Baltic J. Modern Computing, Vol. 10 (2022), No. 3, pp. 511–524
https://doi.org/10.22364/bjmc.2022.10.3.23

Abstractive Summarization of Broadcast News Stories for Estonian

Henry HÄRM, Tanel ALUMÄE

Institute of Software Science, Tallinn University of Technology, Estonia

henry.harm@taltech.ee, tanel.alumae@taltech.ee

Abstract. We present an approach for generating abstractive summaries for Estonian spoken news stories in a low-resource setting. Given a recording of a radio news story, 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 and applying a state-of-the-art text summarization system to generate the result. We evaluated a number of models, with the best-performing model leveraging the large English BART model pre-trained on CNN/DailyMail dataset and fine-tuned on machine-translated in-domain data, and with the test data translated to English and back. The method achieved a ROUGE-1 score of 17.22, improving on the alternatives and achieving the best result in human evaluation. The applicability of the proposed solution might be limited in languages where machine translation systems are not mature. In such cases multilingual BART should be considered, which achieved a ROUGE-1 score of 17.00 overall and a score of 16.22 without machine translation based data augmentation.

Keywords: Abstractive summarization, low-resource languages, pre-trained models, multilingual models, machine-translation

1 Introduction

The growth of online multimedia, such as talks, presentations, lectures and news, has created a significant need to provide easy access to these resources (Furui et al., 2001). 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 signals (Furui, 2003). Automatic summarization has the potential to efficiently generate concise and fluent summaries while preserving critical information from the original media.

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. One of many extractive techniques

is LexRank, which analytically computes the relative importance of words and sentences to produce the summary (Erkan and Radev, 2011). Abstractive summarization, on the other hand, can generate novel sentences by either rephrasing or using the new words, instead of simply extracting the important sentences (Rachabathuni, 2018). The method better approximates human summaries, however abstractive summarization is an exceedingly non-trivial and challenging task (Allahyari et al., 2017).

As news broadcasts primarily contain spoken-word content, summarization can be performed in the text domain on the transcript of an episode, as shown with PodSumm (Vartakavi and Garg, 2020). Deep learning-based neural summarization performs well when applied to abstract text summarization (Salakhutdinov, 2014) compared to structure-based and semantic-based abstractive summarization approaches. In general, neural summarization is solved using an encoder-decoder architecture with recurrent neural networks or self-attention (Sutskever et al., 2014). 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 (Kurniawan and Louvan, 2019). The state-of-the-art neural abstractive summarization models are trained with annotated datasets of hundreds of thousands or millions of data points. At the same time, such quality and quantity are not feasible for most languages, including Estonian.

This work focuses on achieving state-of-the-art low-resource abstractive summarization results for the Estonian language radio news stories. We propose an approach that consists of two steps: automatically generating the transcript, and applying a state-of-the-art text summarization system to generate the result. To overcome the problems of limited available data for training, transfer learning methods on pre-trained models, multilingual models and machine translation were explored and included in the summarization pipeline.

2 Related Work

The majority of neural abstractive summarization models use the encoder-decoder architecture (Sutskever et al., 2014). In order to reduce the bottleneck between encoder and decoder, the attention mechanism is employed, where the decoder is given a weighted average view over the encoded source words at each auto-regressive generation step (Bahdanau et al., 2015; Rush et al., 2015; Nallapati et al., 2016). However, to achieve strong performance, large task specific datasets of up to hundreds of thousands of training documents are required. For example, neural summarization research is often performed in the news domain where numerous large datasets exist, predominantly in the English language. For example, the CNN/DailyMail and the New York Times (NYT) datasets contain around 300k and 700k documents, respectively. Manually annotating training datasets is an expensive and time consuming endeavor, requiring professionals with domain knowledge. Moreover, available sources have a large variety of writing styles and forms, such as news articles, social media posts, and scientific papers. Therefore, improving the performance of abstractive models with limited labeled training examples has become an important problem, as large-scale human-annotation is not feasible in most practical situations.

The most common way to solve the low-resource problem in abstractive summarization is to use some form of pretraining using unlabelled data. Several pretraining techniques have been proposed during the recent years. One relatively early approach was to pretrain the encoder part of the model on unlabelled data, using the language modeling (i.e., next word prediction) objective, and later use it as a starting point for training an encoder-decoder model, while training the rest of the parameters from scratch (Tilk and Alumäe, 2017). In (Ramachandran et al., 2017), the encoder, as well as the embedding and first recurrent layers of the decoder were initialized from a pretrained language model, while the only the encoder layers needing attention over encoder outputs were trained from scratch.

Using large pretrained models for a variety of NLP tasks gained strong momentum with the introduction of BERT (Devlin et al., 2019) that uses the Transformer (Vaswani et al., 2017) encoder architecture. The model uses two training objectives: masked language model (MLM), and next sentence prediction (NSP). 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 (Nozza et al., 2020). BERT was primarily developed for encoding text representations as an encoder only architecture.

Rothe et al. (2020) proposed a Transformer-based sequence-to-sequence model that allows to combine pre-trained Transformer encoder and decoder models. This resulted in new state-of-the-art results in tasks such as machine translation and text summarization (Rothe et al., 2020). Several combinations of model initializations can be used, such as BERT2BERT, a BERT-initialized encoder and decoder with randomly initialized encoder-decoder attention. Several language-specific BERT models have been trained, such as CamemBERT (Martin et al., 2020) and FlauBERT (Le et al., 2020), which have shown improvements over multilingual BERT models, such as XLM-RoBERTa (Lewis et al., 2020). In (Tanvir et al., 2021), an Estonian language-specific BERT model, EstBERT, was described. The evaluation showed that the model outperforms multilingual BERT in most NLP task and proves the usefulness of language-specific models.

Lewis et al. (2020) proposed a self-supervised training method BART. A BART model is trained by firstly corrupting text with an arbitrary noising function and secondly learning a model to reconstruct the original text. Lewis et al. (2020) show that the model achieves state-of-the-art results for abstractive dialogue, question answering and summarization tasks. The multilingual mBART model (Liu et al., 2020) is obtained by applying the BART training method to large-scale monolingual corpora across many languages. As mBART is trained once for all languages as a complete model, it can be fine-tuned for any of the languages in both supervised and unsupervised settings, without any task-specific or language-specific modifications or initialization schemes.

Another method to improve abstractive summarization performance in low-resource scenarios is to augment the available labelled in-domain training data with synthesized data (Fabbri et al., 2021; Loem et al., 2022). For example, abstractive summarization models can be pretrained using the task of headline generation, for which there is more labelled data available, or out-of-domain labelled data can be used to pretrain a summarization model, before finetuning with limited amount of in-domain data (Yu et al., 2021; Magooda and Litman, 2020). Another direction of research is multilingual train-

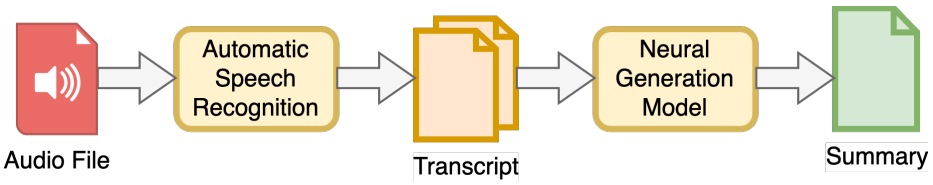

Fig. 1. Illustration of the summarization system. The original audio is transcribed with automatic Speech recognition (ASR), which produces a transcript. A fine-tuned text summarization neural model is used to generate the final abstract summary.

ing. Hasan et al. (2021) introduced XL-Sum, a largescale, high-quality multilingual text summarization dataset. It was also demonstrated that multilingual training can help towards better summarization, most likely due to the positive transfer between morphologically similar languages. Nevertheless, studies have shown that a huge gap still exists between the low-resource and high-resource settings in abstractive summarization.

Isbister et al. (2021) proposed a method of translating task data into English in order to enable the use of large English models. Whereas multilingual models aim to transfer the model to other languages, the proposed approach, on the other hand, aims to transfer the target language test data via machine translation to a high-resource language and apply the well-trained model directly in this language. It was shown that that such approach outperforms native language models in most Scandinavian languages that were experimented with. The exception was the Finnish language, where translation quality was inferior.

3 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 1. In other use-cases where the task data is already in text domain the ASR step is not necessary. A data collection tool is used to build the corpora for training the neural model.

3.1 Automatic Speech Recognition

Automatic speech recognition converts speech into text, enabling the summarization to be solved in the text domain. In our experiments, the publicly available Estonian speech transcription system developed in the Tallinn University of Technology (Alumäe et al., 2018) is used. The word error rate (WER) of the system on broadcast news and broadcast conversations is around 8%. The system also performs automatic punctuation recovery to the recognized stream of words, allowing models trained on punctuated texts to be used for summarization.

3.2 Summarization Models

Several neural abstractive summarization methods were investigated and compared to baseline approaches.

3.2.1 Extractive Baselines. In broadcast news, the first sentence often gives a general overview of the rest of the story. Hence, an easy way to summarize the story is to simply use the first sentence. Extractive text summarization 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 (El-Kassas et al., 2021).

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 (Erkan and Radev, 2011). 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 (Erkan and Radev, 2011).

3.2.2 BERT2BERT. BERT2BERT is a pretraining method for sequence-to-sequence Transformer models where the weights of both encoder and decoder are initialized from a pretrained BERT-style models (Rothe et al., 2020) Only the the weights of the encoder-decoder attention layers are initialized randomly.

We experiment with the following pretrained models as BERT2BERT initialization checkpoints:

1. mBERT[1] is a pre-trained BERT (Devlin et al., 2019) model trained on a Wikipedia corpus containing 104 languages (including Estonian).
2. EstBERT[2] (Tanvir et al., 2021) is a BERT model pre-trained on the Estonian language. For training the EstBERT, the Estonian National Corpus 2017 (Kallas and Koppel, 2020) was used. It consists of four sub-corpora: Estonian Reference Corpus 1990-2008, Estonian Web Corpus 2013, Estonian Web Corpus 2017 and Estonian Wikipedia Corpus 2017.
3. XLM-RoBERTa[3] (Conneau et al., 2020) is a RoBERTa (Liu et al., 2019) model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages, including Estonian. RoBERTa training method is very similar to that of BERT, but it only uses the MLM training objective and adds some algorithmic improvements with regard to BERT, such as dynamic masking and larger batch size.

3.2.3 BART. BART (Lewis et al., 2020) is a self-supervised pretraining method for sequence-to-sequence Transformer models. BART model is trained to perform as a denoising autoencoder: the training data includes “corrupted” or “noisy” text, which has

[1] https://huggingface.co/bert-base-multilingual-cased
[2] https://huggingface.co/tartuNLP/EstBERT
[3] https://huggingface.co/xlm-roberta-base

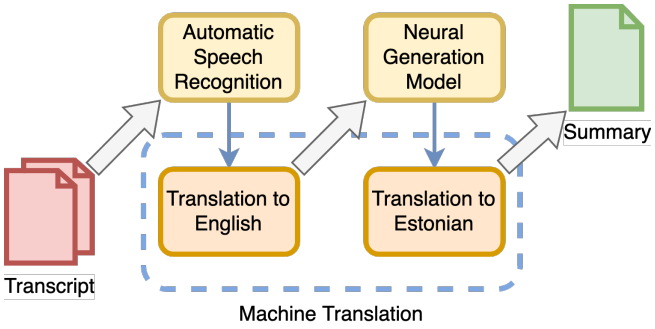

Fig. 2. Machine translation summarization architecture. The broadcast transcription will be machine translated to English and fed into the summarization model. The output will be machine translated back to Estonian for the final result.

to be mapped to clean or original text by the sequence-to-sequence model. The noising schemes that are used by BART are token masking, token deletion, text infilling, sentence permutation, and document rotation.

We use the following pretrained BART models as initialization checkpoints:

1. mBART25[4] (Liu et al., 2020), which is a multilingual BART model with 12 encoder and decoder layers trained on monolingual data from 25 languages, including Estonian.
2. BART, pre-trained on 160 GB of English CommonCrawl data. More specifically, we use the pre-trained BART model that is already finetuned on the CNN/DailyMail summary corpus[5]. We use this model in the "translate-test" setup, i.e., input test data is machine-translated from Estonian to English and the generated summaries are machine-translated back to Estonian, as described on Figure 3.2.3. We also experiment with finetuning the model with in-domain data (i.e. ERR transcripts and the corresponding summaries) translated from Estonian to English.

3.3 Machine Translation

As explained above, several experiments require translating between Estonian and English texts. The Google Cloud Translation API is used in all experiments.

4 Datasets

To train the neural abstractive summarization model, a large number of documents must be available in the target language. As the training is supervised, every document needs to be annotated with a handwritten summary. This section describes the datasets used in this paper. Table 1 lists some numerical facts about each dataset.

[4] https://huggingface.co/facebook/mbart-large-cc25
[5] https://huggingface.co/facebook/bart-large-cnn

Table 1. Statistics about used datasets. Dataset sizes with train, test, validation splits and source and target average word count.

Dataset	Train	Test	Validate	Source avg. len.	Target avg. len.
ERR	4758	595	595	341	19
ETNC19	268 472	-	-	242	9
Translated CNN/DM	9 359	-	-	497	33
English CNN/DM	286 817	13 368	11 487	766	53

4.1 ERR

The ERR "Uudised" news story archive is used as the main experimentation dataset. It contains around 9000 annotated spoken news recordings. The statistics for the datasets are given in Table 1. This dataset was selected for its well-formed and short structure (around 2-minute stories). Each episode consists of one news story, often containing one or more relevant interviews. As each episode contains a single story, we do not need to apply topic segmentation. The recordings do not contain advertisements which simplifies the task, as these should be removed or ignored in the pipeline.

The recording is transcribed with the ASR system and added to the dataset. Using visual dataset analysis, some common anomalies and problems with the summaries were identified, and necessary filters were implemented and added to the pre-processing phase. 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 were added, and unwanted characters such as line breaks were cleaned for well-formed sentences. The dataset consists of 5948 data points, each containing the following properties: episode identifier, generated transcript, summary and headline. An example datapoint can be seen in Table 2.

4.2 ETNC19

To increase multilingual models general understanding of the Estonian language, second phase pre-training experiments are conducted with the largest Estonian text corpus, the Estonian National Corpus 2019 (ETNC19) (Kallas and Koppel, 2020). The corpus consist of Estonian articles, periodicals, blogs, Wikipedia and web pages. Most of the documents in the corpus include a headline. When training models using this corpus, we use the provided headline as the supervision (i.e., as a very condensed summary), although we acknowledge that the headlines are stylistically and grammatically often very different from summaries.

4.3 CNN/DailyMail

The English-language CNN/DailyMail corpus (Nallapati et al., 2016) contains human-generated abstractive summaries from news stories on the CNN and Daily Mail web-

Table 2. Example from the collected ERR news broadcast dataset. The episode has been trascribed by the ASR system which is used as the source text of the summarization.

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.
Headline
Koalitsioonierakonnad järgmise aasta riigieelarvest.
Id
5760

sites. The summaries are automatically produced from the summary bullets accompanying each story, where each bullet is treated as a sentence. We use this corpus in two ways. The original English data is used for training the English BART model. We also machine-translate a subset of the dataset (9000 document-summary pairs) to Estonian and use it as additional training data for Estonian summarization models.

5 Results

5.1 Experiments

The results are shown in Table 3 and example summaries are shown in Table 4. Sample transcript translated into English and the corresponding summary generated using BART is shown in Table 5. The models described in Section 3.2 are finetuned or pretrained with the datasets, as described in Section 4. The multilingual models are firstly pre-trained with the ETNC19 and translated DailyMail datasets. The ETNC dataset is used with headlines as the target word sequence. The training is done for two epochs for both datasets respectively. The Estonian BERT omits pre-training on ETNC as it has already been self-trained on the corpus. After pre-training the models are finetuned on the ERR dataset and evaluated. Fine tuning is done with our ERR dataset (or its translated version with English BART models) for 32 epochs. The learning rate has been set to 5e-05, batch size to 24 with 2 gradient accumulation steps and with Adam optimizer selected. Two NVIDIA Tesla V100 GPUs, 128GB of RAM and 16 CPU cores were utilized for training the models. With the specified hardware, the fine-tuning takes around 4 hours. For generation the beam size is set to 5 and for BART the length penalty is set to 1.0 and minimum length is set to 10.

Table 3. Experimentation results with the models and their ROUGE scores.

Model	Training data	ROUGE-1	ROUGE-2	ROUGE-L
Extractive baselines				
First sentence		12.03	3.45	10.14
LexRank		10.88	2.86	9.66
BERT2BERT				
EstBERT	Translated CNN/DM, ERR	11.72	3.13	10.88
mBERT	ETNC19, Translated CNN/DM, ERR	12.03	3.45	10.14
XLM-RoBERTa	ETNC19, Translated CNN/DM, ERR	12.07	3.35	10.43
BART				
mBART	ERR	16.22	5.03	13.43
mBART	ETNC19, Translated CNN/DM, ERR	17.00	**5.52**	14.30
Testset translated into English and back				
BART	CNN/DM	13.02	3.33	9.97
BART	CNN/DM, Translated ERR	**17.22**	5.15	**14.51**

In addition, baseline LexRank and first sentence implementations were tested. For the LexRank implementation, Estonian stopwords compiled by Uiboaed (2018) are imported. To tokenize and split the sentences, the EstNLTK (Orasmaa et al., 2016) toolkit is used, and the computing of ROUGE scores are done with the Hugging Face datasets library.

5.2 Analysis

The results for the experimentation show that using "round-trip" machine translated test data together with the BART model pre-trained on CNN/DailiyMail and fine-tuned on our task data outperforms the native and multilingual models. The system achieves a ROUGE-1 score of 17.22. The model does not need extensive fine-tuning as it is pre-trained for downstream summarization tasks, significantly reducing the time and resources required.

The second best performing model is the multilingual BART with a ROUGE-1 score of 17.00, which is not significanlty lower than the BART model. However, during human evaluation the model achieved a significanlty lower result of 3.0. Other models needed extensive fine-tuning with larger datasets and produce lower results. The simple first sentence model generally performed well for the given task with a ROUGE-1 score of 12.03. 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. LexRank method achieved a ROUGE-L score of 10.88.

The applicability of using the proposed BART based approach is dependent on the existence of a high-quality machine translation solution for the target language. Testing with the Estonian language shows that the translation is adequate; however, languages

Table 4. Generated summaries for the best performing models for the same reference episode.

Source	Summary
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. *[...]* 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.	**BART + translate-test** Kõik koalitsioonierakonnad tahavad, et järgmise aasta riigieelarve jääks ülejäägiks. Selle saavutamine on aga keeruline, sest osa juba seadustatud maksutõusudest pööratakse tagasi ja osa jääb puutumata. **mBART** Valitsus otsustas teha lisaeelarvesse järgmise nelja aasta jooksul.
Majandusminister Taavi Aas andis täna valitsuse pressikonverentsil hoiatuse Viljandi ja Haapsalu elanikele. Nende linnade reoveeproovidest on leitud koroonaviiruse jälgi. *[...]* 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.	**BART + translate-test** Viljandi ja Haapsalu reoveeproovidest leiti koroonaviiruse jälgi. **mBART** Valitsus andis täna valitsuse pressikonverentsil hoiatuse Viljandi ja Haapsalu elanikele koroonaviiruse jälgi sisaldavate reoveeproovide eest.
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. Tasku keskuse aktsiaseltsi tegevjuht Annika Oja räägib, et kinoketi lahkumine oli kahepoolne otsus. *[...]* Ei, meil ei ole olnud selliseid detailseid läbirääkimisi, aga mõtteid selles suunas teid on erinevates suundades, aga, aga ei ole ei ole otsuseid ja ei ole ka konkreetseid plaane, mida, mida saaks kommenteerida. Samas kui juubee Investi soov ära osta Foorum Cinemas Baltikumi kino äri läbi läheb ehk ettevõte selleks Konkurentsiametilt heakskiidu saab, kuuluks kinoketile tar tuska kinoekraan. Teisisõnu kasvaks Apollo kino mõjuvõim Tartu turul veelgi. Äärime ütleb, et ei oska kommenteerida, kas ja kui kaugel Foorum Cinemas ja Baltikumi ostmisega ollakse.	**BART + translate-test** Alates septembrist on Tartu kinoturul vaid kaks kino: Cinamoni keskus ja Tasku kaubanduskeskuse keskus. Ei kontsern ega kaubanduskeskus pole kinnitanud, kas ja kui kaugel nad on raskustes oleva Forum Cinemase väljaostmisest. **mBART** Septembrist suleti Tartu Tasku kaubanduskeskuse filmikeskus Cinamon. Cinamon Group põhjendab oma kodulehel Tartu kinokeskuse sulgemist lausetega, et kino on küll pidevalt täiustada tööd, aga pärast korduvalt nurjunud katseid leida koostööpartnereid Tasku kaubanduskeskusega ei jäänud muud võimalust, kui lõpetada kino opereerimine ja sulgeda lõplikult filmikeskus.

Table 5. Automatically translated example for BART + translate-test

Source	Summary
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. *[...]* 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.	**BART + translate-test** All coalition parties want next year's state budget to be as a surplus. However, achieving this is difficult, because some of the already legalized tax increases will be reversed and some will be left untouched.

that are typologically vastly different from English and do not have suitable translation models need to consider if the solution is applicable. In such cases, multilingual BART can be considered as it was the best performing solution that achieved a score of 16.22 without machine-translation. The next best performing model is the native BERT

Table 6. Generated transcript example with news anchor and phone interview sections

News anchor section	Phone interview section
Ehitusturu langust ootavad ettevõtjad järgmistes kvartalites. Paremas seisus on riigitellimusi täitvad firmad, rohkem eratellijatest sõltujatel võib sügiseks töö otsa lõppeda, ütles Mitt.	Hoonet ehitusest peaaegu 80 protsenti oli eratellimus ja circa 20 protsenti riigi tellimus ja eratellimused on väga paljuski edasi lükatud. Kukkumine tuleb kindlasti suur sektorid, targem on riigil rohkem ehitada. Neid sektorid on 50000 tööd. Kui need tööd saavad, siis riik saab makse, teistpidi saab uut infrastruktuuri uusi hooneid.

Table 7. Human summary evaluation for BART, mBART model and first sentence baseline on a random subset of ERR test dataset. The survey shows the percentage of participants who prefer a given summary.

Model	Percentage
BART + translate-test	42%
First sentence	3%
mBART	26%

model with ROUGE-1 score of 11.72. The multilingual XLM-RoBERTa model should be considered where native language models are not available, achieving comparable performance of ROUGE-1 12.07. 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 order to verify the results we conducted a survey aimed to compare the two best performing models. The participants were shown a subset of 20 transcripts with 3 generated summaries with BART, mBART models and baseline first sentence summary in a random order. The participants were asked to select the best summary considering the following aspects:

1. State the main ideas of the article, not just the superficial details.
2. Identify the most important details that support the main ideas.
3. Summary is as concise as possible.
4. Does not have grammatical or factual errors.

The results of the survey with 15 participants are given in Table 7. The responses show that the preferred summaries are generated with the BART model which also had the highest ROUGE score from the experiments. The baseline was preferred over the mBART results, however the difference between them was not large.

We reviewed the experiment results manually and noted that that the ASR process is not perfect. The ASR accuracy is high with only minor mistakes in the sections where the news story is being read out. Issues mainly arise in interview sections, where guests are commenting on the topic. In these sections the audio quality is lower with disruptions and the speech is more spontaneous. Example given in Table 6 shows that

interview section can have low readability. However, as the sections mostly give extra context, the generated summaries are not affected.

The machine translation used with the BART model performs well in general, but can decrease readability. The translation can choose words that do not suit the context of the sentence or the structure of the sentence is not optimal. However, our human evaluation survey would suggest that the shortcomings are minor or infrequent compared to other models.

6 Conclusions

In this work, we introduce a system for the news broadcast abstractive summarization task under low resource conditions. The system consists of a automatic speech recognition system that produces a transcript, and a neural summarization model, which generates the final summary. Three possible models with their variations were investigated and compared to find the most suitable solution. The best performing approach is the English BART model pre-trained on the CNN/DailyMail dataset and fine-tuned on translated in-domain data, with test data "round-trip" translated to English and back. The model outperforms the alternative models with ROUGE-1 score of 17.22 and achieves the best result in human evaluation. In target languages where machine translation systems are not mature, the multilingual BART model fine-tuned on a task specific dataset should be considered. Due to time constraints this work concentrated on BART and mBART models. In future work we will continue to expand our experimentation with additional models such as mT5 (Xue et al., 2020) and DeltaLM (Ma et al., 2021).

References

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., Kochut, K. (2017). Text summarization techniques: a brief survey, *ArXiv preprint* **abs/1707.02268**.

Alumäe, T., Tilk, O., Asadullah (2018). Advanced rich transcription system for Estonian speech, *Baltic HLT*.

Bahdanau, D., Cho, K., Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate, *Proc. ICLR*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale, *Proc. ACL*, pp. 8440–8451.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding, *Proc. NAACL-HLT*, pp. 4171–4186.

El-Kassas, W. S., Salama, C. R., Rafea, A. A., Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey, *Expert Systems with Applications* **165**, 113679.

Erkan, G., Radev, D. R. (2011). Lexrank: Graph-based lexical centrality as salience in text summarization, *Journal of Artificial Intelligence Research* **22**, 457–479.

Fabbri, A., Han, S., Li, H., Li, H., Ghazvininejad, M., Joty, S., Radev, D., Mehdad, Y. (2021). Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation, *Proc. HLT-NAACL*, pp. 704–717.

Furui, S. (2003). Recent advances in spontaneous speech recognition and understanding, *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

Furui, S., Iwano, K., Hori, C., Shinozaki, T., Saito, Y., Tamura, S. (2001). Ubiquitous speech processing, *ICASSP*, p. 13–16.

Hasan, T., Bhattacharjee, A., Islam, M. S., Mubasshir, K., Li, Y.-F., Kang, Y.-B., Rahman, M. S., Shahriyar, R. (2021). XL-sum: Large-scale multilingual abstractive summarization for 44 languages, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4693–4703.

Isbister, T., Carlsson, F., Sahlgren, M. (2021). Should we stop training more monolingual models, and simply use machine translation instead?, *Proc. NoDaLiDa*, pp. 385–390.

Kallas, J., Koppel, K. (2020). Estonian National Corpus 2019.

Kurniawan, K., Louvan, S. (2019). IndoSum: A new benchmark dataset for Indonesian text summarization, *Proceedings of the 2018 International Conference on Asian Language Processing*, p. 215–220.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., Schwab, D. (2020). FlauBERT: Unsupervised language model pre-training for French, *Proc. LREC*, pp. 2479–2490.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, *Proc. ACL*, pp. 7871–7880.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation, *Transactions of the Association for Computational Linguistics* **8**, 726–742.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach, *ArXiv preprint* **abs/1907.11692**.

Loem, M., Takase, S., Kaneko, M., Okazaki, N. (2022). Extraphrase: Efficient data augmentation for abstractive summarization, *ArXiv preprint* **abs/2201.05313**.

Ma, S., Dong, L., Huang, S., Zhang, D., Muzio, A., Singhal, S., Awadalla, H. H., Song, X., Wei, F. (2021). Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders, *arXiv preprint arXiv:2106.13736* .

Magooda, A., Litman, D. (2020). Abstractive summarization for low resource data using domain transfer and data synthesis, *The Thirty-Third International Flairs Conference*.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., Sagot, B. (2020). CamemBERT: a tasty French language model, *Proc. ACL*, pp. 7203–7219.

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B. et al. (2016). Abstractive text summarization using sequence-to-sequence RNNs and beyond, *Proc. CoNLL*, pp. 280–290.

Nozza, D., Bianchi, F., Hovy, D. (2020). What the [MASK]? making sense of language-specific BERT models, *ArXiv preprint* **abs/2003.02912**.

Orasmaa, S., Petmanson, T., Tkachenko, A., Laur, S., Kaalep, H.-J. (2016). EstNLTK - NLP toolkit for Estonian, *Proc. LREC*, pp. 2460–2466.

Rachabathuni, P. K. (2018). A survey on abstractive summarization techniques, *Proc. ICICI* pp. 762–765.

Ramachandran, P., Liu, P., Le, Q. (2017). Unsupervised pretraining for sequence to sequence learning, *Proc. EMNLP*, pp. 383–391.

Rothe, S., Narayan, S., Severyn, A. (2020). Leveraging pre-trained checkpoints for sequence generation tasks, *Transactions of the Association for Computational Linguistics* **8**, 264–280.

Rush, A. M., Chopra, S., Weston, J. (2015). A neural attention model for abstractive sentence summarization, *Proc. EMNLP*, pp. 379–389.

Salakhutdinov, R. (2014). Deep learning, *Proc. of KDD*, p. 1973.

Sutskever, I., Vinyals, O., Le, Q. V. (2014). Sequence to sequence learning with neural networks, *Advances in Neural Information Processing Systems 27*, pp. 3104–3112.

Tanvir, H., Kittask, C., Eiche, S., Sirts, K. (2021). EstBERT: A pretrained language-specific BERT for Estonian, *Proc. NoDaLiDa*, pp. 11–19.

Tilk, O., Alumäe, T. (2017). Low-resource neural headline generation, *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 20–26.

Uiboaed, K. (2018). Eesti keele stoppsõnad / Estonian stop words.

Vartakavi, A., Garg, A. (2020). Podsumm – podcast audio summarization, *ArXiv preprint* **abs/2009.10315**.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need, *Advances in Neural Information Processing Systems 30*, pp. 5998–6008.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C. (2020). mt5: A massively multilingual pre-trained text-to-text transformer, *arXiv preprint arXiv:2010.11934* .

Yu, T., Liu, Z., Fung, P. (2021). AdaptSum: Towards low-resource domain adaptation for abstractive summarization, *Proc HLT-NAACL*, pp. 5892–5904.

Received August 19, 2022 , accepted August 27, 2022