, and A. P. Parikh, “BLEURT: Learning Robust Metrics for Text Generation,” *arXiv*, Apr. 2020, Accessed: May 09, 2021. [Online]. Available: http://arxiv.org/abs/2004.04696.

[42] H. Karlbom and A. C. Spotify, “Abstractive Podcast Summarization using BART with Longformer attention.” Accessed: May 05, 2021. [Online]. Available: https://huggingface.co/facebook/bart-large-cnn.

[43] K. M. Hermann *et al.*, “Teaching Machines to Read and Comprehend,” *Advances in Neural Information Processing Systems*, vol. 2015-January, pp. 1693–1701, Jun. 2015, Accessed: May 05, 2021. [Online]. Available: http://arxiv.org/abs/1506.03340.

[44] A. Clifton *et al.*, “100,000 Podcasts: A Spoken English Document Corpus,” in *Proceedings of the 28th International Conference on Computational Linguistics*, Jan. 2020, pp. 5903–5917, doi: 10.18653/v1/2020.coling-main.519.

[45] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The Long-Document Transformer,” *arXiv*, Apr. 2020, Accessed: May 05, 2021. [Online]. Available: http://arxiv.org/abs/2004.05150.

[46] T. Yu, Z. Liu, and P. Fung, “AdaptSum: Towards Low-Resource Domain Adaptation for Abstractive Summarization,” Mar. 2021, Accessed: May 06, 2021. [Online]. Available: http://arxiv.org/abs/2103.11332.

[47] Y.-S. Chen and H.-H. Shuai, “Meta-Transfer Learning for Low-Resource Abstractive Summarization,” 2021. Accessed: May 06, 2021. [Online]. Available: www.aaai.org.

[48] A. Magooda and D. Litman, “Abstractive Summarization for Low Resource Data using Domain Transfer and Data Synthesis,” 2020. Accessed: May 06, 2021. [Online]. Available: http://www.coursemirror.com/download/dataset2.

[49] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. v Le, “XLNet: Generalized Autoregressive Pretraining for Language Understanding.” Accessed: Apr. 30, 2021. [Online]. Available: https://github.com/zihangdai/xlnet.

[50] T. B. Brown *et al.*, “Language Models are Few-Shot Learners,” 2020. Accessed: Apr. 30, 2021. [Online].

[51] F. Nooralahzadeh, G. Bekoulis, J. Bjerva, and I. Augenstein, “Zero-Shot Cross-Lingual Transfer with Meta Learning.” Accessed: Apr. 30, 2021. [Online]. Available: https://github.com/.

[52] T. Isbister and M. S. Rise, “Should we Stop Training More Monolingual Models, and Simply Use Machine Translation Instead?,” 2021. Accessed: Apr. 29, 2021. [Online]. Available: https://cloud.google.com/translate/docs/advanced/translating-.

[53] O. Tilk and T. Alumäe, “Low-Resource Neural Headline Generation.” Accessed: Feb. 06, 2021. [Online].

[54] Ingrid. Mårdh, *Headlinese: On the grammar of English front page headlines*, vol. 58. Lund : Liberläromedel/Gleerup, 1980.

[55] S. Parida and P. Motlicek, “Abstract Text Summarization: A Low Resource Challenge.” Accessed: May 06, 2021. [Online]. Available: http://opennmt.net/OpenNMT-py/.

[56] W. Liu, L. Li, Z. Huang, and Y. Liu, “Multi-lingualWikipedia Summarization and Title Generation On Low Resource Corpus,” Dec. 2019, pp. 17–25, doi: 10.26615/978-954-452-058-8_004.

[57] J. Kallas and K. Koppel, “Estonian National Corpus 2019,” *Center of Estonian Language Resources*, 2020. https://doi.org/10.15155/3-00-0000-0000-0000-08565L (accessed Mar. 14, 2021).

[58] A. See, P. J. Liu, and C. D. Manning, *Get To The Point: Summarization with Pointer-Generator Networks*. Vancouver, Canada: Association for Computational Linguistics, 2017.

[59] “Eesti keele stoppsõnad / Estonian stop words.” https://datadoi.ee/handle/33/78 (accessed Apr. 20, 2021).

[60] S. Orasmaa, T. Petmanson, A. Tkachenko, S. Laur, and H.-J. Kaalep, “EstNLTK - NLP Toolkit for Estonian,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, N. C. (Conference C. and K. C. and T. D. and M. G. and B. M. and J. M. and A. M. and J. O. and S. Piperidis, Ed. Paris, France: European Language Resources Association (ELRA), 2016.

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Henry Härm

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "Abstractive Summarization of news Broadcasts for low resource languages", supervised by Tanel Alumäe.
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

08.05.2021

1 The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 – ERR dataset sample datapoints

{"id": 5760,
"transcript":

“Riigieelarve juures on teadagi oluline, millised on prioriteedid, mille peale raha kulutada ning millised ja kui suured on maksud, kust see raha saadakse. Aga oluline on ka lähtepunkt, kui palju saadakse tulu ja kui palju kulutatakse, ehk kui tasakaalus on eelarve. Käesoleva aasta eelarvet on puudujäägi pärast korduvalt kritiseeritud. Mis saab järgmisel aastal sotside juht Jevgeni Ossinovski. Et nendel aegadel, kui majandus on kasvanud aeglaselt on riik investeerinud rohkem, mille tõttu on on eelarve tulnud kerges miinuses siis nüüd, mil majanduskasv kiire, tuleb kindlasti reserve koguda. Meie hinnangul tuleks järgmise aasta eelarve koostada nominaalse ehk reaalse ülejäägiga. See on kindlasti võimalik ja, ja ma arvan, et, et me jõuame, suudame valitsuses selles osas ka kokkuleppele jõuda. Keskerakonna aseesimees Kadri Simson kinnitab, et tegelikult on see kogu koalitsiooni huvi. Tõepoolest, me lähme riigieelarve strateegiat vormistama siis nii et järgmisel aastal tulud oleksid suuremad kui kulud, ehk et me saavutaksime nominaalse ülejäägi, mis siis tähendab, et järgmisel aastal riigireservid kasvaksid ja seda on võimalik saavutada, aga see on keeruline, sest me oleme kokku leppinud, et mõned juba seadustatud maksutõusud tühistatakse. Ja samas on meie peaminister öelnud, et ühtegi uut maksutõusu järgmine aasta ei tule. Seega me peame leidma lisaks majanduskasvule ka viise, kuidas kulutusi vähendada. Isamaa ja Res Publica Liidu esimees Helir-Valdor Seeder tuletab meelde, et nemad võitlesid juba käesoleva aasta eelarve tasakaalu eest kuid järgmise aasta eelarve osas väljendab ta ennast mõnevõrra ettevaatlikumalt. Loomulikult 2019 aasta riigieelarvele mõeldes ja riigieelarvestrateegiale järgmisel neljal aastal me kindlasti peame seadma eesmärgi vähemalt tasakaalus eelarve. Ja sõltuvalt siis võimalustest, milline on järgmisel neljal aastal majanduskasv, aga ikkagi olukorras, kus majanduskasv on seal üle kolme protsendi, peaksime olema võimelised tulevikus hakkama koguma reserve. Et see on Isamaa ja Res Publica liidu üks niisugune, ütleme siis makromajanduslik või suurem eesmärk, jättes kõrvale valdkondade poliitikad, vajadused, investeeringud, palgapoliitikate küsimused, aga see peab olema üks suur eesmärk. Nagu Kadri Simson juba ütles, eesmärgi saavutamine ei ole hoolimata east majanduskeskkonnast sugugi lihtne sest kõik koalitsioonierakonnad on aru saanud, et maksutõusust tuleb loobuda. Uued maksukavad, puudutagu need siis suhkrut või autosid esialgu unustada.”,

"summary":

“Koalitsioonierakonnad valmistuvad riigieelarve strateegia aruteluks. Üksmeelsed ollakse selles, et miinuses riigieelarvet ei tohi järgmiseks aastaks teha.”,
"name":
“Koalitsioonierakonnad järgmise aasta riigieelarvest.”}

{"id": 4971,
"transcript":

“Majandusminister Taavi Aas andis täna valitsuse pressikonverentsil hoiatuse Viljandi ja Haapsalu elanikele. Nende linnade reoveeproovidest on leitud koroonaviiruse jälgi. Sotsiaalminister Tanel Kiik. Vahepeal on viidud läbi selliste joove, uuringud üle 10 Eestimaa ja on teatud omavalitsusi, kus need näitajad tuli, ütleme siis esialgsetel signaalidele, alarmeerivaid, tõsi on see, et, et Jõhvi puhul näiteks räägime signaalidest, siis see on ootuspärane, seotud ilmselt Ida-Virumaal laiemalt hetkel leviva viiruse puhanguga. Mis puudutab nüüd Haapsalu Viljandit mitte tõesti pressiooniks ka mainiti, siis nende puhul tuleb natuke analüüsida, et kas on tegemist ikkagi konkreetselt koha peale viirusega või on tegemist näiteks sealt linnast läbi liikuvate inimestega ehk et on see siis Tartu inimesed, kes Viljandis peatuste Intalin inimesed, teised teistpidi liikudes ja sama Haap see näide ehk need on piirkonnad, kus tihtipeale tehakse kas vahepeatusi või ka näiteks võidetakse aega lihtsalt siseturismi mõttes ehk ampsu näitajale kindlasti ei ole võimalik. Veendu veendunult öelda, et tegemist on kohalike elanike vabalt, võib olla tegemist ka hoopiski siseturistidega, kes on sinna tulnud teistest Eesti maakondadest. Silm natuke analüüsime uurimist seda, et kas see nii-öelda näitaja on nende korduv ajas või on see täiesti ühekordse sellise episoodiga ja, ja seejärel on võimalik, võib anda teadusnõukoja Terviseameti koostöös selline täpsem põhjalikku selgitus sellele. Bioloogia professor Irja Lutsari sõnul proovide tulemustest järeldusi teha veel ei saa. Ma arvan, et see hoiatus on praegusel ajahetkel ennatlik, et me uurime ja vaatame, see pigem on, eks ole, teadlastele, et teadlased võtavad sealt nendest kohtadest järgmine nädal jälle proovida ja vaatavad neid, aga pigem näitas reoveeuuring, et Tartus signaali üldse enam ei ole, aga, aga seda me juba näeme ka, et Tartus on ilmselt see puhang nagu kontrolli alla praeguseks saadud. Lutsari sõnul võivad proovid näidata tulemusi ka vaid ühe nakatunu korral. Nad võtavad proove erinevatest kollektoritest ja määravad seal siis viiruse RNA-d. Ja, ja see meetod on, on nagu hästi tundlik, nii et kui seal on juba mingi teatud hulk üle üle arvatava või üle määratu vampiiri, siis me juba seda viiruse RNA-d, näeme nüüd, mida see nüüd täpselt tähendab. Selleks on meie tulemusi praegu liiga vähe ja neid peaks nagu vaatama rohkem nädalate kaupa või päevade kaupa, mitte niisugusest ühekordsest signaalist, mis võibki pärineda ju tegelikult ühelt inimeselt, et, et selle järgi veel järeldusi teha ei saa.”,

"summary":

“Hiljem teatas Tartu Ülikool, et valitsuse pressikonverentsil jagati vananenud andmeid.”,
"name":
“Reovee proovid näitavad koroona jälgi Haapsalus ja Viljandis.”}

Appendix 3 – Estonian national dataset sample datapoints

<doc id="1" src="Reference Corpus" filename="aja_EPL_2002_02_22.ma" texttype_nc="periodicals" newspaperNumber="Eesti Päevaleht 22.02.2002" heading="Majandus" title="Hotell Tallinna kõrval asuv auk saab detailplaneeringu" texttype="Journals" texttype_src="source data">

Tallinna linn algatab Paldiski maantee ääres Hotell Tallinna kõrval asuva suure vundamendiaugu ja tühermaa detailplaneeringu koostamise, ehitustööde alustamist takistavad aga ala segased omandisuhted. Piirkonnaarhitekt Alice Laanemägi ütles, et OÜ Maranello Vara on taotlenud linnalt selle ala detailplaneeringu koostamist, et ehitada sinna tulevikus Grand Hotel Tallinna laiendus, veepark ja parkla. Ehitisi planeeritakse kaks, korruseid 20 ja krundi täisehitamise protsendiks 80. Laanemägi lisas, et taotlus on küll olemas, kuid planeeringu koostamist ei ole veel jõutud alustada. OÜ Maranello Vara esindaja Olev Kasak ütles, et nad palusid küll linnal detailplaneering algatada, kuid võimalikust ehitamisest on veel ennatlik rääkida. Krundil on väga palju omanikke ja nende kõigiga tuleb ehitamise või maa ostmise üle läbi rääkida. Paraku ei ole firmal õnnestunud kõiki maaomanikke veel üles leida. AS Amerest Hotels kavatses sellele krundile 10 aastat tagasi ehitada Sheraton hotelli. 1992. aastal sai hoone ehitusloa ja 1993. aastal kaevati valmis vundamendiauk. Seejärel tekkisid ehitajal kohtuvaidlused ühe kinnistu õigusjärgse omanikuga ja kogu ehitus jäi seisma. Vundamendiauk on krundil siiani alles.

</doc>

<doc id="2" src="Reference Corpus" filename="aja_EPL_2002_02_22.ma" texttype_nc="periodicals" newspaperNumber="Eesti Päevaleht 22.02.2002" heading="Majandus" title="Kaitseliit on saanud Liiva kõrtsile kaks pakkumist" texttype="Journals" texttype_src="source data">

Kaitseliit on saanud Tallinnas Viljandi maanteel asuva endise Liiva kõrtsi hoonestusõigusele seni kaks pakkumist, kuid kuulutuse hilinemise tõttu pikendas Kaitseliit konkursi tähtaega järgmise reedeni. Kaitseliit kuulutas 14. veebruaril välja konkursi Tallinnas Viljandi maantee 18 asuval endise Liiva kõrtsihoone kinnistul hoonestusõiguse seadmiseks 99 aastaks alghinnaga 750.000 krooni. Kinnistu suurus on 8745 ruutmeetrist, maa sihtotstarve on seni olnud riigikaitsemaa. Kinnistu territooriumi kasutatakse praegu küttepuude müügiplatsina. Kinnistul asub muinsuskaitsealune Liiva kõrtsihoone. Enampakkumiskonkursi läbiviija Argo Hollo ASist Kodu Haldus ütles, et praeguseks on tehtud kaks pakkumist, kuid konkursi tähtaja pikendamisega nädala võrra loodetakse pakkumisi saada veel. Konkursi tingimuste kohaselt peab hoonestaja püstitama uued ehitised ja renoveerima olemasolevad hooned kümne aasta jooksul. Samuti tuleb kinnistu hoida alaliselt heas seisukorras. Kui ehitamiskohustust ei täideta tähtajaks, peab hoonestaja tasuma 10.000 krooni kuu kohta.

</doc>

Appendix 4 – Translated ERR dataset sample datapoints

{"id": 5760,
"transcript":

"Of course, in the case of the state budget, it is important what are the priorities, on which to spend the money and what and how big are the taxes, where does this money come from. But the starting point is also how much revenue is received and how much is spent, ie how balanced the budget is. This year's budget has been repeatedly criticized for its deficit. What will happen next year to the head of the Socialist Yevgeny Ossinovsky. That in those times when the economy has been growing slowly, the state has invested more, due to which the budget has come in a slight deficit now that economic growth is fast, reserves must be accumulated. In our opinion, next year's budget should be drawn up with a nominal or real surplus. That is certainly possible, and, and I think that we will be able to reach an agreement in government on that. Kadri Simson, the deputy chairman of the Center Party, confirms that this is in fact in the interest of the entire coalition. Indeed, we are going to formulate the state budget strategy so that next year's revenues will be higher than expenditures, that is, we will achieve a nominal surplus, which then means that next year's state reserves will increase and can be achieved, but it is difficult because we have agreed that some of the already legalized tax increases will be reversed. And at the same time, our Prime Minister has said that there will be no new tax increases next year. So, in addition to economic growth, we also need to find ways to reduce spending. Helir-Valdor Seeder, chairman of the Union of Fatherland and Res Publica, reminds that they have already fought for the balance of this year's budget, but he is somewhat more cautious about next year's budget. Of course, with the 2019 state budget in mind and the state budget strategy for the next four years, we must definitely set the goal of at least a balanced budget. And then, depending on the possibilities for economic growth in the next four years, but still in a situation where economic growth is over three percent, we should be able to start accumulating reserves in the future. As this is one such union of the Fatherland and Res Publica, let us say a macroeconomic or larger goal, leaving aside sectoral policies, needs, investment, wage policy issues, but it must be one big goal. As Kadri Simson has already said, achieving the goal is not easy at all, regardless of the economic environment, because all coalition parties have understood that the tax increase must be abandoned. New tax schemes, whether for sugar or cars, will be forgotten for the time being.",

"summary":

"Coalition parties are preparing for the discussion of the state budget strategy. There is a consensus that the minus state budget should not be made for next year."}

{"id": 4971,
"transcript":

"The Minister of Economic Affairs Taavi Aas issued a warning to the residents of Viljandi and Haapsalu at a government press conference today. Traces of coronavirus have been found in wastewater samples from these cities. Minister of Social Affairs Tanel Kiik. In the meantime, such intoxication has been carried out, surveys have been conducted in more than 10 Estonia and there are certain local governments where these indicators came from, say the initial signals, alarming, it is true that in Jõhvi, for example, we are talking about signals, it is expected. With a virus outbreak currently spreading in Virumaa more widely. As for Haapsalu Viljandi, which was not really mentioned for the press, it is necessary to analyze a bit whether it is a virus or a person moving through the city, ie it is the people of Tartu who stop in Viljandi, the people of Intalin, others moving in the opposite direction and the same Haap this example, ie these are areas where stops are often made or, for example, time is saved simply in the sense of domestic tourism, ie it is certainly not possible for the shooter indicator. Make sure to say with confidence that these are local residents freely, they may also be domestic tourists who have come there from other Estonian counties. A little bit of analysis of the investigation is whether this so-called indicator is their recurring over time or whether it is a completely one-off such episode and, and then it is possible, can provide such a more detailed and detailed explanation in collaboration with the Health Board. According to Irja Lutsar, professor of biology, it is not yet possible to draw conclusions from the results of the samples. I think this warning is premature at the moment that we are researching and watching, it is rather, isn't it, for scientists that scientists will take these places to try and watch them again next week, but rather the wastewater study showed that the signal in Tartu is no longer not, but, we can already see that this outbreak in Tartu is probably as under control. According to Lutsar, the samples can show results only in case of one infected person. They take samples from different collectors and then determine the viral RNA there. And, and this method is, as sensitive as it is, so if there's already a certain amount of vampire over and above that, we already see that viral RNA, we'll see exactly what it means now. There are too few of our results for that at the moment, and they should be looked at more by the week or by the day, rather than by a one-off signal that may actually come from one person, so that no conclusions can yet be drawn from it.",

"summary":

"Later, the University of Tartu announced that outdated data was shared at a government press conference."}

Appendix 5 – Translated CNN/Dailymail dataset sample datapoints

{“id”: “0044e296ecfe3ba57a351ad2a36d034491e878ce”,
“article”:

“(CNN) Laulja ja laulukirjutaja David Crosby lõi pühapäeva õhtul autoga sõrkjooksu, ütles pressiesindaja. Õnnetus juhtus Californias Santa Ynezis, kus asub Crosby. Crosby sõitis sõrkjooksu löönud umbes 50 miili tunnis. Seda kinnitas California maantee patrullide esindaja Don Clotworthy. Lähetatud kiirusepiirang oli 55. Sõrkjooksja sai mitu luumurdu ja viidi lennukiga Santa Barbara haiglasse, ütles Clotworthy. Tema vigastusi ei peeta eluohtlikeks. "Hr Crosby oli ametivõimudega koostöös ning ta ei olnud mingil moel kahjustatud ega joobes. Hr Crosby ei näinud sõrkijat päikese tõttu," ütles Clotworthy. Pressiesindaja sõnul olid sõrkjooksja ja Crosby samal teepoolel. Jalakäijad peaksid olema tee vasakul küljel, liikudes liikluse poole, ütles Clotworthy. Sõrkjooksjaid peetakse jalakäijateks. Crosby on tuntud magusate meloodiate üle mitmekihiliste harmooniate kudumise poolest. Ta kuulub kuulsasse rokigruppi Crosby, Stills & Nash. "David Crosby on ilmselgelt väga häiritud, et ta kogemata kedagi lõi. Esialgsete teadete põhjal tunneb ta kergendust, et härra vigastused ei olnud eluohtlikud," ütles Crosby esindaja Michael Jensen. "Ta soovib sõrkjooksjale väga kiiret taastumist.”,

"highlights":

“Õnnetus juhtub Californias Santa Ynezis, kus asub Crosby. Sõrkjooksja sai mitu luumurdu; tema vigastusi ei peeta eluohtlikeks.”}

{“id”: “00716be72be8cf48cc23ac3b4b8924e569628be2”,

“article”:

“(CNN) Sigma Alpha Epsilon on tule all video eest, mis näitab parteisse sattunud vennaskonna liikmeid rassistlikku laulu laulmas. SAE riiklik peatükk peatas üliõpilased, kuid Oklahoma ülikooli president David Boren astus sammu edasi, öeldes, et ülikooli kuuluvus vennaskonnaga on püsivalt tehtud. Uudised on šokeerivad, kuid pole SAE-le esimest korda poleemikat. SAE asutati 9. märtsil 1856 Alabama ülikoolis viis aastat enne Ameerika kodusõda, vennaskonna veebisaidi andmetel. Kui sõda algas, oli rühmal vähem kui 400 liiget, kellest "369 läksid sõtta konföderatsiooniriikide ja seitse liidu armee eest", seisab veebisaidil. Vennaskonnas on nüüd uhke enam kui 200 000 elavat vilistlast ning umbes 15 000 üliõpilast, kes asustavad 219 peatükki ja 20 "kolooniat", kes soovivad ülikoolidesse täisliikmeks saada. SAE on pidanud viimasel ajal pärast rida liikmesurmaid kõvasti vaeva nägema, paljud süüdistasid uute töötajate värbamist, kirjutas SAE riiklik president Bradley Cohen vennaskonna veebisaidil saadetud teates. Vennaskonna veebisaidil on loetletud enam kui 130 peatükki, millele on viidatud või mis on peatatud tervisekaitse ja tööohutuse juhtumite tõttu alates 2010. aastast. Vähemalt 30 juhtumist hõlmas udustamist ja veel kümned alkoholi. Nimekirjas puudub aga arvukalt viimaste kuude juhtumeid. Nende hulgas vastavalt erinevatele meediaväljaannetele: Yale'i ülikool keelas SAE-d eelmisel kuul ülikoolilinnaku tegevustest pärast seda, kui liikmed üritasid väidetavalt sekkuda initsiatsiooniriitusega seotud seksuaalse väärkäitumise uurimisse. Stanfordi ülikool peatas detsembris SAE eluasemeõigused pärast seda, kui leiti, et vennastekoguduses osalevad korporatsiooniliikmed on graafiliselt seksuaalse sisuga. Ja novembris Johns Hopkinsi ülikool peatas vennaskonna alaealiste joomise tõttu. "Meedia on meid sildistanud" rahva surmavaimaks vennaskonnaks "," ütles Cohen. Näiteks 2011. aastal suri üliõpilane, kui teda sunniti ülemäärase alkoholi tarvitamise vastu, vastavalt kohtuasjale. SAE eelmine kindlustusandja vedas vennaskonna maha. "Selle tulemusena maksame Londoni Lloyd'sile Kreeka tähtede maailmas kõrgeimaid kindlustusmakseid," ütles Cohen. Ülikoolid on keeldunud SAE katsetest avada uusi peatükke ja vennaskond pidi ähvardavate intsidentide tõttu 18 kuuga 12 sulgema.”,

"highlights":

“Sigma Alpha Epsiloni viskab Oklahoma ülikool välja. Samuti on viimastel kuudel Yale'i, Stanfordi ja Johns Hopkinsi ametnikud vastuolus olnud.”}