DanSumT5: Automatic Abstractive Summarization for Danish

Anonymous Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Anonymouser Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Anonymousest Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Anonymous Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Anonymous Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
email@domain

Abstract

Automatic abstractive text summarization is a challenging task in the field of natural language processing. This paper presents a model for domain-specific summarization for Danish news articles, DanSumT5; an mT5 model fine-tuned on a cleaned subset of the DaNewsroom dataset consisting of abstractive summary-article pairs. The resulting state-of-the-art model is evaluated both quantitatively and qualitatively, using ROUGE and BERTScore metrics and human rankings of the summaries. We find that although model refinements increase quantitative and qualitative performance, the model is still prone to factual errors. We discuss the limitations of current evaluation methods for automatic abstractive summarization and underline the need for improved metrics and transparency within the field. We suggest that future work should employ methods for detecting and reducing errors in model output and methods for referenceless evaluation of summaries.

1 Introduction

1.1 Automatic text summarization

Automatic text summarization is the automatic generation of short text which condenses the most salient points of a longer text. Much of the research in this field to date has focused on automatic extractive summarization (El-Kassas et al., 2021), which directly extracts and concatenates sentences from the original text. Various methods have been developed for selecting sentences for extractive summaries. Some, such as TextRank (Mihalcea and Tarau, 2004) rely on measures of sentence importance; others rely on sentence location, such as the simple LEAD-3 heuristic, which selects the first three sentences of a given article (Varab and Schluter, 2020). Extractive summarization therefore ensures grammaticality but tends to suffer from a lack of coherence. Additionally, extractive summaries can be afflicted with dangling anaphoras, in which an extracted sentence refers to a preceding sentence not included in the summary (Gupta and Gupta, 2019).

In recent years, the move towards deep learning in natural language processing (NLP) has brought to the forefront automatic abstractive summarization. These summaries paraphrase the main points from the original text by generating artificial text which contains potentially novel sequences not found in the source document. This approach therefore views summarization as a text-to-text problem on which models can be trained and fine-tuned. In many cases, abstractive methods enable more informative summarization since it is possible to convey more information with fewer sentences by rephrasing and compressing the text. However, factual credibility is not ensured and the generated summary may contain statements that are inconsistent with the original text (Maynez et al., 2020; Zhao et al., 2020).

The majority of existing work on automatic summarization has been in English (Azmi and Al-Thanyyan, 2012; Khan et al., 2019). In this paper, we develop an abstractive text summarization

model for Danish news data. We achieve state-of-the-art by fine-tuning mT5 on the DaNewsroom dataset (Varab and Schluter, 2020) consisting of news articles and their corresponding summaries.

1.2 Previous work

This work builds on earlier attempts to employ a similar solution for Danish[1]. This previous work utilized different subsets of the DaNewsroom dataset. However, all subsets were less thoroughly filtered and therefore resulted in more repetitive summaries, possibly inflating performance. An mT5 model trained on a subset representative of the full dataset performed similarly to extractive baselines on the same dataset (Varab and Schluter, 2020). However, due to a high degree of extractive summaries in the dataset, the resulting summaries were predominantly extractive. Models trained on more abstractive subsets of the full dataset displayed more abstractive behaviour, though the resulting summaries were of much lower quality.

The previous studies yielded two tentative but significant insights. Firstly, we established that fine-tuning an mT5 model capable of producing abstractive summaries does not ensure abstractive summarization. Secondly, our work emphasized the need for employing more nuanced quantitative metrics, as well as qualitative inspection of model output, to determine the nature and quality of the generated summaries.

The current work expands on these previous efforts by implementing several changes to the data, model, evaluation, and fine-tuning procedure.

1.2.1 Multilingual language models

In this work, we use mT5 (Xue et al., 2021), a multilingual T5 architecture (Raffel et al., 2020), pre-trained on data from 101 languages. There are several reasons for this, beyond the T5 architecture being well-suited to text summarization tasks. While there is evidence to suggest that monolingual models generally perform better on monolingual tasks (Nozza et al., 2020; Popa and Stefănescu, 2020; Rust et al., 2021) (see also section 6), multilingual models seem to increase performance for smaller languages, likely by leveraging cross-lingual transfer (Conneau et al., 2020; Lauscher et al., 2020). Additionally, it has been suggested that this effect depends on the size of the vocabulary during pre-training for a specific language, as well as lexical and typological proximity between included languages (Lauscher et al., 2020; Rust et al., 2021). Indeed, performance appears to be related to both size of the pre-training corpora of specific languages, as well as linguistic similarity between the target and source language (Arivazhagan et al., 2019; Lauscher et al., 2020).

1.3 Quantitative evaluation metrics

ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L) denote the co-occurrence of unigrams (R-1) and bigrams (R-2) in the generated and reference summary (Lin, 2004; Varab and Schluter, 2020), while R-L identifies the Longest Common Subsequence (LCS) (Briggs, 2021; Lin, 2004). R-L is thereby the only ROUGE measure capable of considering syntax since it rewards longer identical sequence overlaps, while R-1 and R-2 might inflate performance due to many co-occurrences even if the generated summary is syntactically incoherent. ROUGE scores have displayed some correlation with human evaluation of summary fluency and adequacy (Lin and Och, 2004), and are the most commonly used metric for automatic summarization. For example, previous research employing mT5 architectures in non-English languages have obtained ROUGE scores, ranging from R-1 = 15.33, R-2 = 5.12, R-L = 13.85 for Amharic to R-1 = 58.05, R-2 = 26.76, R-L = 37.38 for Vietnamese (Ahuir et al., 2021; Eddine et al., 2022; Farahani et al., 2021; Hasan et al., 2021; Nguyen et al., 2021).

It should be noted that ROUGE scores assess lexical overlap of strings, with no reference to semantic similarity. This is a problem for abstractive summarization, as a qualitatively acceptable summary could consist of completely novel strings with no overlaps relative to the source text, which would be penalised by the ROUGE metric. Alternative evaluation metrics instead utilize the semantic and syntactic relations captured in contextualised word embeddings produced by transformer-based architectures, such as BERT (Devlin et al., 2019). BERTScore (Zhang et al., 2020) calculates the similarity between generated and reference summaries as the sum of cosine similarities between the contextual embeddings for the tokens. Greedy matching is used to compute a score for each embedding in the generated summary and the most similar embedding in the refer-

[1] LINK REDACTED FOR PEER REVIEW

ence summary. In studies reporting BERTScores for non-English mT5 models, performance ranged from 65.80 to 71.55 (Ahuir et al., 2021; Eddine et al., 2022)). In what follows, we evaluate model performance using a combination of both ROUGE and BERTScore.

2 Dataset

2.1 DaNewsroom dataset

To train our models we utilize the DaNewsroom dataset (Varab and Schluter, 2020), inspired by the English Newsroom dataset (Grusky et al., 2018). Danewsroom is currently the only publicly available dataset for Danish summarization and consists of 1.1 million article-summary pairs from various Danish news outlets published over the past 20 years. The summaries were retrieved using a metadata tag and in most cases correspond to the subheading of the article.

2.1.1 Data cleaning

To quantify the degree of abstractiveness of a summary we use the density score (Varab and Schluter, 2020; Grusky et al., 2018), where a lower score indicates a more abstractive summary while higher scores indicate increasingly more extractive summaries, i.e. longer identical sequence overlaps with the original text.

Based on the density measure, we find that the reference summaries in the DaNewsroom dataset are primarily extractive. For the purpose of abstractive summarization, we follow binned density threshold categories (Grusky et al., 2018; Varab and Schluter, 2020), and filter our dataset to contain only abstractive text-summary pairs with a density score between 0 and 1.5.

Several reference summaries in DaNewsroom are short and/or incomplete as in multiple cases, the web scraping used to collect the dataset extracted incorrect or incomplete reference summaries. Examples of short or single-word summaries include '2008', 'Et', 'Behandling', and 'P'. Similarly, the dataset also contains some extremely short articles. In particular, a number of the dataset articles are just one-liners about television scheduling or a paywall:

- Der er lukket for nye kommentarer til denne artikel (*New comments are closed for this article*)

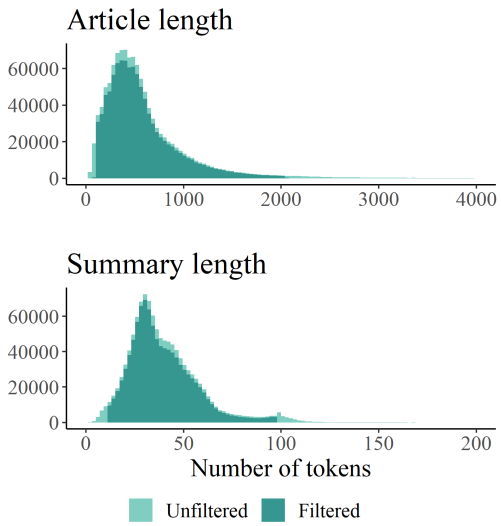

Figure 1: Distribution of summary and article token lengths before and after data cleaning

- Svanerne i Slotsparken har fået fem unger (*The swans in Slotsparken have had five cygnets.*)
- 'DR1 — Tirsdag d. 26. august kl. 20:00 - 20:45'. (*DR1 — Tuesday 26 August at 20:00 - 20:45*)

On the other hand, some of the articles are very long. In some cases, the web scraping concatenated the article with a long thread of comments, resulting in articles consisting of several thousand tokens, including some English content. We further cleaned the dataset by filtering the summaries and articles based on token length, with lower and upper cutoffs defined respectively as the 2nd and 98th percentiles. Additionally, the data was standardized using ftfy (Speer, 2019) and filtered using various heuristic quality filters, including filters for removing repetitious text, text with a high ratio of non-alphabetic tokens, and text with less than two stopwords. This filtering was performed using an implementation similar to textdescriptives (Hansen and Enevoldsen, 2023) and follow a similar approach as used by Rae et al. (2022). The resulting cleaned subset contained 258,146 article-summary pairs.

3 Model

3.1 Infrastructure

We fine-tuned all of our models using the Transformers library (Wolf et al., 2020) and PyTorch (Paszke et al., 2019) as the backend on 4 RTX8000

GPUs. We used Weights and Biases (Biewald, 2020) for experiment tracking and visualizations.

3.2 Training and models specifications

For our hyperparameter tuning, we trained for 1800 steps with a batch size of 120, and validated with the same batch size, using the full validation set. All models were trained on our filtered dataset using an 80-10-10 split. Due to time constraints and limited resources, the hyperparameter search was performed on a small subset of the data (2%), which was deemed large enough to be representative of model performance on the full dataset.

Articles and reference summaries were truncated to a maximum length of 1024 and 128 tokens. Mixed precision training was employed to lower the memory impact during training. The models were trained using the AdamW optimiser (Kingma and Ba, 2017) with a polynomial learning rate schedule using a learning rate of 3.0×10^{-4}, 2000 warmup steps, and a batch size of 16. The models were trained for ten epochs with a dropout rate of 0.01.

For decoding, we used a beam search with two beams and a repetitive 3-gram penalty. Additionally, to encourage longer summaries, we employed minimum and maximum generation lengths of 9 and 128, as well as a length penalty of 5. The best model was determined according to the cross entropy loss and the resulting model was tested using a held-out test set. From this sample of 25,830 article-summary pairs, generated summaries were evaluated by calculating mean density and mean F1-scores for R-1, R-2, R-L (Lin, 2004), and BERTScore (Zhang et al., 2020). The large XLM-RoBERTa model was used to create the embeddings for the BERTScore (Conneau et al., 2020), where our previous work used a base-sized multilingual BERT for the same purpose (Devlin et al., 2019).

The mT5 model has been made available in different sizes, with larger sizes generally leading to improved performance (Xue et al., 2021). We fine-tuned three different model sizes; small, base and large. All other parameters were kept constant.

4 Results

4.1 Quantitative results

Table 1 shows R-1, R-2, R-L, and BERTScore results for our three models. We see that the models perform in accordance with expectation, with larger models performing better.

Models	R-1	R-2	R-L	BERTScore
DanSumT5$_{small}$	21.41	6.21	16.10	88.28
DanSumT5$_{base}$	23.22	7.11	17.65	88.77
DanSumT5$_{large}$	**23.75**	**7.46**	**18.25**	**88.97**

Table 1: Performance metrics by model. Best score is highlighted in bold

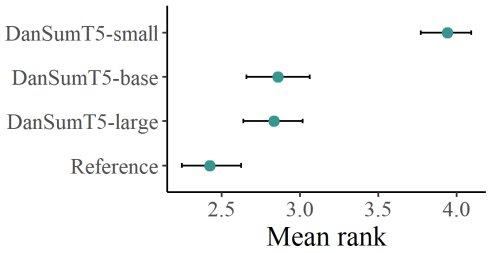

Figure 2: The mean rank obtained for each model. Error bars display the 95% bootstrapped confidence intervals.

The average density of the model-generated summaries was 2.67 for DanSumT5$_{small}$, 2.32 for DanSumT5$_{base}$, and 1.91 for DanSumT5$_{large}$, with lower numbers indicating higher levels of abstractiveness.

4.2 Human evaluation

To manually evaluate the models, we selected generated summaries for 100 randomly sampled articles in the test set. Two raters were then tasked with reading the articles and ranking the three model summaries along with the reference summary according to preference. The raters were unaware of whether the summary was generated by a model or which model produced which summary. The two raters obtained an inter-rater reliability of 0.68 (Cohen's Kappa). Both raters generally preferred the reference summaries over the generated summaries, as shown in Figure 2.

Most of DanSumT5's generated summaries are grammatically correct and cover the main content of the article, though they tend to suffer from factual inconsistencies (See Appendix for five randomly chosen examples of reference and model-generated summaries).

5 Discussion

5.1 Quantitative performance

As seen in Table 1, DanSumT5$_{large}$ achieves higher BERTScores than our previous work, while ROUGE performance is similar or slightly

higher[1]. We have thus achieved state-of-the-art results for Danish abstractive summarization. ROUGE performance still falls short of some similar models fine-tuned for abstractive summarization in other languages, though it remains within the range of published performances (Ahuir et al., 2021; Eddine et al., 2022; Farahani et al., 2021; Hasan et al., 2021; Nguyen et al., 2021).

5.2 Qualitative performance

Human evaluation revealed that the DanSumT5 models generally create summaries of high subjective quality. For instance, the DanSumT5$_{large}$ summaries were preferred over the reference summaries in 21.43% of cases. One limitation of all model summaries is that many of them suffer from factual inconsistencies, as illustrated in Appendix, Table 4, where DanSumT5$_{small}$ states that artists were hit by a snowstorm, though in actuality, an artificial snowstorm was part of their performance.

5.3 Limitations and Future Directions

5.3.1 ROUGE and BERTScore

Given the inherent limitations of ROUGE scores for evaluating abstractive summaries, we opted to employ BERTScore which does not penalise lexical diversity. Since BERTScore is fully differentiable, it could be used to compute a loss metric for optimisation of both training and evaluation (Zhang et al., 2020). However, some limitations of ROUGE also apply to BERTscore, inasmuch as precision, recall and F1 scores depend on the length of the generated and reference summary. For both ROUGE and BERT metrics, generating very short summaries, or using very short reference summaries, might inflate performance since co-occurences in short sequences are disproportionately rewarded for high relative overlap between reference and model output. Additionally, a high BERTScore is also not a guarantee of factual or grammatical consistency.

Quantitative performance indicates similarity with the reference summary; however, this might not be optimal. For instance, since many reference summaries in the DaNewsroom dataset contain dangling anaphoras, these issues could transfer to the generated summaries. A possible way to remedy this would be to utilise anaphoric information, for instance by checking coreferences of the generated summary, and locating errors relating to anaphoric resolution (Steinberger et al., 2007; Sukthanker et al., 2020). Also, despite additional filtering, some reference summaries and articles in the dataset were still incomplete or incorrect. Mismatched or incomplete summary-reference pairs complicate the task of the model, leading to nonsensical or unrelated outputs: In Appendix, Table 3, the reference summary appears unrelated to the accompanying article (not included in Appendix), making comparisons between generated and reference summaries illogical. Alternative quantitative metrics suggest omitting the reference summary and measure performance using only the original text, for instance by calculating the increase in model performance gained by access to the generated summary (Vasilyev et al., 2020).

Recent research shows that most quantitative metrics do not correlate well with human evaluations of generated summaries in important dimensions such as coherence, consistency, fluency, and relevance (Fabbri et al., 2021; Liu et al., 2017). Indeed, it has also been argued that there is no best practice for reliable human evaluation of summaries, and that human evaluations often do not correlate with other human evaluations of the same summaries (Fabbri et al., 2021; Iskender et al., 2021). Though human evaluations are often presented as the gold standard, evaluator demographics, expertise, and task design hugely affect human evaluations (Harman and Over, 2004; Louis and Nenkova, 2013). Different summaries may focus on different aspects of the same article, with no way to objectively conclude which is better.

5.3.2 Quantifying abstractiveness

While state-of-the-art for Danish abstractive summarization, DanSumT5$_{large}$ ROUGE performance falls short of state-of-the-art mT5 models for other languages. Importantly, though, there is no clear definition of what counts as an 'abstractive summarization model'. Many studies on abstractive summarization do not report or evaluate the density of the summaries in their dataset, or of their model-generated summaries. A high ROUGE score could correspond either to an extractive or simply very short summary, thereby inflating ROUGE performance. We note that our model-generated summaries have a wide range of density scores, with the largest model producing more abstractive (low-density) summaries. We obtained high BERTScores, indicating semantic consistency, but lower range ROUGE scores compared to similar literature. Since abstractive sum-

marization allows high lexical diversity, it is not likely to achieve as high ROUGE scores as extractive summarizers, and thereby low ROUGE performance need not be strongly indicative of poor abstractive summarization.

Not all papers present translated examples of model output. Though translation might not be the optimal reflection of non-English summarization, it allows readers to evaluate an approximation of the qualitative results. Lack of transparency, therefore, makes it extremely challenging to evaluate whether the reported summarizers are truly abstractive.

5.4 Model limitations

5.4.1 Factors limiting practical implementation

Automatic summarization requires the model to be factual to the source text, especially for real-world practical implementations. However, none of the evaluated metrics considers the factual correctness of a generated summary (Falke et al., 2019) and does not reward the model for being factually faithful (Maynez et al., 2020).

Many of the reference summaries suffer from dangling anaphoras since they are scraped from the article's subheading, often lacking the context of the title. These were likely not intended to be read as summaries, or even read in isolation from the article's title. This underlines the importance of data quality, since it defines the upper limit for model performance. We found that most of the 100 DanSumT5 summaries inspected for evaluation avoid dangling anaphoras, likely due to only a minority of the dataset suffering from this artefact, and could thus be argued to be better than the reference summaries in this aspect. For example, in the Appendix, Table 2, the reference summary refers to "the superstar", while the model summaries mention the actor by name.

Since DanSumT5 sometimes generates summaries with obviously incorrect content (see Appendix), it is unsuitable for practical implementations where factual accuracy is important, such as summaries of news articles. One possible solution could be to use a separate system to detect such errors in the generated summaries (Falke et al., 2019), such as already existing systems for detection of errors related to quantities (Zhao et al., 2020). Fine-tuning according to this metric or even using it for reinforcement learning could alleviate concerns around accuracy and quality. Another approach uses question asking to evaluate factual consistency by checking if the generated summary and article yield the same answer (Wang et al., 2020). Future work could extend this approach to Danish.

5.4.2 Data limitations and considerations

The current paper demonstrates that it is possible to fine-tune a multilingual model to create a performant text summarization model for a specific domain of Danish language. Other directions for future work must therefore include further experimentation with different datasets. The practical costs of creating high-quality datasets for this task are a challenge for a language such as Danish which is relatively model-rich, compared to similarly sized languages, and data-poor in terms of high-quality data, compared to larger languages. However, as shown in this paper, good results can be obtained by fine-tuning a multilingual model on a web-scraped dataset with minimal data cleaning.

6 Negative Results

During the training of these models, we attempted a few additional ideas, most of which were shown to be unpromising. This section briefly describes these attempts:

1. As suggested by (Abdaoui et al., 2020) we reduced the model by restricting the vocabulary to Danish and English tokens. Meaningful tokens were estimated using a filtered version of the Danish Gigaword (Strømberg-Derczynski et al., 2021)[2] and English Gigaword (Graff and Cieri, 2003). Reducing the model size allowed us to train these models with a larger batch size. While the reduced size led to similar performance to the original architecture for the small and base-sized models, the large pruned model proved highly unstable during training.

2. During the early phases of development, we also experimented with the recently released Danish T5 model, danT5 (Ciosici and Derczynski, 2022). As part of the initial grid search over hyperparamters, we discovered that this monolingual model performed consistently worse compared to the similarly

[2]More information about the specific dataset can be found at `https://huggingface.co/datasets/DDSC/dagw_reddit_filtered_v1.0.0`

sized multilingual mT5 models. This could be due to danT5's novel warm-starting approach which utilises an English T5 model checkpoint, but further experimentation is required in this area.

3. This work seeks to train models for abstractive summarization and thus filters out extractive summaries from the training data based on a density threshold. We experimented with lowering the density threshold for the training set summaries to include increasingly more extractive summaries. While this obtained a lower loss when including moderately more extractive references, the resulting summaries were notably more extractive.

4. In an attempt to avoid potential overfitting on the filtered dataset, we also experimented with training a large mT5 model for only 1 epoch using similar hyperparameters with the exception of a dropout of 0.1. In the human evaluation, this model placed slightly below the base-sized model of DanSumT5.

7 Conclusion

This paper presents DanSumT5, a model achieving state-of-the-art results in automatic abstractive summarization for Danish news articles. These results were achieved by fine-tuning an mT5 model and implementing more robust cleaning of the Danewsroom dataset. While ROUGE performance of our models are in the lower end, BERTScore results are generally high. Human inspection of the generated summaries revealed grammatical and coherent summaries of news articles. This disparity between performance evaluations could reflect that ROUGE penalizes the lexical diversity of abstractive summarization. This emphasizes the need for more transparency in the field, and we argue that research should include more nuanced metrics, as well as manual evaluation of the density and overall quality of the model output. Limitations of our work include data quality and the prevalence of factual errors in the generated summaries. All code related to this project is open-sourced via Github[3], and the model is made freely available via Huggingface for public use[4].

[3]LINK REDACTED FOR PEER REVIEW
[4]LINK REDACTED FOR PEER REVIEW

References

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. Load What You Need: Smaller Versions of Multilingual BERT. *arXiv preprint arXiv:2010.05609*.

Vicent Ahuir, Lluís-F. Hurtado, José Ángel González, and Encarna Segarra. 2021. NASca and NASes: Two Monolingual Pre-Trained Models for Abstractive Summarization in Catalan and Spanish. *Applied Sciences*, 11.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. Technical Report arXiv:1907.05019, arXiv.

Aqil M Azmi and Suha Al-Thanyyan. 2012. A text summarizer for Arabic. *Computer speech & language*, 26(4):260–273. Place: Kidlington Publisher: Elsevier Ltd.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

James Briggs. 2021. The Ultimate Performance Metric in NLP.

Manuel R. Ciosici and Leon Derczynski. 2022. Training a T5 Using Lab-sized Resources. ArXiv:2208.12097 [cs].

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116 [cs]*. ArXiv: 1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs].

Moussa Kamal Eddine, Nadi Tomeh, Nizar Habash, Joseph Le Roux, and Michalis Vazirgiannis. 2022. AraBART: a Pretrained Arabic Sequence-to-Sequence Model for Abstractive Summarization. Technical Report arXiv:2203.10945, arXiv. ArXiv:2203.10945 [cs] type: article.

Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Mehrdad Farahani, Mohammad Gharachorloo, and Mohammad Manthouri. 2021. Leveraging ParsBERT and Pretrained mT5 for Persian Abstractive Text Summarization. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–6.

David Graff and Christopher Cieri. 2003. English Gigaword. Artwork Size: 4089446 KB Pages: 4089446 KB Type: dataset.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Som Gupta and S. K Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Lasse Hansen and Kenneth Enevoldsen. 2023. TextDescriptives: A Python package for calculating a large variety of statistics from text. ArXiv:2301.02057 [cs].

Donna Harman and Paul Over. 2004. The Effects of Human Variation in DUC Summarization Evaluation. In *Text Summarization Branches Out*, pages 10–17, Barcelona, Spain. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. *arXiv:2106.13822 [cs]*. ArXiv: 2106.13822.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of Human Evaluation for Text Summarization: Lessons Learned and Challenges Ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96, Online. Association for Computational Linguistics.

Rahim Khan, Yurong Qian, and Sajid Naeem. 2019. Extractive based Text Summarization Using KMeans and TF-IDF. *International Journal of Information Engineering and Electronic Business*, 11:33–44.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. ArXiv:1412.6980 [cs].

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers. Technical Report arXiv:2005.00633, arXiv. ArXiv:2005.00633 [cs] type: article.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, pages 605–es, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2017. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. Technical Report arXiv:1603.08023, arXiv. ArXiv:1603.08023 [cs] type: article.

Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics*, 39(2):267–300.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. *arXiv:2005.00661 [cs]*. ArXiv: 2005.00661.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Hieu Nguyen, Long Phan, James Anibal, Alec Peltekian, and Hieu Tran. 2021. VieSum: How Robust Are Transformer-based Models on Vietnamese Summarization? *arXiv:2110.04257 [cs]*. ArXiv: 2110.04257.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [MASK]? Making Sense of Language-Specific BERT Models. Technical Report arXiv:2003.02912, arXiv. ArXiv:2003.02912 [cs] type: article.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Cristian Popa and Vlad Stefănescu. 2020. Applying Multilingual and Monolingual Transformer-Based Models for Dialect Identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 193–201, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. ArXiv:2112.11446 [cs].

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv:1910.10683 [cs, stat]*. ArXiv: 1910.10683.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. Technical Report arXiv:2012.15613, arXiv. ArXiv:2012.15613 [cs] type: article.

Robyn Speer. 2019. ftfy. Zenodo. Version 5.5.

Josef Steinberger, Massimo Poesio, Mijail Kabadjov, and Karel Jezek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43:1663–1680.

Leon Strømberg-Derczynski, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rystrøm, and Daniel Varab. 2021. The Danish Gigaword corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. Anaphora and coreference resolution: A review. *Information Fusion*, 59:139–162.

Daniel Varab and Natalie Schluter. 2020. DaNewsroom: A Large-scale Danish Summarisation Dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6731–6739, Marseille, France. European Language Resources Association.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and Answering Questions to Evaluate the Factual Consistency of Summaries. Technical Report arXiv:2004.04228, arXiv. ArXiv:2004.04228 [cs] type: article.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*. ArXiv: 1910.03771.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934 [cs]*. ArXiv: 2010.11934.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. Technical Report arXiv:1904.09675, arXiv. ArXiv:1904.09675 [cs] type: article.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing Quantity Hallucinations in Abstractive Summarization. *arXiv:2009.13312 [cs]*. ArXiv: 2009.13312.

8 Appendix

This appendix consists of five examples of automated summaries generated by our models, along

with the reference summary included in DaNewsroom. Table 3 demonstrates what happens when the article text seems not to be correctly extracted, as discussed in subsection 5.2. In this case, incorrectly scraped web data results in the DanSumT5 models generating seemingly irrelevant summaries, relative to the reference.

Origin	**Summary**
Reference	Viborg-Esbjerg 0-5 Et hattrick fra den tidligere OB-spiller Rajko Lekic sender Esbjerg på førstepladsen i Superligaen. *Viborg-Esbjerg 0-5 A hat-trick from former OB player Rajko Lekic sends Esbjerg to first place in the Super League.*
DanSumT5$_{large}$	Esbjerg fører nu Superligaen efter to sejre på stribe og en 5-0-sejr over Viborg på egen bane. *Esbjerg now leads the Super League after two wins in a row and a 5-0 win over Viborg at home.*
DanSumT5$_{base}$	Esbjerg satte Viborg på plads med en 5-0 sejr over Esbjerg, der dog stadig har svært ved at score i de seneste to kampe i Superligaen. *Esbjerg put Viborg in their place with a 5-0 victory over Esbjerg, who still have difficulty scoring in the last two games in the Super League.*
DanSumT5$_{small}$	Esbjerg satte Viborg på plads i Superligaen efter en skuffende 7.-plads. Viborg er ude af sæsonen med en 6-0-sejr på 6-0 over Viborg. - Sport *Esbjerg put Viborg in place in the Super League after a disappointing 7th place. Viborg is out of the season with a 6-0 win of 6-0 over Viborg. - Sports*

Table 1

Origin	**Summary**
Reference	Superstjernen havde sine hunde med til Australien på ulovlig vis – det kan komme til at koste dyrt *The superstar brought his dogs to Australia illegally – it could be expensive*
DanSumT5$_{large}$	Hollywood-stjernen risikerer fængselsstraf for at have smuglet sine to hunde til Australien, hvor han filmede med "Pirates of the Caribbean"-filmene. • Johnny Depp og Amber Heard sender hundene hjem • Billedserie: Johnny Deeps venner i Australien • Depp: Jeg elsker dem • Politiker: De er slemme at smide os ud • Politi: De elsker hundene • Politiet: De har svært ved at slippe for straf • Depps hunde er farlige for *The Hollywood star faces jail time for smuggling his two dogs to Australia, where he filmed the "Pirates of the Caribbean" movies. • Johnny Depp and Amber Heard send the dogs home • Photo series: Johnny Depp's friends in Australia • Depp: I love them • Politician: They are mean to throw us out • Police: They love the dogs • Police: They have a hard time escaping punishment • Depp's dogs are dangerous to*
DanSumT5$_{base}$	Johnny Depp og hans kone Amber Heards ulovlige medbringelse af deres hunde til Australien kan koste dem op til ti års fængsel. - Kendte *Johnny Depp and his wife Amber Heard's illegal bringing of their dogs to Australia could cost them up to ten years in prison. - Famous people*
DanSumT5$_{small}$	Johnny Depps ulovlige medbringning af deres to hunde i Australien må måske gå i fængsel for deres ulovlig medbringing. - Film og TV *Johnny Depp's illegal bringing of their two dogs in Australia may have to go to jail for their illegal bringing. - Movies and TV*

Table 2

Origin	**Summary**
Reference	Hvad får voksne mennesker over 40 til at bruge flere timer om ugen på at vække eventyret til live i den virtuelle verden? De gængse svar på spørgsmålet bunder i fordomme, mener to inkarnerede computerspillere. Her følger deres egen forklaring fra det virtuelle univers *What makes adults over 40 spend several hours a week bringing the adventure to life in the virtual world? The usual answers to the question are rooted in prejudice, say two inveterate computer gamers. Here follows their own explanation from the virtual universe*
DanSumT5$_{large}$	Interview: Det er svært at forestille sig, at livet uden for murene i Danmark kan blive bedre, end man umiddelbart skulle tro. Det kræver i hvert fald en særlig indsats, siger forfatteren Mathiador, der er aktuel med bogen "Kroniken" om en ung mands liv i skoven bagved ved Grønland. Han fortæller, hvordan det er at blive klogere på livet ude i verden, og hvorfor det er sådan en *Interview: It is hard to imagine that life outside the walls of Denmark can be better than one would initially think. In any case, it requires a special effort, says the author Mathiador, who is current with the book "The Chronicle" about a young man's life in the forest behind in Greenland. He tells what it is like to learn more about life out in the world, and why it is such a*
DanSumT5$_{base}$	Interview: Journalisten Mathiador er en af de danske forfattere, der har skrevet bogen "journalisten", der handler om livet som journalist. Han er aktuel med en ny bog om bogen, der er fyldt med højdepunkter, fortæller han i denne uges udgave af'Journalisten'. I dag er han aktuel i en bog, som er baseret på bogen 'Avisen', som handler direkte om journalisten og skuespilleren, der skriver klummer i Lokalavisen Assens. *Interview: The journalist Mathiador is one of the Danish authors who wrote the book "the journalist", which is about life as a journalist. He is current with a new book about the book, which is full of highlights, he tells in this week's edition of 'Journalisten'. Today he is current in a book which is based on the book 'Avisen', which is directly about the journalist and actor who writes columns in the local newspaper Assens..*
DanSumT5$_{small}$	Den danske forfatter og forfatter Mathiador fortæller om sin evne til at blive kendt for sin kærlighed til livet i Danmark. Han er kendt som " Journalisten", men i dag er han aktuel med at gøre det svært for ham at komme i gang med at finde en plads i rummet, hvor han er i centrum af en række af verdens mest prestigefyldte kunstnere, der elsker at skabe kærligheden. Det er en sjælden kærlighed, når han rejser sig til landet. *The Danish writer and author Mathiador talks about his ability to become known for his love of life in Denmark. He is known as "The Journalist", but today he is current with making it difficult for him to start finding a place in the room where he is at the center of a number of the world's most prestigious artists who love to create love. It is a rare love when he stands up to the country.*

Table 3

Origin	**Summary**
Reference	De danske vikinger leverede en fejlfri præstation under den første prøve ved Eurovision. Men ikke alle elementer fungerede optimalt. *The Danish Vikings delivered a flawless performance during the first rehearsal at Eurovision. But not all elements worked optimally.*
DanSumT5$_{large}$	Det danske Grand Prix-hold havde problemer med sneen til prøverne tirsdag eftermiddag. *The Danish Grand Prix team had problems with the snow for the rehearsals Tuesday afternoon.*
DanSumT5$_{base}$	Rasmussen og hans vikinger havde tirsdag første prøve i Eurovision Song Contest, men sneen var det største problem. Se videoen her på siden og læs mere. *Rasmussen and his Vikings had their first rehearsal in the Eurovision Song Contest on Tuesday, but the snow was the biggest problem. Watch the video on this page and read more.*
DanSumT5$_{small}$	Det danske Grand Prix-håb Rasmussen blev tirsdag eftermiddag ramt af en snestorm, der ifølge DRs underholdningschef blev en rigtig god oplevelse for de danske deltagere i Melodi grand Prix i Portugal. *The Danish Grand Prix hope Rasmussen was hit by a snowstorm on Tuesday afternoon, which according to DR's entertainment manager was a really good experience for the Danish participants in the Melodi Grand Prix in Portugal.*

Table 4

Origin	**Summary**
Reference	Flere anmeldere roser filmen om forhistorien til Harry Potter for at være underholdende og velsmurt. *Several reviewers praise the film about the prequel to Harry Potter for being entertaining and well-oiled.*
DanSumT5$_{large}$	J. K. Rowlings Potter-filmatisering får ros af anmelderne, der roser instruktøren, men savner dramatik og en knap så nuttet hovedrolleindehaver. *J. K. Rowling's Potter film adaptation is praised by critics, who praise the director but missing drama and a not-so-cute lead.*
DanSumT5$_{base}$	Harry Potter-fans kan godt begynde at glæde sig, hvis de følger instruktøren David Yates' særlige præstation i 'Fantastiske skabninger og hvor de findes', mener flere anmeldere - og det er svært at se filmen, siger de. *Harry Potter fans can start to rejoice if they follow the director David Yates' special performance in 'Fantastic Beasts and Where to Find Them', several reviewers believe - and it is difficult to watch the film, they say.*
DanSumT5$_{small}$	J.K. Rowlings manuskript til Harry Potter-serien er særligt hendes præstation, men filmen er i fuld gang med at underholde publikum. *J. K. Rowling's screenplay for the Harry Potter series is especially her achievement, but the film is busy entertaining audiences.*

Table 5