Applying the transformer architecture on the task of headline selection for Finnish news texts

Maria Adamova
St Petersburg State University
Universitetskaya emb., 7-9-11
199034 St Petersburg, Russia
mariia.gorokhova@ya.ru

Maria Khokhlova
St Petersburg State University
Universitetskaya emb., 7-9-11
199034 St Petersburg, Russia
m.khokhlova@spbu.ru

Abstract

The paper evaluates the possibilities of using transformer architecture in creating headlines for news texts in Finnish. The authors statistically analyse the original and generated headlines according to three criteria: informativeness, relevance and impact. The study also substantiates for the first time the effectiveness of a fine-tuned text-to-text transfer transformer model within the task of generating headlines for news articles in Finnish. The results show that there is no statistically significant difference between the scores obtained by the original and generated headlines on the mentioned criteria of informativeness, relevance and impact.

1 Introduction

The headline of any text plays one of the most important roles. Today, online media are ahead of their paper counterparts in terms of popularity, timeliness and mass appeal. The diversity of Internet media emphasises the importance of the task of creating unique and attractive headlines. More and more attention is paid to search engine optimisation, the main goal of which is to attract new users and increase website traffic. Since electronic media is a sphere of accumulation of huge arrays of text data, the question of its optimisation and automation is among the problems of modern computational linguistics. Creating headlines specially adapted to search engines can help increase the visibility of news articles and, consequently, increase traffic. Moreover, creating headlines manually is a labour-intensive and time-consuming process, which does not meet the requirements of the responsiveness of modern electronic media.

The aim of this paper is to critically evaluate the effectiveness of transformer architecture in creating headlines for news texts in Finnish. The paper statistically analyses original and generated headlines according to three criteria: informativeness, relevance and impact. It is also the first attempt to substantiate the effectiveness of a fine-tuned Text-to-Text Transfer Transformer model in the task of generating headlines for news articles in Finnish.

2 Related Work

The popularity of transformers has led to the development of a wide range of pre-trained language models based on this architecture. Transformers' strengths, such as its deep understanding of natural language and its ability to focus on particularly meaningful information in input data, have made models based on it effective tools, including for headers genetation. Applying fine-tuning to pre-trained models on specific tasks allows the knowledge gained from pre-training to be extended to a new task.

The results of fine-tuning pre-trained models to generate news headlines in Russian were demonstrated in (Bukhtiyarov & Gusev, 2020) using the mBART and BertSumAbs models as examples. The Multilingual BART model (mBART) is a language model based on the transformer architecture, consisting of an encoder and an autoregressive decoder, and pre-trained on large-scale monolingual corpora covering 25 languages (Liu et al., 2020). The training process is built on reconstructing a document from its noisy version, which leads to significant improvements in machine translation quality at both sentence level and whole document level (ibid).

Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages, pages 115–122

The BertSumAbs model uses Bidirectional Encoder Representations from Transformers (BERT) as the encoder and a randomly initialised 6-layer transformer as the decoder (Liu & Lapata, 2019). The decoder is pre-trained and the decoder is trained from scratch, so the tuning may be unstable. To overcome the mismatch, the optimisers of the encoder and decoder are separated (ibid).

The fine-tuning of the selected models led to a significant improvement in the results: the ROUGE metric scores increased on average by 2.9 points compared to the previous leading edge performance of the Phrase-Based Attentional Transformer model (Sokolov, 2019) and by 2.0 points compared to CopyNet (Gusev, 2019). The performance of BertSumAbs was also shown to be higher within the task of generating headlines in Russian; moreover, the BertSumAbs model produced headlines of a more abstract nature, while mBART was more prone to copying (Bukhtiyarov & Gusev, 2020). Human evaluation confirmed the effectiveness of the developed models: headlines produced by BertSumAbs were selected by five or more experts in 32% of cases, while original headlines were selected by five or more experts in 28% of cases (ibid).

In (Koppatz et al., 2022), the authors discuss Finnish news headline generation using GPT-2. The key issue in developing the model was the possibility of using the proposed system as an auxiliary tool in real journalistic practice. The generated headlines were expertly evaluated by journalists from a Finnish publishing house. As a result, the generated headlines were very close to being practically usable, and although the concrete implementation is not yet ready to become a fully automated headline generation system, as it still needs human control, the algorithm may well be applicable to potential needs.

3 Models

As part of our work, it was decided to use fine-tuning of the pre-trained model. During the training, experiments were conducted on different variations of the GPT-2 and T5 models for Finnish, provided by the Finnish-NLP community and publicly available on Hugging Face (Finnish-NLP, Hugging Face). In the first stages, both GPT-2 and T5 demonstrated the ability to generate coherent results for the task at hand, but in order to increase the novelty of the study it was decided to continue working with the T5 models.

T5 (Text-to-Text Transfer Transformer) is a model introduced in 2020 by the Google AI team, which has fundamentally the same standard structure as the original transformer, consisting of 12 pairs of encoder-decoder blocks, using the self-awareness mechanism, direct communication network and ecoder-decoder attention described in (Raffel, 2020: 11). One of the characteristics of the structure of T5 models is the use of relative scalar embeddings, which are a type of positional coding. Unlike absolute positional encodings, which assign a unique encoding to each position in a sequence, relative positional encodings encode the relative position between two tokens. To do this, the difference between the positions of the two tokens is computed, which is later used to compute a scalar value that is added to the embedding of the token.

The language model for Finnish, *Finnish T5*, was pre-trained on a combination of 6 datasets: the Finnish language subset of the mC4 dataset and Wikipedia, Yle Finnish News Archive 2011-2018 and 2019-2020, Finnish News Agency Archive (STT) and The Suomi24 Sentences Corpus. The raw datasets were automatically cleaned to improve quality and weed out non-Finnish examples. The result was a raw dataset containing about 76 GB of text. The texts were not lower-cased, so the model retained case sensitivity. A masked language modelling (*MLM*) task was used in the pre-training process.

All variations of the T5 model presented by the Finnish-NLP community were studied and the following ones were selected as a result of preparatory experiments:

1. t5-large-nl36-finnish
2. t5-mini-nl8-finnish
3. byt5-base-finnish

ByT5 is an extension of the T5 model. Compared to other known models (*BERT, T5, GPT*) that rely on learned vocabulary, the ByT5 model works with UTF-8 bytes, thus eliminating the need for text preprocessing. The underlying assumption is that text data is typically stored as a sequence of bytes, the passing of which to the model will allow arbitrary text sequences to be processed (Xue, 2022: 291). In the context of the ByT5 model, Text-to-Text framework reduces to the task of generating a byte sequence based on

Model	Parameters	Number of layers	Dimension of embedding vector (output vector of transformers block)	Dimension of intermediate vector within transformer block (size of feed-forward projection matrix)	Dimension of key/value projection matrix	Number of attention heads
t5-mini	72M	8	384	1536	64	8
byt5-base	582M	18	1536	3968	64	12
t5-large	1425M	36	1024	4096	64	16

Table 1: Main characteristics of the selected models.

some input bytes. Models of this kind are more robust to the presence of noise in the data, since they do not depend on the preprocessing step and are also free from the problems that arise when processing words that are absent in the dictionary. In addition, the parameters that account for the word matrix in large dictionary models can be allocated in byte models for other purposes, such as increasing the number of layers. The characteristics of the models are summarised in Table 1.

4 Methodology

4.1 Data collection and preprocessing

A key element in the fine-tuning process is collecting and organising the data needed to train the model to perform a new task. The study collected a relatively small corpus of 1,600 examples of news text-headline pairs. The news text was not represented in all cases by the whole news article; in most cases the first paragraphs were extracted.

Six online sources of Finnish-language news were selected for data collection (due to their popularity and accessibility):

1. selkosanomat.fi;
2. aamulehti.fi;
3. suomenuutiset.fi;
4. iltalehti.fi;
5. is.fi;
6. mtvuutiset.fi.

The corpus included articles from 2021 to 2024 on various topics. The principle of thematic division did not coincide in all cases in different sources. From each resource we selected from 1 to 6 of the most widely presented thematic headings, for which we then randomly selected about 70 articles. The exceptions were the section on culture (Kultuuri) from suomenuutiset.fi with 51 articles, as there were no more news items in this category, and the political section (Politiikka) from mtvuutiset.fi with 79 articles, as it was necessary to complete the planned number of examples.

Pre-processing of texts included removing paragraph and line breaks, adding the prefix ‘header:’ before the text of the news article, putting dots in headings to indicate the end of a sentence, if necessary.

The text of the news article did not include technical information such as captions to illustrations, information about the author and/or hero of the article, date of publication, subject tags and links. It is typical for aamulehti.fi and is.fi to put the first word of the first paragraph of the article in upper case. Such cases were brought to the standard spelling of words in a sentence. The whole text was not reduced to a single lower case, as the model is case sensitive as a result of pre-training.

As a result, a table in the csv format was generated, consisting of two columns: ‘input’, which contains the news article excerpt with the task prefix, and ‘target’, which contains the corresponding headline with a full stop or other terminating punctuation at the end of it.

4.2 Evaluation

The ROUGE series of metrics is considered to be the baseline for evaluating the performance of the text summarisation task (Maples, 2017: 2), so the metrics are reported in this study. The ROUGE metrics package is based on counting the number of matched units in human-generated and generated texts. The counted units are combinations of n words and the longest matched word sequence (Cohan & Goharian, 2016: 807). ROUGE is calculated as follows:

$$ROUGE[n] - recall = \frac{N_{grampred.} \cap N_{gramref.}}{N_{gramref}}$$

$$ROUGE[n] - precision = \frac{N_{grampred.} \cap N_{gramref.}}{N_{grampred}}$$

$$ROUGE[n] - F1 = 2 \times \frac{recall \times precision}{recall + precision}$$

where $N_{gram\ pred.}$ — n-grams in the generated text; $N_{gram\ ref.}$ — n-grams in the original text.

We used the *Rouge* package to calculate the scores for pairs of original and generated titles and the *FilesRouge* package to calculate the mean when comparing all original and all generated titles.

The ROUGE metric is easy to compute and versatile as it can be applied to data in any language. However, ROUGE considers n-gram matches without considering semantics and grammar, so the results of such metrics cannot give a complete picture of the suitability of the generated materials.

4.3 Questionnaire parameters

There are subjective indicators that are particularly important when it comes to the quality of headlines. The key characteristics of headlines are informativeness, which is not always expressed by the number of N-grams matched, and potential attractiveness, which is entirely based on human perception. For these reasons, human judgement is still considered to be the most reliable way to assess the quality of the generated text. So in this paper, the main reference point for assessing the quality of the model's performance is the results of an expert questionnaire.

The experts were non-first year undergraduate students with a relevant major in linguistics or philology, on the basis of which an assumption is made about the adequacy of their Finnish language proficiency. In the questionnaire given to the participants, they are first asked to read an excerpt of a news article of the same size that the model receives as input, then each of a pair of headlines (original and generated) is evaluated according to three criteria: *informativeness*, *relevance* and *impact*. It is assumed that these are the parameters that are fundamental to the creation of quality headlines and cannot be reliably assessed using metrics.

The criterion of informativeness refers to the extent to which the headline reflects the content of the news article. The informativeness of the headline is considered as "the ability to provide the reader with a relatively adequate representation of the main topic or idea" of the text (Chekut', 2015: 81).

The relevance criterion is understood as the degree of actual correspondence of the proposed headline to the content of the text.

The criterion of impact implies the degree of produced emotional impact on the reader, the degree of "attracting the reader's attention to the subject of the message" (ibid: 82).

To ensure unbiasedness, no indication of which headline was composed by a human and which was suggested by the model is provided before the questionnaire is run. Evaluating each headline against multiple criteria seems more appropriate as it allows for a more nuanced comparison and indicates areas for further work to improve the model. Each headline for each criterion is scored on a five-point scale, where

- 1 - not expressed at all;
- 2 - insufficiently expressed;
- 3 - weakly expressed;
- 4 - well expressed;
- 5 - strongly expressed.

5 Experiments

For the experiments, it was decided to select 10 non-training articles from each of the six sources that were included in the fine-tuning corpus. 10 news articles from the news resource Yle were also added, as this company is one of the leading news providers in Finland: according to Yle's annual report 2023, the weekly reach of Yle's online and mobile services (including Yle Areena) was 81% of the population (Yle's annual reports, 2023).

The fine-tuning did not emphasise learning any particular topic category, so articles for the experiments were also selected from different topics. To fully analyse the model, headlines were generated using the most sophisticated implementation of Finnish T5 (t5-large-nl36-finnish) to see the results of the untuned model. The following hyperparameters were used for this network (see Table 2).

Hyperparameter	Value
max_length (tokenization)	512
truncation	True
max_length (generation)	40
length_penalty	0.2
num_beams	5

Table 2: Hyperparameters for t5-large-nl36-finnish

A case of correct formation of a compound word deserves a positive assessment. In the input text there were the words "*ravintolamoguli*", consisting of two "*ravintola*" ('restaurant') and "*moguli*" ('tycoon'), and "*yökerhojen*", where "*yökerho*" ('nightclub'), '*-jen*' is the plural ending of genitive). Taking these words as a basis, the model generated a new compound word in the title "*yökerhomoguli*".

It is also worth noting the case where the model attempted to capture the headline formation style characteristic of Finnish news outlets. For the resources participating in the study, the following headline structure is typical: a word or phrase summarising the essence of the message, containing a key named entity or indicating the source of information, followed by ':' and the main part of the headline, e.g. "*Ranskalaistutkijat: Kännykkäkielto ei riitä - rajat tarvitaan kaikkeen ruutuaikaan, myös television*" ('French researchers: banning mobile phones is not enough — we need restrictions on all screen time, including TV') or "*Ennuste sen kuin paranee: Luvassa jopa 24 astetta*" ('The forecast is improving: up to 24 degrees is expected'. The model generated the following example headline: "*Tiktok: Nuoret etsivät tietoa ja seuraavat uutisia useimmiten*" ('Young people are more likely to seek information and follow the news').

For the generated and original headlines we calculated the average ROUGE (see Table 3).

	recall	precision	F-measure
Rouge 1	0.08	0.09	0.07
Rouge 2	0.03	0.03	0.02
Rouge 3	0.08	0.09	0.07

Table 3: Evaluation of the selected models.

However, the overall quality of the generated material is far from satisfactory: in most cases, the generated headline was accompanied by superfluous tokens that do not carry any semantics, for example: "*Moni suomalainen europarlamentaarikot eli mepit luopuvat paikastaan EU-parlamentissa. query: query: query:*". In this case, the sequence can be seamlessly cleaned of unwanted tokens without affecting the main part of the generated header.

This resulted in very low scores, indicating low efficiency, but one should keep in mind the already mentioned lack of indicativeness of the evaluation metrics. The rouge_2 values are significantly lower than rouge_1, which is quite expected for the special case of abstract summarisation. The values of rouge_l completely coincide with those of rouge_1, which suggests that the longest common sequence consists of a single word.

5.1 Model fine-tuning: parameters

In the next step, the t5-mini-nl8-finnish network was fine-tuned using the following hyperparameters (see Table 4).

Hyperparameter	Value
test_size	0.2
num_train_epochs	5
per_device_train_batch_size per_device_eval_batch_size	16
eval_steps	40
warmup_steps	50
max_len	100
max_length	20
num_beams	3

Table 4: Hyperparameters for the t5-mini-nl8-finnish network

The training loss (training_loss), which reflects how well the model performs on the training data, was 4.29. The loss on the test data (eval_loss) decreased from 16.92 (before training) to 1.04 (after training).

No further attempts were made to reduce the loss rates, as the error observed in the generated data made it meaningless to continue the search for more optimal training parameters.

Hyperparameter	Value
test_size	0.2
num_train_epochs	5
batch_size	28
max_input_length (TTTrainArgs)	50
max_output_length (TTTrainArgs)	50

max_length (TTSettings)	50
num_beams	8
do_sample	True
top_k	0
top_p	0.8

Table 5: Hyperparameters for byt5-base-finnish

Fine-tuning of the byt5-base-finnish network was carried out with the following values of hyperparameters (see Table 5).

5.2 Model fine-tuning: headlines analysis

As a result of training, the training loss (training_loss) decreased from 2.12 to 1.22, the loss on the test data (eval_loss) from 2.04 (before training) to 1.09 (after training).

Nominatives in this study refer to headings containing a proper name, e.g. "*Koskinen tunnusti syyllistyneensä törkeään talousrikokseen*" ('Koskinen pleaded guilty to aggravated financial crime') or "*Windows95man kertoo, että esiintymisasuun ei tule muutoksia*" ('Windows95man reports that there will be no changes to the suit for the speech').

Dotted headlines are slightly less common. Punctuated headlines were considered to be those that indicate the subject of the news item but do not reveal it in full, e.g. "*Dramaattiset tapahtumat saivat alkunsa lapsen syntymästä*" ('Dramatic events caused by the birth of a child') or "*Demokratian tulevaisuus on turvattava kaikilla tasoilla*" ('The future of democracy must be protected at all levels').

Most of the headlines were of the predicative type, in which the subject of speech and the predicate are included, thus forming an extended thesis, e.g. "*Sateet tulevat maan etelä- ja keskiosassa*" ('Rains will take place in the south and centre of the country') or *'Mestaruusjuhlat alkavat Tampereella maanantaina*" ('Championship celebrations start in Tampere on Monday').

Approximately half of all the headlines received can be categorised as noun headlines. They contain an indication of the general topic of the news article and actively fulfil the function of attracting attention, as the main details of the question asked are not disclosed. For example, "*Millainen on paras leivonnainen?*" ('Which baked goods are the most delicious?') or "*7. tammikuuta 2023 kuolleen naisen lähiomainen kertoo, miten se toimi*" ('A relative of the woman who died on 7 January 2023 tells how it happened').

Slightly fewer headlines can be called transitive, characterised by a direct statement of the main facts. For example, "*Tulppaanifestivaali järjestetäään Amsterdamissa*" ('Tulip Festival will be held in Amsterdam') or "*Tappara voitti Suomen mestaruuden*" ('Tappara won the Finnish championship').

A small proportion of headlines were categorised as opinion. These headlines consist of a reference to a famous person or expert and a subject heading. For example, "*Riikka Purra on huolissaan siitä, mitä hallitus tekee*" ('Riikka Purra is concerned about what the government is doing') or "*Laurence des Cars toivoo, että Mona Lisa saisi oman huoneen*" ('Laurence de Carse hopes Mona Lisa will have her own room').

We also found a few examples of clickbait headlines that aim to evoke emotions in readers as much as possible: "*Italian hallituksella ei ole mitäään tekemistä*" ('The Italian government has nothing to do') or "*Maahanmuuttopolitiikka on karannut käsistä*" ('Immigration policy is out of control').

The network also demonstrated the ability to generate new compound words. For example, there were two words "*taitouinnin maajoukkue*" ('synchronised swimming team'), which in the generated headline merged into one – "*taitouintimaajoukkue*" with a similar meaning. Special attention should be paid to the observance of the rules of alternation of the steps '*nt-nn*'. An example of the actual use of the word '*taitouintimaajoukkue*' was found in the news section of the Finnish-language resource (uimaliitto.fi).

Nevertheless, the occurrence of factual errors could not be avoided. In the headline "*Suomen taitouintimaajoukkue kilpailee EM-kisoissa*" ('The Finnish national team will compete at the European Championship'), the information was distorted, as the news item stated that the Finnish national team would qualify for the European Championship if they had enough points, which had not happened yet. In another case, the second key person was omitted and the meaning of the message was not fully disclosed: "*Lepistö pääsi opiskelemaan musiikkia*" ('Lepistö has been accepted to study music'). In reality, it was about the directors of Sastamala Music College, Sini-Mari Lepistö and Tuomas Honkkila, who are also

students of the institution. The last example points to a surname error: "*Johanna ja Samuel Glassar ovat olleet yhdessä jo vuosia*" ('Johanna and Samuel Glassar have been together for many years'), when in fact the news story refers to "*Johanna Puhakka*" and "*Samuel Glassar*".

There was no difference in the generation of news for Yle and other sources whose materials were used in the training. The only thing that can be noted is that Yle is characterised by shorter headlines with simple constructions, so the generated headlines seem to be closer to the original ones. Equally important is that all generated headlines are characterised by varying degrees of extractiveness, and no cases of direct copying of a fragment from a news article were found. The results of calculating the mean ROUGE score were as follows (see Table 6).

	recall	**precision**	**F-measure**
Rouge 1	0.09	0.10	0.09
Rouge 2	0.02	0.03	0.02
Rouge 3	0.09	0.10	0.09

Table 6: Evaluation of the selected models.

We can note a slight increase in rouge_1 scores compared to the results of the t5-large-nl36-finnish network. Otherwise, the conclusions remained unchanged: rouge_2 is expectedly lower than rouge_1, the longest common sequence presumably consists of one word. The results of byt5-base-finnish were included in the expert questionnaire.

6 Results

The simpler network t5-mini-nl8-finnish as a result of fine-tuning stably generated the token '*Äijä*', recognised as problematic because it does not carry any semantic load, but nevertheless fulfills the role of a full member in the generated headers.

The most productive was the byt5-base-finnish network fine-tuning, which is based on processing text directly at the byte level. This is most likely the reason why the network does not allow the generation of a problematic token. For the previously discussed Finnish T5 implementations, it is assumed that there were errors in the training of the model or tokenisers.

The results obtained with the pre-trained byt5-base-finnish network are diverse: by content, the generated headings were categorised into classes such as nominative, predicative and punctuated; by the techniques used, they were categorised as transitive, nominative, opinion and clickbait headings. Among the 70 headings generated, only three factual errors were found.

The questionnaire included 5 randomly selected news articles that participated in the study. A total of 20 experts participated in the survey. For each criterion, the maximum amount of points a headline could receive if each of the 20 experts gave a score of 5 was 100 points.

The average score on a five-point scale for original headlines on the informativeness was 3.90, for generated headlines reached 3.57. The Student's t-test resulted in a p-value above 0.05, so the null hypothesis of no statistically significant differences between the mean scores on the informativeness is accepted.

For the relevance, the original headlines received a mean score of 3.86 on a five-point scale, while the generated headlines received a mean score of 3.75. As in the case of testing the previous criterion, the t-test confirmed the null hypothesis that there are no statistically significant differences between the sample scores on the relevance criterion.

For original headlines, the mean score on a five-point scale on the impact is 3.69, while for generated headlines it is 3.67. For the t-test, the null hypothesis is also confirmed, namely that there is no statistically significant difference between the sample scores on the impact criterion.

Thus, the hypothesis put forward in this study is confirmed: there is no statistically significant difference between the scores obtained by the original and generated headlines according to the criteria of informativeness, relevance and impact. This means that the results of neural network work are close in quality to the results of human work.

7 Conclusion

In this paper, the results of a practical application of the Transformer-based Finnish T5 model were studied in a task of generating Finnish headlines from input news text. A critical evaluation was carried out, noting both the strengths and promising aspects of the different implementations of the model, as well as points still in need of improvement.

The byt5-base-finnish network performed the best. The experts' evaluations indicate that the

network shows sufficiently high potential in all the criteria considered (informativeness, relevance and impact) to be useful, for example, as an auxiliary tool for creating headlines in Finnish. News content authors can use the headline variant proposed by the network as a basis for further work.

To be fully self-sufficient, the network still needs to improve its reliability, namely to get rid of factual errors. Such a problem can be solved by learning from a larger example dataset, but this also requires more powerful computational resources.

Acknowledgments

Maria Khokhlova acknowledges St etersburg State University for a research project 124032900006-1 (inner ID 95435961).

References

Bukhtiyarov A., Gusev I. 2020. Advances of Transformer-Based Models for News Headline Generation. In Communications in Computer and Information Science. P. 54-61

Chekut' E.P. 2015. Functions of the headline as an actualiser of textual categories 'Funkcii zagolovka kak aktualizatora tekstovyh kategorij'. In *Actual questions of Germanic philology and methods of teaching foreign languages. XIX International Scientific and Practical Conference* 'Aktual'nye voprosy germanskoj filologii i metodiki prepodavanija inostrannyh jazykov. XIX Mezhdunarodnaja nauchno-prakticheskaja konferencija'. P. 81-83.

Cohan A., Goharian N. 2016. Revisiting Summarization Evaluation for Scientific Articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).* P. 806-813. https://aclanthology.org/L16-1130.pdf.

Finnish-NLP. Hugging Face. https://huggingface.co/Finnish-NLP.

Gusev I. O. 2019. Importance of copying mechanism for news headline generation. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*. P. 228-236.

Koppatz M., Alnajjar Kh., Hämäläinen M., Poibeau Th. 2022. Automatic Generation of Factual News Headlines in Finnish. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 100–109, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

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. In *Transactions of the Association for Computational Linguistics.* Vol. 8. P. 726-742.

Liu Y., Lapata M. 2019. Text Summarization with Pretrained Encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, P. 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Maples S. 2017. The ROUGE-AR: A Proposed Extension to the ROUGE Evaluation Metric for Abstractive Text Summarization. 10 p. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2761938.pdf.

Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., Liu P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In *Journal of Machine Learning Research 21*. P. 1-67. https://arxiv.org/pdf/1910.10683.pdf.

Sokolov A. M. 2019. Phrase-Based Attentional Transformer For Headline Generation. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*. P. 615-621

Xue L., Barua A., Constant N., Al-Rfou R., Narang S., Kale M., Roberts A., Z Raffel C. 2022. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. In *Transactions of the Association for Computational Linguistics* 10(01). P. 291-306.

Yle's annual reports 2023. https://drive.google.com/file/d/1wkBa5zWLGC3hh2FUwWHfFG9Vd8jHXMbg/view.