Entropy-based Sampling for Abstractive Multi-document Summarization in Low-resource Settings

Laura Mascarell and **Ribin Chalumattu** and **Julien Heitmann**
ETH Zurich
{lmascarell,cribin,julien.heitmann}@inf.ethz.ch

Abstract

Research in Multi-document Summarization (MDS) mostly focuses on the English language and depends on large MDS datasets that are not available for other languages. Some of these approaches concatenate the source documents, resulting in overlong model inputs. Existing transformer architectures are unable to process such long inputs entirely, omitting documents in the summarization process. Other solutions address this issue by implementing multi-stage approaches that also require changes in the model architecture. In this paper, we introduce various sampling approaches based on information entropy that allow us to perform MDS in a single stage. These approaches also consider all source documents without using MDS training data nor changing the model's architecture. Besides, we build a MDS test set of German news articles to assess the performance of our methods on abstractive multi-document summaries. Experimental results show that our entropy-based approaches outperform previous state-of-the-art on German MDS, while still remaining primarily abstractive. We release our code[1] and MDS test set[2] to encourage further research in German abstractive MDS.

1 Introduction

In light of the ever-growing volume of available information, it becomes essential to be able to automatically summarize information from several sources. Multi-document Summarization (MDS) aims at condensing the most important information from different documents. Despite the advances in single-document summarization (Zhang et al., 2020), summarizing multiple related documents remains a greater challenge due to its input length and the presence of redundant information (Fan et al., 2019; Song et al., 2022). Therefore, some research focuses on implementing multi-stage approaches that first identify the relevant information to then feed it into a summarization model (Lebanoff et al., 2018; Liu and Lapata, 2019a). More recent works utilize pre-trained language models (Lewis et al., 2020; Raffel et al., 2020; Xiao et al., 2022) fine-tuned for the summarization task and feed them with the source documents concatenated (Johner et al., 2021; Xiao et al., 2022). However, these approaches pose two major issues. First, concatenated inputs exceeding the length limit of the model are truncated, which might lead to the omission of entire documents. Second, they rely on multi-document datasets that are scarce or unavailable in languages other than English.

Hokamp et al. (2020) introduce a decoding strategy that adapts single- to multi-document summarization without using additional training data nor applying changes to the single-input model architecture. At every decoding timestep, it averages the output probabilities of a single-document summarization model for each individual document, combining them into a single output. Instead of averaging all log-probabilities, which favours highly frequent tokens, we propose to make a more informed decision. In particular, we leverage entropy to measure the model confidence in the next token prediction and thus select the most informative output. We implement different entropy-based approaches and evaluate their performance on MDS of German text. Our main contributions are:

- We present different entropy-based sampling approaches for the MDS task. These are specially well-suited for languages like German that have limited or unavailable MDS data.

- We build and release a new German MDS test set in the news domain that is more suitable for evaluating abstractive summarization than the existing MDS German dataset auto-hMDS (Zopf, 2018). We expect our dataset to foster research on German abstractive MDS.

[1] Link to GitHub repository.
[2] Link to Multi-GeNews repository.

- The experimental results demonstrate that our method achieves the state-of-the-art performance in German abstractive MDS in terms of ROUGE scores and manual evaluation.

2 Related Work

Multi-document Summarization Some prior work approaches MDS as a multi-stage process (Liu et al., 2018; Zhu et al., 2021) that first extracts salient sentences from the source documents to then distill them into a summary using different methods such as graph-based modeling (Li et al., 2020; Chen et al., 2021) or modifying the attention mechanism (Perez-Beltrachini and Lapata, 2021). In contrast, Lebanoff et al. (2018) highlights the importance of adapting Single-document Summarization (SDS) models to summarize multiple documents and propose an approach that adapts their attention weights. Similarly, other works propose various changes in the model architecture (Liu and Lapata, 2019a; Elsahar et al., 2021). The main disadvantage of these approaches is that they are tailored to specific model architectures.

More recently, Xiao et al. (2022) introduce PRIMERA, a pre-trained model for MDS that can be applied in zero- or few-shot settings. The source documents are concatenated and fed into the Longformer Transformer model, which can handle long inputs up to 4,096 or even 16k tokens with current GPUs. Nevertheless, PRIMERA is only available for English and there is no alternative for other languages. Similarly, Johner et al. (2021) performs MDS on German text using the pre-trained language model BART (Lewis et al., 2020) and concatenating the source documents as input.[3] However, BART input length is restricted to 1,024 tokens, which may end up excluding entire documents from the summarization process.

Overall, our entropy-based approaches present the following advantages over prior work: (a) they do not require a pre-step to extract salient information (b) nor changes in the SDS model architecture with (c) no need for additional MDS training data, and (d) still considering all source documents in the summarization process. This work is built upon the dynamic ensemble approach from Hokamp et al. (2020), improving the decoding strategy by sampling on more informative predictions.

Entropy in Summarization Xu et al. (2020) leverage entropy to analyze the performance of Transformer-based models in the SDS task. Later, van der Poel et al. (2022) use entropy to determine when the model is uncertain about the next token prediction and apply Pointwise Mutual Information (PMI) instead to alleviate hallucination. Similarly, we apply the conditional PMI approach to MDS. Instead of finding the conditional entropy threshold through hyperparameter search as in van der Poel et al. (2022), we apply maximum probabilistic information entropy (Li et al., 2021). This novel entropy definition has been successfully used to reduce the size of image datasets by selecting the most informative samples.

3 Entropy Background

In information theory, the entropy of a random variable denotes the amount of information, or lack thereof (i.e. uncertainty), associated with its possible outcomes. Thus, given a probability distribution p over all possible outcomes $x_1, \dots, x_n$ of a random variable X, we quantify the entropy of X using the standard Shannon entropy equation:

$$H(X) = -\sum_{i=1}^{n} p(x_i)\ \log p(x_i) \qquad (1)$$

The entropy is then maximum for uniform distributions, where all outcomes are equally likely, indicating high uncertainty.

In the context of automatic text generation, we can leverage entropy to quantify the confidence of probabilistic models in their predictions (Xu et al., 2020; van der Poel et al., 2022). More specifically, summarization models aim at generating a summary string $\mathbf{y}^*$ of a given source document $\mathbf{x}$ that maximizes the scoring function:

$$\mathbf{y}^* = \underset{\mathbf{y} \in \mathcal{Y}}{\text{argmax}}\ \log p(\mathbf{y} \mid \mathbf{x}), \qquad (2)$$

where $\mathbf{y}$ is the sequence of tokens $y_0, \dots, y_T$ from the model vocabulary $\mathcal{V}$ generated at every timestep t, $0 < t < T$. During decoding, that is, the prediction of each sequence token $y_t \in \mathcal{V}$, the model provides a probability distribution $p(\cdot \mid \mathbf{y}_{<t}, \mathbf{x})$ over $\mathcal{V}$ that also takes into account the context of the previous tokens. According to Equation 1, we can then use such distribution to measure the model's confidence in the prediction:

[3] To the best of our knowledge, Johner et al. (2021) is the only work that tackles MDS in German besides Zopf (2018) with the auto-hMDS dataset.

$$H(p(\cdot \mid \mathbf{y}_{<t}, \mathbf{x})) = -\sum_{y \in \mathcal{V}} \Big(p(y \mid \mathbf{y}_{<t}, \mathbf{x}) \times \log p(y \mid \mathbf{y}_{<t}, \mathbf{x}) \Big) \tag{3}$$

4 Entropy-based MDS

Given a set of documents $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, the dynamic-ensemble approach (*DynE*) described in Hokamp et al. (2020) adapts single- to multi-document summarization as follows: at every decoding timestep t, it computes the output probabilites for each individual source document using a single-document summarization model; next, it averages these outputs to obtain a single log-probability distribution assigned to the token y:

$$p(y \mid \mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{\mathbf{x} \in \mathcal{X}} p(y \mid \mathbf{y}_{<t}, \mathbf{x}) \tag{4}$$

We leverage entropy information to adapt the *DynE* approach and implement various sampling strategies that select the most informative output at each decoding timestep t.

Minimum Entropy (H_{min}) Based on the hypothesis that low entropy indicates a higher confidence in the prediction, this approach picks the token prediction of the model instance with the lowest entropy $\min_{1 \leq i \leq |\mathcal{X}|} H(p(\cdot \mid \mathbf{y}_{<t}, \mathbf{x}_i))$. Note that this approach does not guarantee certainty in all token predictions. In those cases where all model instances exhibit high uncertainty, the selected instance could still have high entropy and thus provide an arbitrary prediction.

Max-predicted Probability Threshold (H_{th}) Li et al. (2021) focus on the maximum-predicted probability p_{max} to determine the model's confidence and reduce redundancy in datasets of images. Specifically, the authors state that a low maximum probability indicates high entropy and consequently, low confidence in the prediction. Therefore, they propose to measure entropy as:

$$H(X) = -p_{\max} \log p_{\max} \tag{5}$$

where $p_{\max} = \max_{x \in X} p(x)$. Figure 1 plots Equation 5, showing the correlation between information entropy and p_{max}. Note that the entropy is highest when p_{max} is 0.35, with a positive correlation for probabilities below this threshold and a negative correlation for probabilities above it.

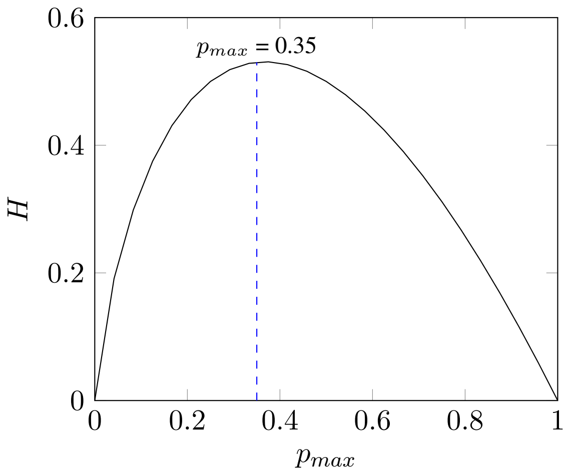


Figure 1: Plot of the maximum probabilistic information entropy (Equation 5), which illustrates the correlation between the maximum-predicted probability p_{max} and information entropy, when $0 \leq p_{max} \leq 1$.

Inspired by Li et al. (2021) approach, we apply maximum probabilistic information entropy in MDS, assuming that values of p_{max} below the threshold 0.35 indicate that the model is essentially guessing. At each decoding step, we obtain the maximum-predicted probability for each input document in $\mathcal{X}$ and proceed as follows:

a) we choose the prediction with the highest p_{max} among those above the threshold. The higher the probability, the lower the entropy.

b) if all probabilities are below the threshold, we conclude that there is not enough information for the current prediction and we average their log-probabilities as in Equation 4.

Mutual Information Decoding (H_{pmi}) Several works apply mutual information approaches during decoding to favor more specific and informative outputs (Li et al., 2016; Takayama and Arase, 2019). Later, van der Poel et al. (2022) observe that highly frequent tokens often indicate hallucinated content and implement a decoding strategy that integrates mutual information to mitigate hallucination in single-document summarization. In particular, their approach optimizes for Pointwise Mutual Information (PMI) when the model is uncertain about its prediction, generating summaries that are more faithful to the source document:

$$p(y|\mathbf{x}) = \log p(y|\mathbf{y}_{<t}, \mathbf{x}) - \lambda \log p(y|\mathbf{y}_{<t}), \tag{6}$$

where $0 < \lambda < 1$ to avoid excessively penalizing high-frequent tokens, which could lead to ungrammatical outputs (Li et al., 2016). Based on these findings, we propose an additional variation of our H_{th} approach, which applies PMI when there is no certainty in any of the predictions, that is, all probabilities are below the 0.35 threshold.[4]

5 Datasets

This section describes the datasets used to train and evaluate our MDS approaches. Specifically, we consider three pre-existing German datasets that are suitable for single-document—GeWiki (Frefel, 2020) and 20m (Rios et al., 2021)—and multi-document summarization—auto-hMDS (Zopf, 2018). Moreover, we build Multi-GeNews, a MDS test set in the news domain that is specifically tailored for abstractive MDS.

5.1 Single-document Summarization

GeWiki This is the largest dataset available for single-document abstractive summarization in German, consisting of 240k summary-article pairs. Here, the lead text of Wikipedia articles are extracted as summaries of the rest of the article.

20m A single-document summarization dataset with 18,305 news articles and their corresponding manually-written summaries collected from the Swiss newspaper *20 Minuten* ('20 Minutes').

5.2 Multi-document Summarization

auto-hMDS This multi-document summarization dataset consists of 2,210 summaries from Wikipedia leads as in GeWiki and 10,454 source documents. The documents were obtained by automatically querying the internet with summary sentences, resulting in a highly extractive dataset. Nonetheless, we consider it in our experiments for comparison with the related work. Despite being the largest MDS dataset in German, auto-hMDS is significantly smaller than its English counterpart Multi-News (Fabbri et al., 2019).[5]

Multi-GeNews Due to the lack of abstractive MDS datasets in German, we built a MDS test set to assess the performance of the proposed approaches. The data comes from the news portal of the Swiss media company SRF[6] and consists of news articles published between January and March 2020.

The articles published on the SRF website are often followed by a *Mehr zum Thema* ('More on the topic') section with related articles on the subject. To build our test set, we first utilize this section to obtain clusters of related articles. Specifically, we collect the related article suggestions and filter those published within one day of each other to ensure that they cover the same news. Next, we generate the reference summaries, which will be used to compute the automatic scores, concatenating the lead paragraphs of the articles in each cluster.[7] Hence, the reference summaries are a combination of lead texts. We finally filter salient sentences and remove duplicated information from the reference summaries using a pretrained extractive summarization model for German text. To build this model, we adapted the BertExt architecture (Liu and Lapata, 2019b) for the German language.[8] The adaption involved initializing the Bert component of the BertExt architecture using a German Bert checkpoint[9] and subsequently fine-tuning the entire model on the newswire 20m dataset.

The resulting dataset consists of 754 unique articles grouped into 402 clusters. Each cluster contains two to six articles with a median of four articles and the corresponding generated reference summary.[10] The average length of the articles and summaries are 593 and 61 tokens, respectively.

6 Experiments

We evaluate the performance of the entropy-based sampling approaches on our Multi-GeNews and the auto-hMDS datasets in terms of automatic ROUGE scores (Lin, 2004) and extractive fragment density ρ (Grusky et al., 2018). Since we focus on abstractive summarization, the latter allows us to measure the degree of extractiveness of the summaries, and in turn, abstractiveness—higher ρ values indicate that the summary is more extractive and contains larger text chunks from the source article. Furthermore, we collect human annotations on a subset of the Multi-GeNews to assess the faithfulness of the generated summaries and get a deeper understanding on their quality (Section 7).

[4] We use a λ of 0.25 in our experiments, which we manually selected based on the impact of various values on the output.

[5] over 56k summaries and 250k source documents.

[6] https://www.srf.ch/news

[7] Similarly to the GeWiki dataset, we consider the lead paragraph of an article as its summary.

[8] https://github.com/nlpyang/BertSum

[9] https://huggingface.co/dbmdz/bert-base-german-uncased

[10] Although an article can belong to different clusters, there are no identical clusters with the same articles.

	100 words				200 words			
Method	**R1**↑	**R2**↑	**RL**↑	ρ↓	**R1**↑	**R2**↑	**RL**↑	ρ↓
mBART concat	18.4	6.2	12.5	27.9	24.5	7.7	15.0	35.6
mBART + *DynE*	**23.4**	6.9	15.1	2.2	26.8	7.0	16.0	1.9
mBART + H_{min}	20.7	8.6	14.7	17.9	26.9	10.4	17.4	16.6
mBART + H_{th}	21.5	**9.0**	**15.3**	16.1	**27.8**	**10.8**	**18.0**	14.7
mBART + H_{pmi}	16.5	6.9	12.3	12.5	21.0	7.9	14.5	10.0

Table 1: Performance of the entropy-based approaches and the baseline models on the **auto-hMDS** dataset in terms of ROUGE scores and extractive fragment density ρ. The mBART model is fine-tuned on the auto-hMDS dataset by concatenating the source articles into a single input. Similarly, the mBART baseline is fed with the concatenated source articles. Overall, H_{th} achives the highest performance among the various methods evaluated.

	100 words			
Method	**R1**↑	**R2**↑	**RL**↑	ρ↓
mBART concat	23.0	6.0	14.8	9.23
mBART + *DynE*	22.2	4.8	14.9	1.5
mBART + H_{min}	23.4	5.6	15.0	2.46
mBART + H_{th}	**24.5**	6.2	15.6	2.72
mBART + H_{pmi}	23.9	**7.2**	**16.1**	2.78

Table 2: Performance of the entropy-based approaches and the baselines on our **Multi-GeNews** test set. The mBART model is fine-tuned on the 20m dataset as described in Section 6.1. The mBART baseline receives as input the source articles concatenated.

6.1 Models

This section describes the implementation details to build the models used in our experiments. Namely, the two summarization models, individually fine-tuned on the newswire 20m and the auto-hMDS datasets, and the language model used by the pointwise mutual information decoding approach H_{pmi}.

Summarization Models We evaluate the performance of our MDS approaches using two summarization models fine-tuned on the news domain dataset 20m[11] and the MDS dataset auto-hMDS, respectively. The latter allows us to compare the performance of our approaches against prior work on German MDS. The models are based on mBART, a multilingual sequence-to-sequence transformer-based model that effectively handles multiple languages including German (Liu et al., 2020) and initialized with the `facebook/mbart-large-cc25` checkpoint available at the Hugging Face Hub.[12]

In particular, we fine-tune the model on the 20m dataset for 10 epochs and batch size of 2 using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of $3e-5$. The gradient accumulation steps is 16, resulting in a total effective batch size of 32. To fine-tune the single-input mBART model with the multi-document summarization dataset auto-hMDS, we follow the work in Johner et al. (2021) and concatenate the source articles in a single input. We train the model with $3e-5$ learning rate and batch size of 2 for 5 epochs.

Language Model We build a language model to apply the mutual information decoding approach. Specifically, we use the GPT-2 (Radford et al., 2019) checkpoint from Hugging Face[13] and fine-tune it on the same in-domain data as the corresponding mBART summarization model. To ensure that both mBART and GPT-2 models share the same vocabulary, we train GPT-2 using the same tokenizer as mBART. We then fine-tune it for 3 epochs using the AdamW optimizer with learning rate of $5e-4$ and batch size of 16. We set the maximum context length of the model to 256 tokens, since we do not generate longer summaries than that. The gradient accumulation steps is set to 8, resulting in a total effective batch size of 128.

6.2 Results

Table 2 compares the performance of our entropy-based methods against the *DynE* (Hokamp et al., 2020) and mBART baselines on our Multi-GeWiki

[11] Since the GeWiki dataset is significantly larger than the in-domain 20m, we also considered to build a model using both datasets through behavioral fine-tuning. However, the performance on the single-document summarization task was inferior than simply fine-tuning on 20m. Several factors could contribute to this results such as a domain shift or a discrepancy in summary length distribution.

[12] `https://huggingface.co/facebook/mbart-large-cc25`

[13] `https://huggingface.co/gpt2`

	Ref.	Das Coronavirus beherrscht die Medien – doch das sei derzeit kaum angebracht, sagt Wissenschaftsredaktor Häusler. Innerhalb eines Tages kletterte die Zahl der bestätigten Infektionen in China auf 2744. In den umliegenden Ländern gibt es bereits Erkrankte. [The coronavirus dominates the media - but that is hardly appropriate at the moment, says science editor Häusler. Within one day, the number of confirmed infections in China climbed to 2744, and there are already people ill in surrounding countries.]
Rank 1	H_{th}	Die Zahl der Corona-Toten in China ist auf 80 gestiegen. Die Zahl der Touristen, die sich mit dem Virus infiziert haben, ist auf überschaubar. Die Tourismusbranche rechnet nicht damit, dass das Virus sich auf der ganzen Welt ausbreitet. SRF-Wissenschaftler Thomas Häusler und Daniel Sulzer von Interlaken erklären, wie gefährlich die Situation für die chinesische Bevölkerung ist.
		[The number of Covid deaths in China has risen to 80. The number of tourists infected with the virus is on manageable. The tourism industry does not expect the virus to spread around the world. SRF scientist Thomas Häusler and Daniel Sulzer from Interlaken explain how dangerous the situation is for the Chinese population.]
Rank 2	H_{pmi}	In China ist die Zahl der Corona-Fälle stark angestiegen. In der Volksrepublik gibt es bereits über 700 Fälle. Die Zahl der Touristen, die sich mit dem Virus infiziert haben, ist doppelt so hoch wie bei der saisonalen Grippe. Ein Tourismusexperte erklärt, wie gefährlich die Situation für die chinesische Bevölkerung ist.
		[In China, the number of Covid cases has risen sharply. There are already over 700 cases in the People's Republic. The number of tourists infected with the virus is twice as high as the seasonal flu. A tourism expert explains how dangerous the situation is for the Chinese population.]
Rank 3	*DynE*	Die Zahl der Corona-Fälle in der Schweiz steigt, die Zahl der Infizierten nimmt zu. Die Gefahr, sich mit dem Virus anzustecken, ist noch nicht bedroht. Trotzdem steigen die Zahlen wieder an. Laut Experten ist die Gefahr grösser, dass sich das Virus noch weiter ausbreitet.
		[The number of Covid cases in Switzerland is rising, the number of infected persons is increasing. The danger of contracting the virus is not yet threatened. Nevertheless, the numbers are rising again. According to experts, there is a greater risk that the virus will spread even further.]
Rank 4	H_{min}	Der Tourismusverband Chinas rechnet mit einem Rückgang der Touristenzahlen. In Interlaken und Luzern gibt es nur noch wenige Berichte über das Coronavirus. In der Schweiz gibt es aber Hoffnung: Vermehrt Japaner und Chinesen berichten von Infektionen mit dem Virus. Ein Tourismusdirektor glaubt, dass der Tourismus in der Volksrepublik eine globale Pandemie auslösen könnte.
		[The Tourism Association of China expects a decline in tourist numbers. In Interlaken and Lucerne there are only few reports of the coronavirus. In Switzerland, however, there is hope: Increasing numbers of Japanese and Chinese report infections with the virus. One tourism director believes that tourism in the People's Republic could trigger a global pandemic.]

Table 3: Example of the summary ranking task for the input articles 18126230, 18127577, and 18130289, where at least two annotators agreed on the ranking position for each summary. In contrast to the entropy-based approaches, *DynE* is susceptible to generate overly general summaries.

test set. We use the single-document summarization mBART model fine-tuned on the in-domain dataset 20min—see details in Section 6.1. To feed the mBART baseline with multiple documents, we concatenate them in a single input as in Johner et al. (2021). Overall, the automatic ROUGE scores indicate that H_{th} and H_{pmi} achieve the highest performance. Similarly, H_{th} outperforms the other approaches on the auto-hMDS dataset (see Table 1).

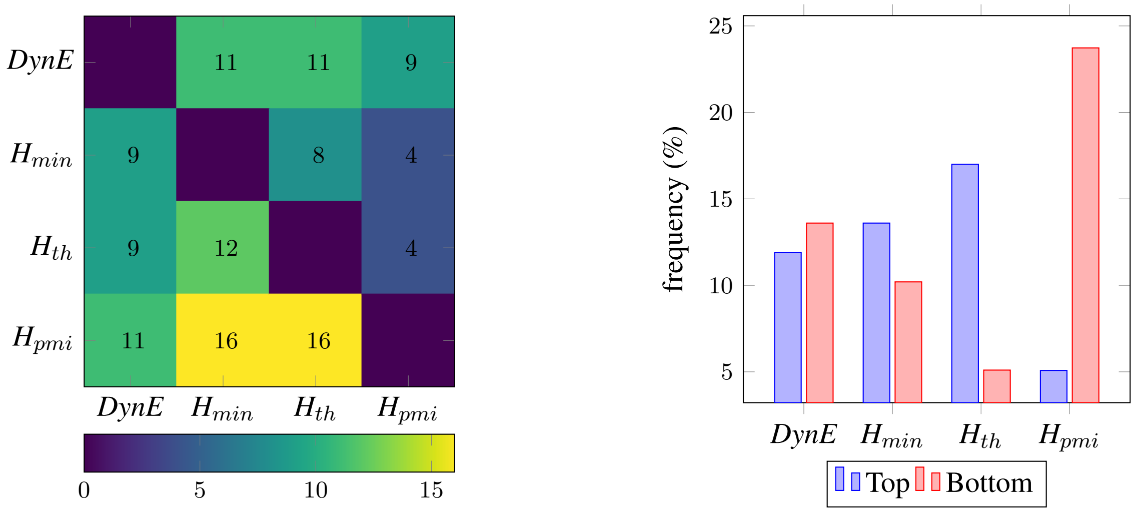
Abstractiveness of the Summaries Table 1 and Table 2 reveal that *DynE* summaries are the most abstractive (lowest ρ scores). In contrast, concatenating the source articles as input results in highly extractive summaries,[14] and the gap is even more significant with mBART fine-tuned on auto-hMDS, since the dataset is highly extractive (Table 1). Although we aim at generating abstractive summaries, the *DynE* approach is prone to generate highly frequent tokens,[15] resulting in general summaries that fail to consider relevant and specific information from the source articles (see example in Table 3). Instead, our entropy-based approaches generate summaries with a moderate level of abstractiveness that also include concrete information.

7 Human Evaluation

We recruited three native German speakers to perform a manual evaluation on the Multi-GeNews test

[14] The results on the extractiveness of mBART summaries are also supported in Johner et al. (2021).

[15] Since the *DynE* approach averages the log-probability outputs at each decoding step, common tokens obtain higher probabilities and are more likely to be predicted.



(a) Heatmap illustrating the distribution of relative preference among the approaches. The x-axis indicates the preferred approach over the y-axis. H_{min} and H_{th} are the most favoured approaches, whereas H_{pmi} ranks as the least preferred.

(b) Percentage of instances where each approach ranked at the top and the bottom positions, according to the annotators. While H_{th} summaries were consistently rated among the top positions, the annotators rated H_{pmi} summaries low.

Figure 2: Evaluation of the quality of the summaries among the different approaches. We only consider those instances where the majority of the annotators agreed on the (a) relative or (b) absolute ranking position.

set.[16] This evaluation task is twofold: (1) assess the relative quality of the summaries, ranking them accordingly (Goyal et al., 2022; Ermakova et al., 2019) and (2) the faithfulness of the generated summaries to the source articles (Krishna et al., 2023), that is, whether the information presented in the summaries is supported by the articles.

Since the task requires to read a considerable amount of text, we ask the participants to annotate a sample of the MDS test set. This sample comprises 20 randomly selected instances that meet the following criteria: (a) each instance consists of three source articles, (b) the generated summaries end with a punctuation mark to avoid incomplete sentences, and (c) the token-level edit distance among the summaries is above five to ensure lexical differences. For each participant, we randomly shuffle the evaluation instances and the required annotations to avoid any biases. Additionally, we do not provide them with any information about which specific approach generated each summary.

Summary Ranking Task The objective of this task is to gain insights into human preferences of the generated summaries. For each instance (i.e. a set of related articles), we ask the participants to rank the generated summaries according to the informativeness of the summaries and their preference. That is, they must evaluate how effectively the summaries are at capturing the essential information from the three source articles. Note that we are not evaluating other linguistic aspects such as cohesion or fluency, since we do not implement any specific methods to improve those. Table 3 provides an example of the annotation task.

This task is specially challenging when multiple summaries either contain similar information or suffer from hallucinations, or both. In fact, the annotators reported that it was hard to decide the rank of a summary between two consecutive positions. This negatively impacted on the inter-annotator agreement, resulting in a final Kendall's tau coefficient of 0.22. In the analysis of this task, we concentrate on the relative performance of the approaches and only consider instances with a majority agreement among annotators.

Figure 2a illustrates the relative preference among the different approaches in this ranking task. The results demonstrate that the summaries from the approaches H_{min} and H_{th} are consistently rated higher than the others. In contrast, the H_{pmi} summaries are the least preferred. These results are also supported in Figure 2b, where we compare the frequency with which each approach was ranked within the top two and the bottom two positions. Furthermore, Figure 2b shows that while the base-

[16] The participants received a voucher worth CHF 75.- as compensation for their participation.

	H_{th}
F	**Donald Trump hielt sich in** der Nacht auf Mittwoch in den beiden Kammern des US-Kongresses seine dritte Rede ab. [Donald Trump delivered his third speech to both chambers of the U.S. Congress on Wednesday night.]
F	Die Rede ist von einem Triumphgehabe gegen die Demokraten. [There is talk of triumphant action against the Democrats.]
F	Das Verfahren gegen Trump ist nach wie vor im Gange. [The case against Trump is still ongoing.]
	H_{pmi}
F	**Donald Trump hielt sich in** den USA nicht an die Corona-Regeln. [Donald Trump did not follow the Covid rules in the USA.]
F	Die demokratische Mehrheit im Kongress hielt sich dagegen und sprach Trump ab. [The Democratic majority in Congress held against this and absolved Trump.]
T	Die Rede ist von einem Triumph für Trump. [The talk is of a triumph for Trump.]

Table 4: Example of the faithfulness annotation task. The boolean in the first column represents whether the text span is factual (T=True) or not (F=False) based on the majority agreement among the annotators. Here, the H_{th} summary was ranked at the top positions of the ranking and H_{pmi} at the bottom positions, even though the latter has a text span annotated as factual. The highlighted text indicate the common tokens between H_{th} and H_{pmi} until H_{pmi} applies PMI, hallucinating on the Covid virus, although Covid is not even mentioned in the source articles.

line summaries *DynE* receive mixed ratings, the H_{th} summaries are consistently ranked in the top positions. This indicates a consistent preference for the H_{th} summaries over the baseline.

Faithfulness Annotation Task van der Poel et al. (2022) leverage PMI to improve faithfulness and evaluate it in terms of automatic metrics (Section 4). The goal of this annoation task is to manually evaluate the faithfulness of our H_{pmi} approach, which applies PMI to MDS, and compare it to the other proposed approaches that do not specifically address faithfulness. Specifically, we follow the guidelines described in Krishna et al. (2023) and split the summaries into text spans to ensure lower inter-annotator variance.[17] We then ask the annotators to judge whether each span is faithful to the source articles, that is, the statements can be verified against the articles. The final Fleiss'κ (Fleiss, 1971) inter-annotator agreement is 0.62.

Overall, the annotations indicate that hallucination is a general issue in all generated summaries. To evaluate the impact of H_{pmi} on hallucination, we only consider those annotations where at least two annotators agree on the factuality label. The results show that H_{min} and H_{th} obtain a factuality rate of 36% and 33.3%, respectively, while H_{pmi} achieves a slightly higher factuality rate of 36.2%. Given the small size of the evaluation sample, we conclude that there is no significant improvement of factuality with the H_{pmi} approach on this task.

Since H_{pmi} is an enhanced version of H_{th}, and H_{th} is consistently preferred over H_{pmi} (Figure 2), we delve deeper into cases where H_{pmi} shows an improvement in factuality, yet it receives a lower rating than H_{th}. The results indicate that H_{pmi} indeed redirects the prediction of the rest of the summary, specially when applied early on as stated in Li et al. (2016). However, it does not necessarily address the issue of hallucination. For example, the first text span of H_{th} in Table 4 hallucinates the moment when the speech occurs *Nacht auf Mittwoch* ('Wednesday night'), and it is therefore annotated as not factual. In contrast, the H_{pmi} generates a sentence about the Covid rules. However, none of the source articles refer to this topic,[18] which results in a more severe hallucination.

8 Conclusion

In this work, we tackle Multi-document Summarization (MDS) in low-resource settings where there is a lack of MDS training data. We therefore present various sampling approaches built upon prior works that use single-document summarization models for the MDS task. Specifically, we leverage information entropy as a metric to measure the model certainty in each token prediction. The experimental results on German MDS show that

[17] Although the guidelines mainly refer to long summaries of at least 150 words, we found them also useful in our setting.

[18] Source articles ids: 18163721, 18160037, and 18160205.

our H_{th} approach, which specifically applies maximum probabilitic information entropy, achieves the state-of-art in German abstractive MDS. In our experiments, we also assessed an extended version of the H_{th} approach that applies Pointwise Mutual Information (PMI) when all predictions exhibit uncertainty. Although PMI has been used in prior work to address hallucination, we observe in the manual evaluation that PMI changes the prediction of the rest of summary, but it does not inherently tackle hallucination. Future work should focus on addressing the issue of hallucination in automatic summarization, including further research on the efficacy of PMI to mitigate hallucinations. Additionally, it would be interesting to explore alternative approaches to enhance the H_{th} approach when there is uncertainty in the prediction. Finally, we built a MDS test set of German news articles that will help the research community to evaluate abstractive MDS on German text.

Ethics Statement

Human Annotation We recruited the annotators for the manual evaluation task on a voluntary basis and provided them with information about the goals and scope of the task. The data was collected anonymously such that no conclusion can be drawn about any particular annotator. This human evaluation obtained the corresponding ethical approval from the Ethics Commission of ETH Zurich university (EK-2023-N-37).

Text Generation Models Ethical considerations documented for natural language generation systems (Smiley et al., 2017; Kreps et al., 2022) also apply to our work. We do not anticipate any additional concerns.

Supplementary Materials Availability Statement: Source code for the presented entropy-based sampling approaches[19] in Section 4 and the Multi-GeNews dataset[20] described in Section 5.2 are available from GitHub.

Acknowledgements

This project is supported by Ringier, TX Group, NZZ, SRG, VSM, viscom, and the ETH Zurich Foundation.

[19] Link to GitHub repository.

[20] Link to Multi-GeNews repository.

References

Moye Chen, Wei Li, Jiachen Liu, Xinyan Xiao, Hua Wu, and Haifeng Wang. 2021. SgSum:transforming multi-document summarization into sub-graph selection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4063–4074, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hady Elsahar, Maximin Coavoux, Jos Rozen, and Matthias Gallé. 2021. Self-supervised and controlled multi-document opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1646–1662, Online. Association for Computational Linguistics.

Liana Ermakova, Jean Valère Cossu, and Josiane Mothe. 2019. A survey on evaluation of summarization methods. *Information processing & management*, 56(5):1794–1814.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2019. Using local knowledge graph construction to scale Seq2Seq models to multi-document inputs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4186–4196, Hong Kong, China. Association for Computational Linguistics.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Dominik Frefel. 2020. Summarization corpora of Wikipedia articles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6651–6655, Marseille, France. European Language Resources Association.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. 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)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Chris Hokamp, Demian Gholipour Ghalandari, Nghia The Pham, and John Glover. 2020. DynE: Dynamic ensemble decoding for multi-document summarization. *arXiv preprint arXiv:2006.08748*.

Timo Johner, Abhik Jana, and Chris Biemann. 2021. Error analysis of using BART for multi-document summarization: A study for English and German language. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 391–397, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Sarah Kreps, R. Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, 9(1):104–117.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4131–4141, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. Leveraging graph to improve abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6232–6243, Online. Association for Computational Linguistics.

Yang Li, Jiachen Yang, and Jiabao Wen. 2021. Entropy-based redundancy analysis and information screening. *Digital Communications and Networks*.

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.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Yang Liu and Mirella Lapata. 2019a. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019b. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Multi-document summarization with determinantal point process attention. *Journal of Artificial Intelligence Research*, 71:371–399.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, Online and in Dominican Republic. Association for Computational Linguistics.

Charese Smiley, Frank Schilder, Vassilis Plachouras, and Jochen L. Leidner. 2017. Say the right thing right: Ethics issues in natural language generation systems. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 103–108, Valencia, Spain. Association for Computational Linguistics.

Yun-Zhu Song, Yi-Syuan Chen, and Hong-Han Shuai. 2022. Improving multi-document summarization through referenced flexible extraction with credit-awareness. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1667–1681, Seattle, United States. Association for Computational Linguistics.

Junya Takayama and Yuki Arase. 2019. Relevant and informative response generation using pointwise mutual information. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 133–138, Florence, Italy. Association for Computational Linguistics.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding neural abstractive summarization models via uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6275–6281, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11328–11339. Proceedings of Machine Learning Research (PMLR).

Fangwei Zhu, Shangqing Tu, Jiaxin Shi, Juanzi Li, Lei Hou, and Tong Cui. 2021. TWAG: A topic-guided Wikipedia abstract generator. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4623–4635, Online. Association for Computational Linguistics.

Markus Zopf. 2018. Auto-hMDS: Automatic construction of a large heterogeneous multilingual multi-document summarization corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).