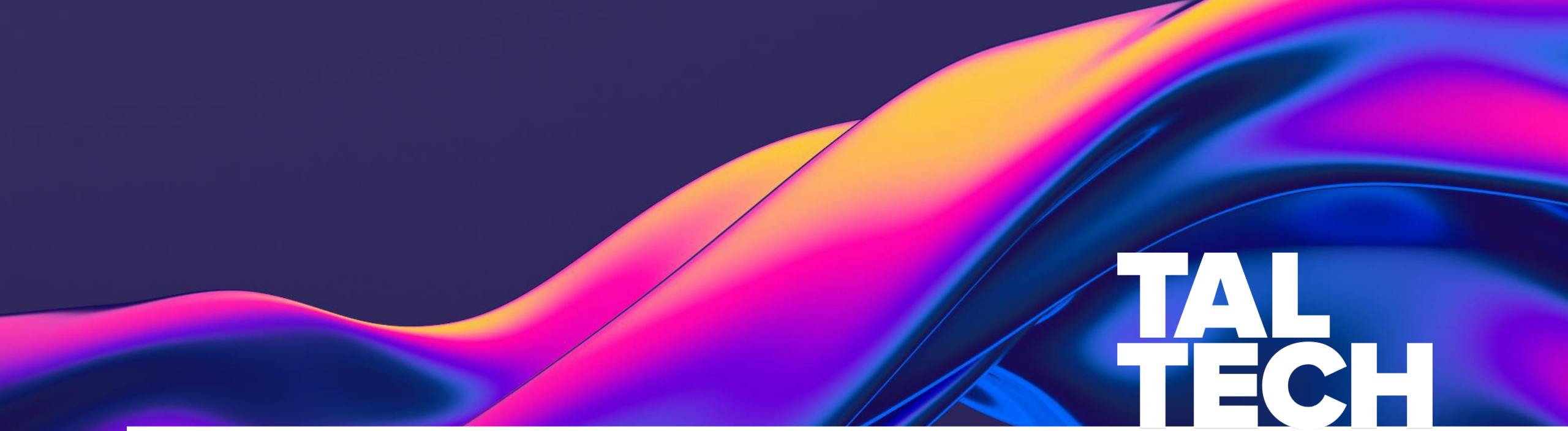

ABSTRACTIVE SUMMARIZATION OF BROADCAST NEWS STORIES FOR ESTONIAN

Henry Härm, Tanel Alumäe
Laboratory of Language Technology
Tallinn University of Technology

TALLINN UNIVERSITY OF TECHNOLOGY

INTRODUCTION

- Generate Abstractive summaries for Estonian language radio news stories.
- Summarization process consists of two steps:
  - Automatic Speech Recognition converts speech into text
  - Neural summarization model generates summary
- Several abstractive models were compared

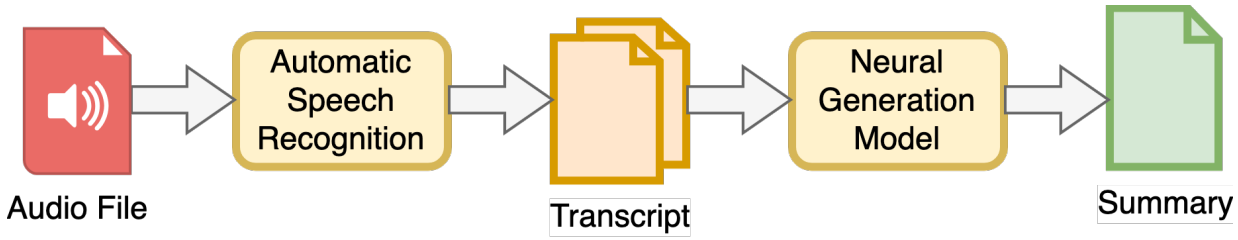

Fig. 1. Summarization process

TAL TECH

MODELS

- BERT
  - mBERT is a pre-trained BERT model trained on a Wikipedia corpus containing 104 languages (including Estonian).
  - EstBERT is a BERT model pre-trained on the Estonian language with the Estonian National Corpus 2017.
  - XLM-RoBERTa is a RoBERTa model pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages, including Estonian.
- BART
  - mBART25, which is a multilingual BART model trained on monolingual CommonCrawl data from 25 languages, including Estonian.
  - BART, pre-trained on 160 GB of English CommonCrawl data. More specifically, we use the pre-trained BART model finetuned on the CNN/DailyMail summary corpus5.

TAL TECH

BART ARCHITECTURE

- BART is a denoising autoencoder that maps a corrupted document to the original document it was derived from
- It is implemented as a sequence-to-sequence model with a bidirectional encoder over corrupted text and a left-to-right autoregressive decoder
- For pre-training negative log likelihood of the original document is optimized
- Large model has 12 layers in the encoder and decoder, and a hidden size of 1024
- Base model has 6 encoder and 6 decoder layers, with a hidden size of 768

TAL TECH

MBART ARCHITECTURE

- mBART is trained by applying the BART to large-scale monolingual corpora across many languages
- mBART is trained once for all languages, providing a set of parameters that can be fine-tuned for any of the language pairs in both supervised and unsupervised settings, without any task-specific or language-specific modifications or initialization schemes.

TAL TECH

ENGLISH BART

- For English BART “translate-test” setup is used
- The input test data is machine-translated from Estonian to English and the generated summaries are machine-translated back to Estonian

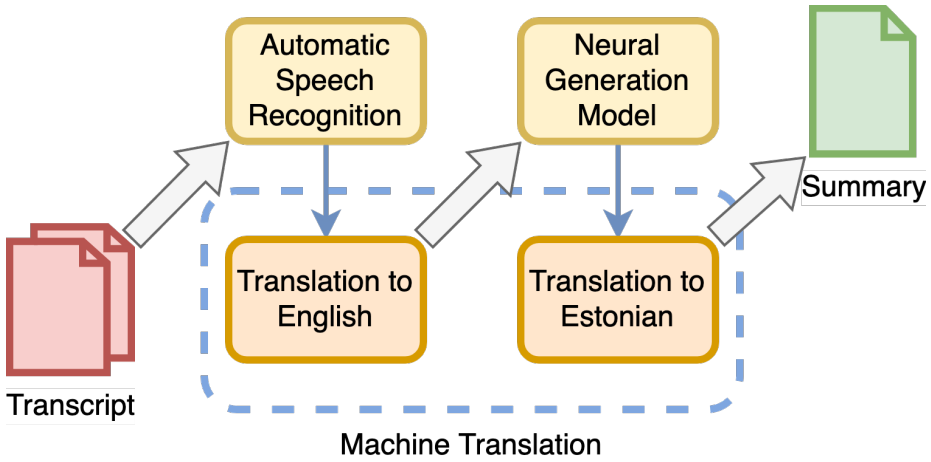

Fig. 2. Machine translation summarization process.

DATASETS

Table 1. Datasets Overview.

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

Table 2. ERR dataset example.

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

TAL TECH

CNN/DM AND ETNC19

- ETNC19 consist of Estonian articles, periodicals, blogs, Wikipedia and web pages and we use headlines for supervised training targets
- Translated CNN/DM consists of 9000 datapoints that are machine translated to Estonian

TAL TECH

BASELINES

- LexRank analytically computes the relative importance of words and sentences to produce the summary.
- First sentence method uses the first sentence of a text. The idea is that the first sentence of a news story highlights the main points.

METRICS

- ROUGE-1 refers to the overlap of unigram (each word) between the system and reference summaries.
- ROUGE-2 refers to the overlap of bigrams between the system and reference summaries.
- ROUGE-L: Longest Common Subsequence (LCS) based statistics identifies longest co-occurring in sequence n-grams automatically.

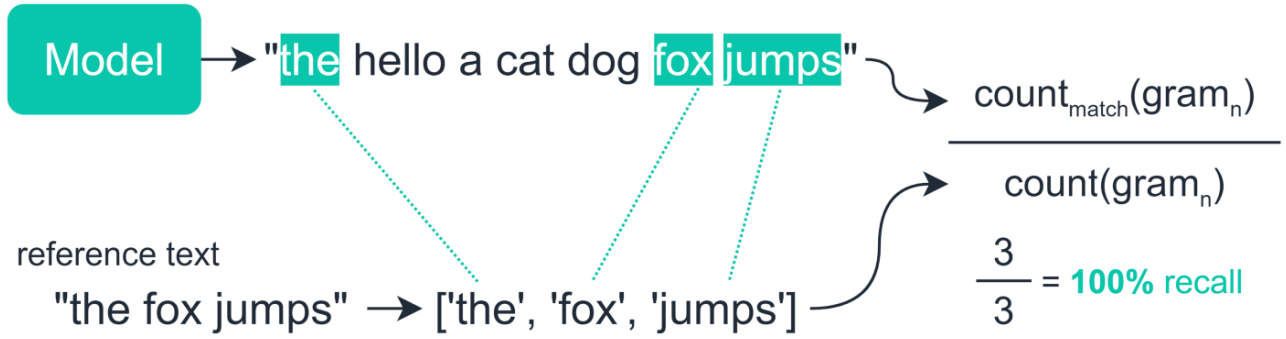

Fig. 3. Rouge score eqample.

RESULTS

- BERT models are used with BERT-initialized encoder and decoder with randomly initialized encoder-decoder attention (BERT2BERT)

Table 4. Human evaluation.

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

Table 3. Experiment results.

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**

TAL TECH

EXAMPLE SUMMARIES

Table 5. Example summaries.

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.

TAL TECH

EXAMPLE SUMMARIES

Table 6. Example english summaries.

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.

TAL TECH

TALLINN UNIVERSITY OF TECHNOLOGY

taltech.ee/en