



**University of
Zurich**^{UZH}

**Zurich Open Repository and
Archive**

University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2023

20 Minuten: A Multi-task News Summarisation Dataset for German

Kew, Tannon ; Kostrzewa, Marek ; Ebling, Sarah

Posted at the Zurich Open Repository and Archive, University of Zurich

ZORA URL: <https://doi.org/10.5167/uzh-234387>

Conference or Workshop Item

Published Version



The following work is licensed under a Creative Commons: Attribution 4.0 International (CC BY 4.0) License.

Originally published at:

Kew, Tannon; Kostrzewa, Marek; Ebling, Sarah (2023). 20 Minuten: A Multi-task News Summarisation Dataset for German. In: SwissText 2023: 8th Swiss Text Analytics Conference, Neuchâtel, 12 June 2023 - 14 June 2023. Association for Computational Linguistics, 1-13.

20 Minuten: A Multi-task News Summarisation Dataset for German

Tannon Kew, Marek Kostrzewa^{*,†} and Sarah Ebling

Department of Computational Linguistics, University of Zurich

Abstract

We present *20 Minuten*, a dataset for abstractive text summarisation of German news articles. This dataset fills a gap in summarisation resources for the German language and includes multiple professionally-written and stylistically distinct summaries, along with captioned images and document-level reading times. In this paper, we conduct baseline experiments with mT5 and compare the performance of fine-tuning in both single- and multi-task settings on six different downstream tasks supported by our dataset. Our results reveal that dedicated models are preferable, especially for the more distinct tasks. We make our dataset available for research purposes and provide the code to facilitate its handling.

Keywords

Abstractive Summarisation, Multi-task learning, mT5

1. Introduction

Automatic text summarisation (ATS) is a central task in natural language processing that aims to reduce a long document into a shorter, concise summary that conveys its key points. Extractive approaches to ATS, which identify and copy the most important sentences or phrases from the original text, have long been a popular choice, but these summaries suffer from being incohesive and disjointed. More recently, abstractive approaches to ATS have gained popularity thanks to advancements in neural text generation [1]. Yet, much of the research on ATS has been limited to English, due to its high-resource dominance.

This work introduces a new dataset for German-language news summarisation. Aside from summarisation, the dataset also allows for addressing additional NLP tasks such as image caption generation and reading time prediction. Furthermore, it is multi-purpose since article summaries cover a range of styles, including headlines, lead paragraphs and bullet-point summaries. In order to showcase the versatility of our dataset for different NLP tasks, we conduct experiments using mT5 [2] and compare the performance on six different tasks under single- and multi-task fine-tuning conditions, providing baselines for future work. Our findings show that dedicated models consistently perform better according to automatic metrics.

2. Background & Related Work

Much of the research on extractive and abstractive summarisation has focused on English. A number of large-scale datasets, covering domains such as news [3, 4, 5, 6], scientific articles [7], and legal texts [8], have been made available, contributing to English’s relatively high-resource status in NLP.

Meanwhile, datasets for German summarisation are considerably smaller and based predominantly on Wikipedia and legal texts. Frefel [9] constructed a dataset of approximately 240,000 Wikipedia articles, wherein an article’s introductory paragraph acts as the summary. To determine whether an introductory paragraph is a suitable summary, content overlap between the body of the article and the first paragraph is measured. Aumiller and Gertz [10] introduced a dataset for joint summarisation and simplification for German. This dataset comprises 2,898 document pairs mapping articles from the German children’s lexicon “Klexicon” to a corresponding article from Wikipedia. As pointed out by the authors, source and target articles in this dataset are written independently of each other resulting in diverging or even unrelated content between article pairs. Glaser et al. [11] provided a dataset for summarisation of 100,000 German court rulings collected from publicly available records, wherein the ‘guiding principle’ serves as the summary. In order to automatically generate the guiding principle based on the detailed facts and reasoning of a court ruling, they experimented with a range of approaches for both extractive and abstractive summarisation based on previous work.

Recently, with the growing trend towards large-scale multilingual models, a number of datasets have been introduced for the purpose of training and evaluating multilingual summarisation models. These models typi-

✉ kew@cl.uzh.ch (T. Kew); marek.kostrzewa@uzh.ch (M. Kostrzewa); ebling@cl.uzh.ch (S. Ebling)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



CEUR Workshop Proceedings (CEUR-WS.org)

task	# of instances			avg. tokens		avg. sents		novel n-grams			tgt items	comp. ratio
	train	valid	test	src	tgt	src	tgt	n=1	n=2	n=3		
Headline	40,886	5,097	5,107	379.2	8.0	20.5	1.0	36.7	79.3	90.0	1.0	0.03
Lead	40,905	5,098	5,112	379.2	26.1	20.5	2.0	35.4	77.9	90.8	1.0	0.11
Summary	22,659	2,821	2,824	386.7	45.4	20.8	3.6	32.6	72.0	86.0	3.5	0.15
Caption	37,232	4,632	4,681	378.9	177.6	20.5	12.0	40.3	68.0	77.6	8.0	0.96
Topic	11,005	1,361	1,342	398.3	4.1	21.4	3.2	51.0	–	–	3.2	–
RT	11,416	1,398	1,388	398.0	1.0	21.3	1.0	98.9	–	–	1.0	–
Summary [15]	17,905	200	200	382.9	45.5	22.6	3.6	29.0	68.8	85.9	–	0.15

Table 1

Overview of the *20 Minuten* dataset. The number of novel n-grams in the target texts measures abstractiveness of the different types of summaries. Target items indicates the average number of distinct items to be predicted, which, depending on the task, may correspond to sentences, bullet-points, captions or words. Compression ratio is computed as the average ratio between the length of source articles and that of the target texts. Note that the number of articles paired with a lead summaries exceeds those paired with a headline due to some headlines constituting invalid strings. For comparison, the bottom row provides statistics from the dataset used by Rios et al. [15] for experiments on German text simplification.

cally aim to improve performance on low-resource languages by leveraging knowledge from higher-resource languages (cross-lingual knowledge transfer) [12]. To this end, Scialom et al. [13] extended the CNN/Daily Mail dataset to cover French, German, Spanish, Turkish and Russian. The German portion of this dataset includes 242,982 news articles collected from the *Süddeutsche Zeitung*¹ with descriptive leading paragraphs as summaries. More recently, Calizzano et al. [14] presented the multilingual WikinewsSum dataset which pairs articles written by community members on Wikinews² with their published source articles. Despite not being strictly summaries, the Wikinews articles are treated as long-form summaries of the source articles. This dataset provides 8,126 instances in German among the seven European languages it covers.

In contrast to these existing resources, our dataset focuses on the domain of German news articles and builds upon an earlier version first used by Rios et al. [15] for the purpose of text simplification. In our dataset, each article is paired with *multiple* stylistically different and professionally written summaries. Additionally, we include human-written labels indicating topical content, estimated reading times and captioned pictures associated with each article. Taken together, this dataset opens new research opportunities for German abstractive ATS and multi-task learning.

3. The *20 Minuten* Dataset

We introduce a new dataset that serves as a valuable resource for German text summarisation research. In its current form, the dataset comprises 51,357 news articles

published online by the Swiss news portal *20 Minuten* (*‘20 Minuten’*)³. Originally available in print, *20 Minuten* presents popular news content, covering topics such as local issues, politics, health, sport, education and entertainment in a short and simplified manner. This commitment to style is reflected in the name – the average time spent commuting to work in Switzerland by public transport. The majority of articles follow a standardised format that features a short article summary in the form of a leading paragraph. Over time, however, the publisher has enhanced the way that content is presented online by adding bullet-point summaries (titled *‘Darum geht’s’*), estimated reading times and extensive image carousels, among others.

In contrast to the dataset collected by Rios et al. [15], which made use of data from *20 Minuten* for experiments on German text simplification, the data collection methods used to construct the current dataset were designed to capture *all* of the content available with a given article in order to enrich this resource for multiple NLP tasks with a particular focus on summarisation. Table 2 provides a detailed overview of the differences between these two datasets.

3.1. Dataset Creation

To construct the dataset, we collected articles from *20min.ch* dating back to July 2010 that contained at least one valid summary section. As a valid summary, we considered either the bullet-point summary, which constitutes an executive summary, or a captioned image carousel, which contains a series of thematically related pictures. In addition to the summary texts, we maintained the document’s paragraph structure as well as additional metadata such as author IDs and estimated

¹<https://www.sueddeutsche.de/>

²<https://www.wikinews.org>

³<https://www.20min.ch/>

Attribute	Our dataset	Rios et al. [15]
Article Id	☒	
URL	☒	
Create Date	☒	
Publish Date	☒	
Update Date	☒	
Reading Time	☒	
Author	☒	
Category	☒	
Title Header	☒	
Title	☒	
Lead	☒	
Summary	☒	☒
Article Content	☒	☒
Picture URLs	☒	
Picture	☒	
Description		
Paragraph Structure	☒	

Table 2
Article attributes in the new *20 Minuten* dataset compared to the initial version used by Rios et al. [15].

reading times. As a result, we have extended the applicability of the corpus, which we aim to demonstrate through our experiments. To facilitate future research, we provide training, test and validation splits. Articles were randomly assigned to one of these splits at a ratio of 80/10/10. An example of a dataset document as a JSON file is shown in Appendix B.

3.2. Dataset Statistics and Tasks

Table 1 presents descriptive statistics of the *20 Minuten* dataset broken down by tasks. We use ‘source’ to refer to the article’s content, while ‘target’ denotes the task-specific target text, e.g. bullet-point summaries, image captions, etc.

Headline Generation Headline generation has typically been studied as a short sentence summarisation task, where the goal is to generate a single sentence that summarises the contents of a larger text [16, 17, 18, 19]. This task is particularly challenging for traditional summarisation methods, as sufficiently communicating the contents of an article in a single sentence requires the generated headlines to be highly abstractive.

Lead Generation Article leads are often used by journalists to grab the reader’s attention and to quickly communicate the gist of an article. Typically, leads resemble short paragraphs consisting of only one to two cohesive sentences. Therefore, similar to headlines, they are

highly abstractive. In our dataset, lead summary generation is reminiscent of the task associated with the XSUM and MLSUM datasets [4, 13].

Summary Generation Bullet-point summaries contain standalone statements highlighting the most important aspects of an article. In some cases, bullet points may be very similar to paragraph initial sentences in the text and thus may benefit from extractive approaches [20, 4]. However, in *20 Minuten*, they are only slightly less abstractive than other types of summaries. To construct a full summary from multiple bullet points, we concatenate them into a single string, similar to the approach taken by Hermann et al. [3].

Caption Generation In *20 Minuten*, articles are often accompanied with a captioned picture gallery. While images are typically closely related to the article’s content, there is no guarantee that the captions accurately reflect the contents of the article. This is also indicated by the high number of novel unigrams found in the target texts. While we restrict our investigation to generating image captions based solely on the article text, we also note that this task presents interesting opportunities for researchers to combine article texts with pictures in multi-modal generation tasks [e.g. 21, 22].

Topic Prediction In our dataset, articles are frequently paired with an three to four topical keywords. Unlike typical text classification scenarios where target labels are drawn from a pre-defined set of candidates in either a ‘closed-world’ [23, 24, 25] or ‘open-world’ setting [26, 27, 28], topical keywords in *20 Minuten* are not restricted to broad umbrella terms like ‘sport’ or ‘politics’. While such labels are used sporadically, most of the topical keywords are reminiscent of trending topic markers that do not necessarily appear in the article content, e.g. ‘Coronavirus’, ‘Fussball’, etc. As a result, topics constitute a truly open set, making this task highly suited to abstractive, generative methods. Table 5 in Appendix A provides a breakdown of the most frequent topics.

Reading Time Prediction A number of articles in the dataset are also annotated with an estimated reading time (RT). Document-level reading times are useful for both potential readers and authors to establish realistic expectations of an article’s complexity and the time investment required to consume it, especially for web-based content [29, 30]. Reading times for *20 Minuten* are estimated according to word count thresholds, typically ranging from one to ten minutes. Previous work on English document-level reading time prediction has shown that simpler linear models typically outperform more complex neural models at this task [31].

4. Experimental Setup

We adopt a unified text-to-text modelling framework to establish baseline performance on a range of tasks supported by our dataset. Specifically, we fine-tune mT5-base [2] on each task, first in a single-task setting and then in a unified multi-task setting. mT5 is a large multi-lingual pre-trained language model (PLM) based on the English-only T5 model [32] and consists of an encoder-decoder Transformer architecture [33]. While earlier PLMs consisted of either a standalone encoder for classification and representation tasks [34] or a decoder for generative tasks [35], mT5’s architecture offers two major benefits. Firstly, it reduces the complexity of multi-task NLP pipelines by removing the need for specialised task-specific frameworks. Secondly, it may also lead to improved performance since sharing a unified set of parameters between related tasks can be beneficial [36].

An important consideration when evaluating the performance of PLMs on new datasets collected from the web is whether or not test examples appear in the pre-training data. We inspected the publicly available mC4 dataset⁴ used to pre-train mT5 for instances from 20min.ch and found extensive overlap. Specifically, we used the descriptive article title included in the URL and identified approximately 560 and 600 compromised articles in the validation and test sets respectively. While mC4’s format does not explicitly make use of supervised labels associated with each of the downstream tasks, evaluating on articles seen during training may still lead to inflated scores. We therefore filtered out any instances from our test and validation sets that appear in mC4, matching these by the unique article headline included in the URL.

As a result of this filtering, our training runs and experiment evaluations are computed on splits that differ slightly from the total number of items for each task in the dataset based on our pre-defined splits (reported in Table 1). Table 3 shows the final numbers of items used in this set of experiments. For replicability and comparison, we provide both validation and test sets as part of our dataset release.

To benchmark the performance for each task in the multi-purpose *20 Minuten* dataset, we compare fine-tuning in single- and multi-task settings. For each of the tasks, we select a distinct natural-language prefix that is intended to aid the model in learning the task (e.g., “summarise:” for bullet-point summary generation, “generate title:” for headline generation). In the single-task setting, we fine-tune mT5 on each of the six tasks in isolation, resulting in one model per task. In the multi-task setting, we concatenate the training, testing and validation splits from all downstream tasks, and shuffle the resulting training set. The training procedure remains the same as in

⁴<https://huggingface.co/datasets/mc4>

task	valid	test
Headline	4,536	4,505
Lead	4,537	4,508
Summary	2,707	2,696
Caption	4,096	4,105
Topic	1,319	1,297
RT	1,396	1,386

Table 3

Number of items for each task after filtering overlapping content found in mC4. Note that the number of training instances is unchanged from the statistics presented in Table 1.

task	Rouge-1	Rouge-2	Rouge-L	Ex. (%)
Headline	19.03	7.77	17.85	–
	14.27	5.27	13.34	–
Lead	26.65	10.18	20.79	–
	24.88	8.92	18.85	–
Summary	29.89	11.35	21.39	–
	29.29	10.86	20.62	–
Caption	32.13	14.15	20.16	–
	31.33	13.53	19.63	–
Topic	–	–	–	44.42
	–	–	–	30.82
RT	–	–	–	66.59
	–	–	–	55.12

Table 4

Performance of mT5 after fine-tuning in single-task settings vs. multi-task settings. Shaded rows indicate the results for the multi-task model.

the single-task setting, but this time training batches consist of a mix of downstream tasks. Therefore, the model has to rely on the task-specific prefix to achieve each task. For details on fine-tuning settings and hyperparameters, see Appendix C.

For automatic evaluation, we follow previous work and report ROUGE-N and ROUGE-L [37] for summarisation tasks. For topic and reading time prediction, we focus on the number of targets the model predicts correctly by reporting the percentage of exact matches. We make our dataset and scripts for efficient handling available at <https://github.com/ZurichNLP/20Minuten>.

5. Results & Discussion

Table 4 presents our experiment results. Comparing the performance of mT5 on each downstream task between the single and multi-task settings shows that the single-task fine-tuning is more beneficial across the board. For longer-form generative tasks (e.g. lead, summary and caption generation) the multi-task model shows more competitive performance. In contrast, for the more concise and constrained generative tasks, such as headline

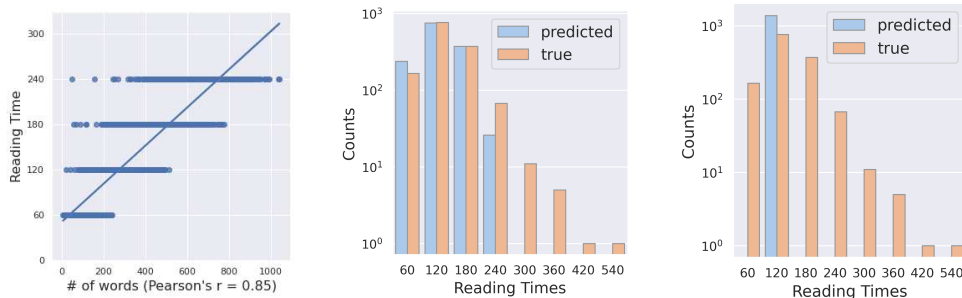


Figure 1: Left: the relationship between ground-truth reading times and word counts in *20 Minuten* articles. Centre, Right: the distribution of predicted reading times from the single-task fine-tuned model and the multi-task fine-tuned model.

and topic generation, the task-specific models perform better.

For the topic prediction task, the dedicated model correctly predicts 44% of the topical keywords on average compared to just 30% in the multi-task setting. However, we also note that relying only on considering exact matches for this task is limiting. Oftentimes, predicted topics could be considered semantically equivalent to the ground truth labels, but they are too harshly penalised due to synonymy, (e.g. ‘Demonstrationen’ vs. ‘Proteste’) or alternative surface realisations (e.g. ‘Russlands Präsident Wladimir Putin’ vs. ‘Wladimir Putin’). Taken together, our results agree with previous works in that training in a multi-task setting is most beneficial for related tasks [36].

Meanwhile, for the uniquely different task of reading time prediction, the model performs poorly in the single-task setup and fails entirely in the multi-task setting where it simply predicts the most common label for all texts in the test set (2 minutes) (see Figure 1). Given the apparent simplicity of the task and a strong correlation with the article word count, we find this result particularly surprising and elaborate on potential causes.

Publicly available tools for estimating document-level reading times typically rely on a simple formula that considers an article’s word-count and the average adult reading speed. While a text’s complexity or thematic content may also play a significant role in true reading times, times provided for articles in *20 Minuten* are estimated by humans based on this simple word-count heuristic and binned into 1-minute intervals.

The plot on the left of Figure 1 shows the spread of word counts for each annotated reading time in the corpus. As can be seen, there are a number of articles with low word counts labelled with longer reading times. A closer inspection of these instances reveals that the article’s textual content is indeed not the only factor used to estimate reading time, but embedded content from external sources is also considered (e.g. videos, images, and

message threads). This results in a large degree of noise, making it hard for the model to learn this task effectively based on the article text via supervised fine-tuning. We suspect that this issue is primarily responsible for mT5’s poor performance in this task and aim to improve these labels in the next iteration of the dataset.

In addition, the plots to the right in Figure 1, show the distribution of reading times predicted by the models, with the single-task model in the middle and the multi-task model on the right. As can be seen, the distribution of predictions from the dedicated model matches the true distribution far better, yet the model tends to undershoot the estimated reading times, while the multi-task model is underfit and always predicts the majority class. We suspect that the multi-task model’s failure to learn this task may also be a consequence of our task sampling method. In this setting, training mini-batches contain a mixture of tasks, which are sampled according to their frequency in the training data. Since reading time labels exist for only a small portion of the total training set, it is an under-represented task and may not have been seen enough during training.

Finally, we provide examples of model-generated outputs in Appendix D. A manual inspection of these outputs reveals that under both types of training conditions, the model is indeed able to accurately produce the stylistically different targets associated with each task. The model successfully learns to produce a set of distinct bullet points for the summary task as opposed to a short, coherent paragraph for the lead generation task, indicating that these surface-level and form features are quickly and easily captured by the model. In most cases, the generated summaries are largely fluent and convincing, however, a deeper investigation is required to assess whether these summaries are satisfactory and accurate to the source text.

We intend to perform an extensive qualitative analysis of mT5’s generations in order to improve upon the baselines set here. In addition, future work could also

investigate the use of semantic clustering techniques for evaluating the topic generation task to avoid overly harsh penalties.

6. Conclusion

We have introduced a new dataset for multi-purpose summarisation of German news articles. By providing multiple, stylistically different summaries for the same input articles, the dataset serves as a valuable resource for research in German-language summarisation. We established benchmark performance using mT5 for each of the tasks laid out in this work and showed that fine-tuning in a single-task setting leads to optimal performance according to automatic metrics. We hope that this dataset will contribute to further research on German summarisation and NLP.

Acknowledgements

We would like to thank TX Group AG for their support in this project and for allowing the broader research community to make use of the *20 Minuten* dataset. In addition, we thank the anonymous reviewers for their valuable feedback.

References

- [1] I. Sutskever, O. Vinyals, Q. V. Le, Sequence to Sequence Learning with Neural Networks, arXiv:1409.3215 [cs] (2014). arXiv:1409.3215.
- [2] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 483–498. URL: <https://aclanthology.org/2021.naacl-main.41>. doi:10.18653/v1/2021.naacl-main.41.
- [3] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, *Advances in neural information processing systems* 28 (2015).
- [4] S. Narayan, S. B. Cohen, M. Lapata, Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1797–1807. URL: <https://aclanthology.org/D18-1206>. doi:10.18653/v1/D18-1206.
- [5] M. Grusky, M. Naaman, Y. Artzi, Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 708–719. URL: <https://aclanthology.org/N18-1065>. doi:10.18653/v1/N18-1065.
- [6] E. Sandhaus, The New York Times Annotated Corpus LDC2008T19, Linguistic Data Consortium 6 (2008). doi:10.35111/77ba-9x74.
- [7] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, N. Goharian, A discourse-aware attention model for abstractive summarization of long documents, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 615–621. URL: <https://aclanthology.org/N18-2097>. doi:10.18653/v1/N18-2097.
- [8] E. Sharma, C. Li, L. Wang, BIGPATENT: A large-scale dataset for abstractive and coherent summarization, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2204–2213. URL: <https://aclanthology.org/P19-1212>. doi:10.18653/v1/P19-1212.
- [9] D. Frefel, Summarization corpora of Wikipedia articles, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6651–6655. URL: <https://aclanthology.org/2020.lrec-1.821>.
- [10] D. Aumiller, M. Gertz, Klexikon: A German Dataset for Joint Summarization and Simplification, arXiv:2201.07198 [cs] (2022). arXiv:2201.07198.
- [11] I. Glaser, S. Moser, F. Matthes, Summarization of German court rulings, in: Proceedings of the Natural Legal Language Processing Workshop 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 180–189. URL: <https://aclanthology.org/2021.nllp-1.19>. doi:10.18653/v1/2021.nllp-1.19.
- [12] Y. Cao, X. Wan, J. Yao, D. Yu, MultiSumm: Towards a unified model for multi-lingual abstractive summarization, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 11–18. doi:10.1609/aaai.v34i01.5328.

- [13] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, J. Staiano, MLSUM: The multilingual summarization corpus, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 8051–8067. URL: <https://aclanthology.org/2020.emnlp-main.647>. doi:10.18653/v1/2020.emnlp-main.647.
- [14] R. Calizzano, M. Ostendorff, Q. Ruan, G. Rehm, Generating extended and multilingual summaries with pre-trained transformers, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 1640–1650. URL: <https://aclanthology.org/2022.lrec-1.175>.
- [15] A. Rios, N. Spring, T. Kew, M. Kostrzewa, A. Säuberli, M. Müller, S. Ebling, A New Dataset and Efficient Baselines for Document-level Text Simplification in German, in: Proceedings of the Third Workshop on New Frontiers in Summarization, Association for Computational Linguistics, Online and in Dominican Republic, 2021, pp. 152–161. doi:10.18653/v1/2021.newsum-1.16.
- [16] M. Banko, V. O. Mittal, M. J. Witbrock, Headline generation based on statistical translation, in: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Hong Kong, 2000, pp. 318–325. URL: <https://aclanthology.org/P00-1041>. doi:10.3115/1075218.1075259.
- [17] A. M. Rush, S. Chopra, J. Weston, A neural attention model for abstractive sentence summarization, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 379–389. URL: <https://aclanthology.org/D15-1044>. doi:10.18653/v1/D15-1044.
- [18] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gülçehre, B. Xiang, Abstractive text summarization using sequence-to-sequence RNNs and beyond, in: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 280–290. URL: <https://aclanthology.org/K16-1028>. doi:10.18653/v1/K16-1028.
- [19] S. Chopra, M. Auli, A. M. Rush, Abstractive sentence summarization with attentive recurrent neural networks, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 93–98. URL: <https://aclanthology.org/N16-1012>. doi:10.18653/v1/N16-1012.
- [20] A. See, P. J. Liu, C. D. Manning, Get to the point: Summarization with pointer-generator networks, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1073–1083. URL: <https://aclanthology.org/P17-1099>. doi:10.18653/v1/P17-1099.
- [21] J. Cho, J. Lei, H. Tan, M. Bansal, Unifying vision-and-language tasks via text generation, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 1931–1942.
- [22] J. Zhu, H. Li, T. Liu, Y. Zhou, J. Zhang, C. Zong, MSMO: Multimodal summarization with multimodal output, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4154–4164. URL: <https://aclanthology.org/D18-1448>. doi:10.18653/v1/D18-1448.
- [23] Y. Meng, Y. Zhang, J. Huang, C. Xiong, H. Ji, C. Zhang, J. Han, Text classification using label names only: A language model self-training approach, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 9006–9017. URL: <https://aclanthology.org/2020.emnlp-main.724>. doi:10.18653/v1/2020.emnlp-main.724.
- [24] P. Petrenz, B. Webber, Squibs: Stable classification of text genres, *Computational Linguistics* 37 (2011) 385–393. URL: <https://aclanthology.org/J11-2004>. doi:10.1162/COLI_a_00052.
- [25] J. Rennie, 20 Newsgroups Data Set, <http://qwone.com/~jason/20Newsgroups/>, 2008.
- [26] L. Shu, H. Xu, B. Liu, DOC: Deep open classification of text documents, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2911–2916. URL: <https://aclanthology.org/D17-1314>. doi:10.18653/v1/D17-1314.
- [27] S. Prakhya, V. Venkataram, J. Kalita, Open set text classification using CNNs, in: Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017), NLP Association of India, Kolkata, India, 2017, pp. 466–475. URL: <https://aclanthology.org/W17-7557>.
- [28] G. Fei, B. Liu, Breaking the closed world assumption in text classification, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp.

- 506–514. URL: <https://aclanthology.org/N16-1061>. doi:10.18653/v1/N16-1061.
- [29] M. Marchese, Why it’s important to add an estimated reading time | Tempesta Media Blog, 2020.
- [30] M. Sall, The Optimal Post is 7 Minutes, 2013.
- [31] O. Weller, J. Hildebrandt, I. Reznik, C. Challis, E. S. Tass, Q. Snell, K. Seppi, You don’t have time to read this: An exploration of document reading time prediction, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 1789–1794. URL: <https://aclanthology.org/2020.acl-main.162>. doi:10.18653/v1/2020.acl-main.162.
- [32] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *Journal of Machine Learning Research* 21 (2020) 1–67.
- [33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is All you Need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., 2017, pp. 5998–6008.
- [34] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *arXiv:1810.04805 [cs]* (2019). [arXiv:1810.04805](https://arxiv.org/abs/1810.04805).
- [35] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners, *OpenAI Blog* 1 (2019).
- [36] S. Ruder, An Overview of Multi-Task Learning in Deep Neural Networks, *arXiv:1706.05098 [cs, stat]* (2017). [arXiv:1706.05098](https://arxiv.org/abs/1706.05098).
- [37] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: *Text Summarization Branches Out*, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: <https://aclanthology.org/W04-1013>.

A. Popular Topics in 20 Minuten

Topic	# of Occurrences	Topic	# of Occurrences
Coronavirus	2,144	⋮	
Wirtschaft	1,463	Frankreich	208
USA	966	Kriminalität	207
Schweiz	721	Kinder	207
Deutschland	582	TV	203
Fussball	579	Pandemie	203
Corona-Impfstoff	403	Österreich	200
Russland	395	Fussball-Nationalteam	187
Sport	370	Arbeit	182
Unfall	369	Royals	177
Ukraine-Krieg	326	Bundesrat	175
Geld	320	China	169
Todesfall	283	Mord	168
Musik	256	Corona-Fallzahlen	164
Impfen	252	Forschung	159
Instagram	240	Tiere	158
News	239	Donald Trump	154
Social Media	225	Politik	151
Gesundheit	222	Polizei	150
Detailhandel	222	Job / Beruf	147
Grossbritannien	220	Gewalt	147
Kantonspolizei	217	Tennis	146
Showbusiness	211	Migros	145
Italien	210	Bern	145
Liebe	209	Super League	143
⋮		Covid-Zertifikat	139

Table 5

Most frequent topical keywords for articles in *20 Minuten*. Unlike typical text classification labels, these topics are more fine-grained and form a truly open set, with many *hapax legomena*. This makes topic prediction in the *20 Minuten* dataset a challenging generative task.

B. Dataset Example

Below we provide an example article taken from the *20 Minuten* dataset. As can be seen, the main content retains the paragraph structure of the published article and the stylistically different types of summaries (title, lead, summary and pictureText).

```
{
  "id": 351770894604,
  "date": "2021-10-27T04:08:28.494485539Z",
  "dateCreated": "2021-10-24T18:06:48.894Z",
  "datePublished": "2021-10-24T20:04:08.000Z",
  "dateUpdated": "2021-10-24T20:04:08.000Z",
  "readingTime": 60,
  "author": "dpa/bho",
  "category": [
    "Ed Sheeran",
    "Musik",
    "Coronavirus"
  ],
  "url": "/story/ed-sheeran-wurde-positiv-auf-corona-getestet-351770894604",
  "titleHeader": "Alle Termine abgesagt",
  "title": "Ed Sheeran wurde positiv auf Corona getestet",
  "lead": "Wie der britische Sänger am Sonntag auf Instagram mitteilte, hat er sich mit dem Coronavirus infiziert. Nun will er Interviews und Auftritte von zu Hause aus abhalten.",
  "picture": [
    "https://cdn.unitycms.io/image/ocropped/1200,1200,1000,1000,0,0/dd-SWUVwxE/01TK2NtQaPjA4B5k1G8S4i.jpg",
    "https://cdn.unitycms.io/image/ocropped/1200,1200,1000,1000,0,0/HPCSKGROswc/Aa07dkFR4z8A8q_GwxJshQ.jpg",
    "https://cdn.unitycms.io/image/ocropped/1200,1200,1000,1000,0,0/fcyNZwJ1A3s/6WFwoC8q4w58H4Q71iX5j_.jpg",
    "https://cdn.unitycms.io/image/ocropped/1200,1200,1000,1000,0,0/khyqyV8kKco/4nOqQQcQK1f8z6TQdY0uTa.jpg",
    "https://cdn.unitycms.io/image/ocropped/1200,1200,1000,1000,0,0/gAguTnK8H9g/2EmZV1cP4j4BZ_Ym6dTHOM.jpg",
    "https://cdn.unitycms.io/image/ocropped/1200,1200,1000,1000,0,0/zn_ohR--o-I/5b1UWjCdq988K0AplFP0kS.jpg"
  ],
  "pictureText": [
    "Der Musiker Ed Sheeran hat sich mit dem Coronavirus infiziert.",
    "Dies teilte der Popsänger am 24. Oktober 2021 auf Instagram mit.",
    "Das Coronavirus macht auch vor diversen Prominenten nicht halt. So zum Beispiel dem Schauspieler Tom Hanks, der im März am Coronavirus erkrankte.",
    "Auch American-Football-Superstar Tom Brady infizierte sich mit dem Virus. Nach seiner Aussage wurde er im Februar nach einer Bootsparade positiv getestet.",
    "Grimes, hier noch mit Ex-Freund und Multimilliardär Elon Musk, hat sich nach eigenen Angaben im Januar 2021 mit Corona infiziert.",
    "Der Talkmaster Larry King starb an den Folgen einer Covid-19-Infektion, die er sich im Dezember 2020 zugezogen hatte."
  ],
  "summary": {
    "summary-351770894604-0": "Ed Sheeran hat sich mit dem Coronavirus infiziert.",
    "summary-351770894604-1": "Dies teilte der britische Pop-Superstar am Sonntag auf Instagram mit.",
    "summary-351770894604-2": "Zurzeit befindet sich Sheeran in Selbstisolation zuhause."
  },
  "content": {
    "content-351770894604-0": "Der britische Sänger Ed Sheeran hat sich mit dem Coronavirus infiziert. Er sei positiv getestet worden, teilte Sheeran am Sonntag bei Instagram mit. Er befinde sich in Selbstisolation. Er werde Interviews und Auftritte von zuhause aus absolvieren. Auf der offiziellen Webseite des 30-Jährigen waren keine Auftritte vor April
```

```

    aufgelistet.",
"content-351770894604-1": "Ob der Sänger gegen das Coronavirus geimpft ist, ist nicht ganz klar.
    Jedoch warb er an einem Auftritt bei Moderator James Corden mit einer abgewandelten Version
    seines Hits «Shape of You» dafür, sich mit einer Impfung gegen das Coronavirus zu schützen.",
"content-351770894604-2": "Das neue Studioalbum «=» von Sheeran soll am 29. Oktober erscheinen.
    Es ist bereits das fünfte Album, das der Brite veröffentlicht. Für 2022 plant der vierfache
    Grammy-Preisträger eine Konzerttournee, bei der er auch im Zürcher Letzigrund auftreten
    wird. 20 Minuten verlost Tickets."
}

```

C. Fine-tuning Hyperparameters

We aimed to use comparable computation budgets for all training runs and thus fine-tune mT5 on each task with a maximum number of update steps. For the multi-task setting, we double the effective batch size and total number of update steps to account for the increased size of the concatenated training set. In both settings, the 1-best model is kept according to loss on the validation set. Finally, in the single-task setting, the maximum target length for each task can be set based on its characteristics, ranging from 4 (for reading time prediction) to 256 (for the caption generation task), whereas in the multi-task setting all tasks must have the same maximum target length (256) when batched together. Table 6 details the hyperparameters used for fine-tuning mt5 on the different downstream tasks. All training runs are done on a single NVIDIA Tesla V100 SXM2 with 32GB memory.

Hyperparameter	Value
Floating point precision	32
Optimisation metric	loss
Learning rate	0.00005
Learning rate scheduler	linear decay
Max grad. norm.	1.0
Optimiser	adamw_hf
Adam betas	0.9, 0.999
Label smoothing	0.0
Early stopping patience	5
Early stopping threshold	0.01
Max update steps	4k (single), 8k (multi)
Save interval	200
Maximum input length	512 (maximum for mt5-base)
Maximum output length	4 (reading time), 32 (topic, headline), 64 (lead), 128 (summary), 256 (caption, multi)

Table 6
Settings used for fine-tuning mT5.

D. Model Output Examples

The tables below show the model generated outputs for all six tasks for two randomly selected test-set articles. Outputs generated by the single-task models are shown on the left, while multi-task model outputs are on the right. Tasks associated with multiple target items (e.g. summary, caption and topic generation) are predicted as a single string, with a dot point, ‘•’, identifying each item. This makes it easy to identify and filter repetitions in postprocessing.

Table 6

Article: “Jede Schweizerin und jeder Schweizer darf monatlich fünf Corona-Selbsttests umsonst in der Apotheke abholen. Doch das gilt nur für ungeimpfte Personen. Wer die erste Impfung erhalten hat, muss für den Corona-Test zahlen. Dabei kosten fünf davon bis zu 60 Franken. Das will Aldi nun ändern: Ab Donnerstag verkauft der Discounter die Selbsttests in seinen Filialen, wie es in einer Medienmitteilung heisst. Eine Packung kostet 19.99 Franken. Darin sind fünf Tests enthalten, pro Stück bezahlt man dann also noch rund vier Franken. Das Angebot gilt solange der Vorrat reicht, heisst es weiter. Dabei handelt es sich laut Aldi um die Corona-Selbsttests, die vom BAG kontrolliert und zugelassen sind. Damit ist Aldi der erste Detailhändler in der Schweiz, der die Corona-Tests verkauft.”

Headline	Aldi verkauft fünf Corona-Selbsttests in seinen Filialen	Aldi verkauft Corona-Selbsttests ab Donnerstag
Lead	Der Discounter Aldi verkauft ab Donnerstag fünf Corona-Selbsttests in seinen Filialen. Das Angebot gilt nur für ungeimpfte Personen.	Aldi verkauft die Corona-Selbsttests ab Donnerstag in seinen Filialen. Das Angebot gilt solange der Vorrat reicht.
Summary	• Aldi verkauft die Corona-Selbsttests in seinen Filialen. • Das Angebot gilt solange der Vorrat reicht. • Das Angebot gilt solange der Vorrat reicht.	• Ab Donnerstag verkauft Aldi die Corona-Selbsttests in seinen Filialen. • Das Angebot gilt solange der Vorrat reicht. • Damit ist Aldi der erste Detailhändler in der Schweiz, der die Corona-Selbsttests verkauft.
Caption	• Ab Donnerstag verkauft Aldi die Corona-Selbsttests in seinen Filialen. • Eine Packung kostet 19.99 Franken. Darin sind fünf Tests enthalten, pro Stück bezahlt man dann also noch rund vier Franken. • Das Angebot gilt solange der Vorrat reicht, heisst es weiter. • Dabei handelt es sich laut Aldi um die Corona-Selbsttests, die vom BAG kontrolliert und zugelassen sind. Damit ist Aldi der erste Detailhändler in der Schweiz, der die Corona-Tests verkauft. • Damit ist Aldi der erste Detailhändler in der Schweiz, der die Corona-Tests verkauft.	• Ab Donnerstag verkauft Aldi die Corona-Selbsttests in seinen Filialen. • Eine Packung kostet 19.99 Franken. Darin sind fünf Tests enthalten, pro Stück bezahlt man dann also noch rund vier Franken. • Das Angebot gilt solange der Vorrat reicht, heisst es weiter. • Dabei handelt es sich laut Aldi um die Corona-Selbsttests, die vom BAG kontrolliert und zugelassen sind. Damit ist Aldi der erste Detailhändler in der Schweiz, der die Corona-Selbsttests verkauft. Damit ist Aldi der erste Detailhändler in der Schweiz, der die Corona-Selbsttests verkauft. • Damit ist Aldi der erste Detailhändler in der Schweiz, der die Corona-Selbsttests verkauft. Damit ist Aldi der erste Detailhändler in der Schweiz, der die Corona-Selbsttests verkauft.
Topic	• Wirtschaft • Aldi • Corona-Test • Coronavirus	• Aldi • Coronavirus
RT	60	120

Table 6

Article: "In Deutschland wütet am Donnerstagmorgen das Sturmtief Ignatz. Es wird von umgestürzten Bäumen und Behinderungen im Bahnverkehr berichtet. Das Sturmfeld ist besonders in der Mitte des Landes aktiv. Vom Bundesland Hessen bis nach Sachsen im Osten gelten Unwetterwarnungen. Ignatz ist das Haupttief, dessen Ausläufer Hendrik zieht in der Nacht und am Donnerstagmorgen über die Schweiz. Die Orkanböen erreichen Windstärken von über 100 km/h. In Feldberg im Schwarzwald wurden gar Windgeschwindigkeiten von 166 km/h gemessen. Die Behörden warnen vor Aufhalten in Wäldern. Auch Brücken solle man unbedingt meiden. Grund: Bei starken Schauern und Gewittern könnten Höhenwinde auftreten, die Geschwindigkeiten von 120 km/h erreichen können. Ebenso wird geraten, Fenster und Türen zu schliessen und sich nicht im Freien aufzuhalten. Es bestehe die Gefahr von herumfliegenden Ästen und Dachziegeln. Der Sturm wird weiter Richtung Osten ziehen. Anschliessend wird die Front an der Ostsee erwartet. In der Millionenstadt Hamburg wird im Verlauf des Tages mit einer Sturmflut gerechnet. Auch das Auftreten von kurzlebigen Tornados sei laut dem Wetterdienst im Norden Deutschlands nicht ausgeschlossen."

Headline	Deutschland wütet mit Sturmtief Ignatz	Sturmtief Ignatz zieht über die Schweiz
Lead	In Deutschland wütet am Donnerstagmorgen das Sturmtief Ignatz. Die Behörden warnen vor Aufhalten in Wäldern.	In Deutschland wütet am Donnerstagmorgen das Sturmtief Ignatz. Es wird von umgestürzten Bäumen und Behinderungen im Bahnverkehr berichtet.
Summary	<ul style="list-style-type: none"> • In Deutschland wütet am Donnerstagmorgen das Sturmtief Ignatz. • Es wird von umgestürzten Bäumen und Behinderungen im Bahnverkehr berichtet. • Die Behörden warnen vor Aufhalten in Wäldern und Brücken. 	<ul style="list-style-type: none"> • In Deutschland wütet am Donnerstagmorgen das Sturmtief Ignatz. • Es wird von umgestürzten Bäumen und Behinderungen im Bahnverkehr berichtet. • Die Behörden warnen vor Aufhalten in Wäldern. • Grund: Bei starken Schauern und Gewittern könnten Höhenwinde auftreten, die Geschwindigkeiten von 120 km/h erreichen können.
Caption	<ul style="list-style-type: none"> • In Deutschland wütet am Donnerstagmorgen das Sturmtief Ignatz. • Es wird von umgestürzten Bäumen und Behinderungen im Bahnverkehr berichtet. • Das Sturmfeld ist besonders in der Mitte des Landes aktiv. Vom Bundesland Hessen bis nach Sachsen im Osten gelten Unwetterwarnungen. • Ignatz ist das Haupttief, dessen Ausläufer Hendrik zieht in der Nacht und am Donnerstagmorgen über die Schweiz. • Die Orkanböen erreichen Windstärken von über 100 km/h. • In Feldberg im Schwarzwald wurden gar Windgeschwindigkeiten von 166 km/h gemessen. • Die Behörden warnen vor Aufhalten in Wäldern. Auch Brücken solle man unbedingt meiden. Grund: Bei starken Schauern und Gewittern könnten Höhenwinde auftreten, die Geschwindigkeiten von 120 km/h erreichen können. Ebenso wird geraten, Fenster und Türen zu schliessen und sich nicht im Freien aufzuhalten. • Grund: Bei starken Schauern und Gewittern könnten Höhenwinde auftreten, die Geschwindigkeiten von 120 km/h erreichen können. 	<ul style="list-style-type: none"> • In Deutschland wütet am Donnerstagmorgen das Sturmtief Ignatz. • Es wird von umgestürzten Bäumen und Behinderungen im Bahnverkehr berichtet. • Die Behörden warnen vor Aufhalten in Wäldern. • Grund: Bei starken Schauern und Gewittern könnten Höhenwinde auftreten, die Geschwindigkeiten von 120 km/h erreichen können. Ebenso wird geraten, Fenster und Türen zu schliessen und sich nicht im Freien aufzuhalten. • Grund: Bei starken Schauern und Gewittern könnten Höhenwinde auftreten, die Geschwindigkeiten von 120 km/h erreichen können. Ebenso wird geraten, Fenster und Türen zu schliessen und sich nicht im Freien aufzuhalten. Es bestehe die Gefahr von herumfliegenden Ästen und Dachziegeln. • Der Sturm wird weiter Richtung Osten ziehen. Anschliessend wird die Front
Topic	• Deutschland • Unwetter	• Deutschland • Sturm
RT	60	120