MIRANEWS: Dataset and Benchmarks for Multi-Resource-Assisted News Summarization

Xinnuo Xu[1], **Ondřej Dušek**[2], **Shashi Narayan**[3], **Verena Rieser**[1] **and Ioannis Konstas**[1]
[1]The Interaction Lab, MACS, Heriot-Watt University, Edinburgh, UK
[2]Charles University, Faculty of Mathematics and Physics, Prague, Czechia
[3]Google Research
{xx6, v.t.rieser, i.konstas}@hw.ac.uk
odusek@ufal.mff.cuni.cz
shashinarayan@google.com

Abstract

One of the most challenging aspects of current single-document news summarization is that the summary often contains 'extrinsic hallucinations', i.e., facts that are not present in the source document, which are often derived via world knowledge. This causes summarization systems to act more like open-ended language models tending to hallucinate facts that are erroneous. In this paper, we mitigate this problem with the help of multiple supplementary resource documents assisting the task. We present a new dataset MIRANEWS and benchmark existing summarization models.[1] In contrast to multi-document summarization, which addresses multiple events from several source documents, we still aim at generating a summary for a single document. We show via data analysis that it's not only the models which are to blame: more than 27% of facts mentioned in the gold summaries of MIRANEWS are better grounded on assisting documents than in the main source articles. An error analysis of generated summaries from pretrained models fine-tuned on MIRANEWS reveals that this has an even bigger effects on models: assisted summarization reduces 55% of hallucinations when compared to single-document summarization models trained on the main article only.

Summary: Kathy Griffin and Howard Stern gather to say goodbye at Joan Rivers funeral in manhattan New York AP. Even in death, Joan Rivers got what she wanted: a star-studded funeral, with the worlds of Hollywood, fashion, media and money all among the mourners on Sunday morning.

Document: on sunday morning, a legion of notables turned out at new york's temple emanu-el to remember rivers, who died thursday at 81: kathy griffin, whose edgy, biting comedy career was largely made possible by rivers; colleague and friend kelly osbourne; sarah jessica parker and whoopi goldberg; howard stern... lined up outside the fifth avenue synagogue and waited for their names to be checked against a list before entering. barricades lined several blocks of manhattan's fifth avenue, and a crowd of fans and media stood watch across the street. the comedian detailed in her 2012 book `` i hate everyone starting with me'' that she hoped for ``a huge showbiz affair with lights, cameras, action "and ``hollywood all the way. "...

Assisting Document: new york city's temple emanu-el; joan rivers all access photo/splash news online updated 09/07/2014 at 12:15 pm edt originally published 09/07/2014 at 11:45 am edt. it was exactly the kind of star-studded send-off she wanted and deserved as crowds of fans packed the sidewalks outside of temple emanu-el on new york's upper east side sunday morning, intimates, relatives and celebrity friends of began to trickle in to the private funeral to pay their last respects to the comedian, who and husband matthew broderick were among the early arrivals...

Figure 1: An example where the summary (top section) contains information that is not explicitly included in its main document (middle section), but is covered in the related assisting document (bottom section). We highlight the information in the summary that is aligned to its corresponding main and assisting documents with yellow and pink colors, respectively.

1 Introduction

The vast majority of current research on abstractive summarization is aimed at single-document news summarization due to the widespread availability of data, e.g. (NY Times; Sandhaus (2008), CNN/DailyMail; Hermann et al. (2015), Newsroom; Grusky et al. (2018), XSum; Narayan et al. (2018a), MLSUM; Scialom et al. 2020). The datasets are curated by pairing a single document with human authored highlights/description as the summary. This task is typically approached using conditional generation models, including sequence-to-sequence architectures with attention and copy mechanisms (See et al., 2017), Transformers (Liu and Lapata, 2019a), and pre-trained language modeling (e.g. Radford et al., 2019; Lewis et al., 2020).

While these SotA summarization models reach a high level of fluency and coherence, they are also highly prone to hallucinating content that is not grounded by the input document. Maynez et al. (2020) classified hallucinations into ***intrinsic*** that mistakenly manipulate information from the source document resulting in *counterfactual* output, and ***extrinsic*** that introduce information not grounded in the document (see Figure 1). Extrinsic halluci-

[1]Our code and data are available at: https://github.com/XinnuoXu/MiRANews

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1541–1552
November 7–11, 2021.

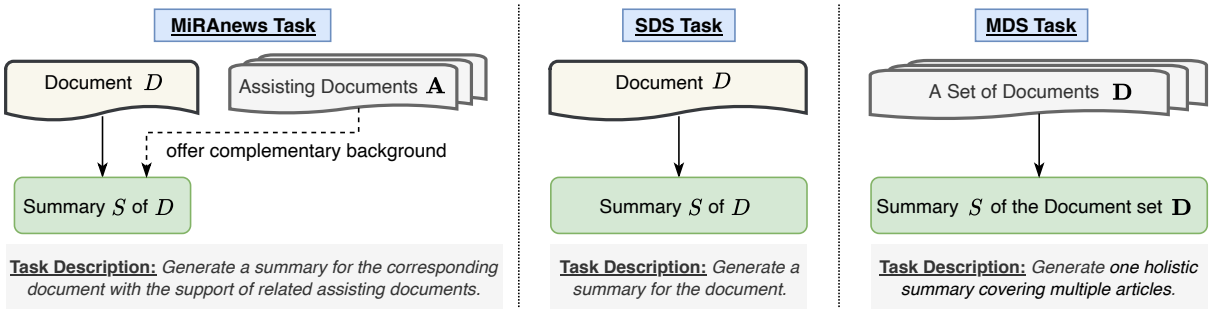

Figure 2: Comparison of Summarization tasks. Single-document Summarization (SDS task) focuses on generating summary S based on a single document D. Multi-document Summarization (MDS task) creates a holistic summary S covering multiple articles $\mathbf{D}$. The MIRANEWS task differs by producing summary S based only on the events pertinent in the *main* article D, while reaching to a set of assisting documents $\mathbf{A}$ for complementary background.

nations are further broken down into 'factual', i.e., holding true in real life, and 'counterfactual'.

Similar to (Maynez et al., 2020), we find not only the models are to blame, but also the datasets: human-written summaries contain up to 36% external facts which are not *faithful*, i.e., covered by the single input document. In other words: the summaries also contain 'extrinsic hallucinations'. Moreover, facts which are present are often re-phrased or shortened in the summary in ways which requires world knowledge. Consider the example in Figure 1, where the surname "*Rivers*" used throughout the document (middle section), is elaborated as the full name "*Joan Rivers*" in the summary (top section), i.e. adding information. Meanwhile, "*celebrities lined up outside the fifth avenue synagogue*" in the document is specified as "*say goodbye at Joan Rivers funeral*" in the summary, which requires world knowledge. Moreover, the fact about an "a star-studded funeral" is not mentioned explicitly in the document. Any summarization model that is agnostic to such data divergence issues between the source and target texts (Dhingra et al., 2019) will function more as an open-ended language model and will be prone to extrinsic hallucinations.

In this work, we tackle the problem of *extrinsic hallucinations* by introducing a new task, Multi-Resource-Assisted News Summarization and a novel dataset (MIRANEWS). Following Maynez et al. (2020), we regard the incorporation of background knowledge within a generated summary as the desired property. However, instead of sourcing this knowledge via pretraining on large datasets,[2] we base our work on the assumption that articles from alternative news resources covering the same news event can complement the background knowledge in an easier to learn, more direct, and explainable way. Consider the example in Figure 1, where the assisting document (bottom section) from another news resource recounts some facts in the summary (highlighted in pink) in a more explicit way.

Note that, as shown in Figure 2 (left), **our task is different from both Single-document Summarization** (SDS, middle) **and Multi-document Summarization** (MDS, right): SDS aims at generating a summary for a single main document, while we aim to generate a target summary S for a single document D with supporting facts from multiple assisting documents $\mathbf{A}$. In this paper:

- We introduce a new task, Multi-Resource-Assisted News summarization, aiming at generating a summary for the corresponding news article with the support of related assisting documents.
- We create and release a new dataset (MIRANEWS) introducing a novel automatic data collection method which gathers multiple assisting news articles from different news resources for a document-summary pair.
- We introduce new *referenceless* metrics, which quantitatively evaluate extrinsic hallucinations both in summarization datasets and output summaries, and confirm that introducing assisting documents offers better grounding to more than 27% of facts mentioned in the reference summaries.
- We report benchmark results using models both fine-tuned and trained from scratch on MIRANEWS. We show that modeling assisting documents effectively introduces external facts in the summaries that are grounded on the assisting docu-

[2] Although they report BERTS2S (Rothe et al., 2020) to output more factual hallucinations in the summary than their non-pre-trained counterparts on XSum (Narayan et al., 2018a), still over 90% of the total hallucinations are incorrect.

ments, resulting in 55% less counterfactual hallucinations than SDS systems.

2 Data Collection

Data Resource. Following Fabbri et al. (2019)'s MDS efforts, we use the news aggregation portal newser.com to collect news articles with their assisting documents, where each webpage reports on a news event and includes editor-picked links to the relevant news articles from other news websites. An example is in Figure 3: three news articles (D_2, D_3, D_4) from *nytimes*, *newser*, and *CNBC* are linked to the webpage (D_1), all of which report on the same event of starship prototype landing.

News Cluster and Content Extraction. We consider each article on newser.com, together with the pages cited therein, as a cluster about one news event. We extract the document and the corresponding summary from each webpage automatically, following the method introduced in NEWSROOM (Grusky et al., 2018).[3] Specifically, the documents are constructed from the HTML main text body excluding HTML markups, inline advertising, images/videos, and captions, while the target summary S is extracted from the document's metadata fields, e.g. *og:description*, *twitter:description*, *description*, which are often written by editors and journalists to appear on social media services or as search engine webpage descriptions. Hence, for each cluster $\mathbf{C}$, we collect paired documents and summaries $\mathbf{C} = \{(D_1, S_1), (D_2, S_2) \cdots (D_m, S_m)\}$, where m is the number of webpages in the cluster.

Collecting Assisting Documents. We first represent all documents in the news cluster $\mathbf{C}$ as $\mathbf{D} = \{D_1, D_2 \cdots D_m\}$. In turn, we take $\mathbf{A}_i = \mathbf{D} - D_i$ as the assisting documents for each document D_i and its summary S_i in the cluster. Thus, for a cluster including m corresponding webpages, we create m examples. Each of them contains one document, its summary, and $m-1$ assisting documents, denoted as $(D_i, S_i, \mathbf{A}_i)$.

Accordingly, we create the full MIRANEWS dataset $\mathcal{D} = \{(D_i, S_i, \mathbf{A}_i)\}_{i=1}^{M}$ by collecting examples from all available 57K newser.com pages following Fabbri et al. (2019). Note that, before creating the clusters, we first randomly split the webpages into training (80%), validation (10%), and test (10%) set, and then generate examples within each set in order to prevent data leaking, i.e. each document is only included in one of the sections (regardless of main/assisting role).

[3] We use the data scraping and data extraction code from https://github.com/lil-lab/newsroom.

3 Data Analysis

MIRANEWS contains 150K examples in total, with an average of 1.7 assisting documents per instance[4] Table 1 compares MIRANEWS with popular large scale summarization datasets. MIRANEWS is similar in size to CNN; document and summary average lengths in MIRANEWS are similar to CNN, DailyMail (Hermann et al., 2015), NY Times (Sandhaus, 2008), and Newsroom (Grusky et al., 2018), but longer than XSum (Narayan et al., 2018a).

3.1 Bias towards Extractive Methods

N-gram novelty. We evaluate the dataset bias towards extractive methods using n-gram novelty introduced in (Narayan et al., 2018a). This metric reports the percentage of novel n-grams in the gold summaries that do not appear in their source documents. Lower values indicate that more n-grams of the summaries appear in the documents, i.e. there is more overlapping information that supports the summary, leading to more extractive summaries.

The left section in Table 2 shows the results in comparison with other commonly used datasets. MIRANEWS(S-D), i.e. the percentage of novel n-grams in the summaries S that do not appear in their main document D, is lower than in other benchmarks. This means that MIRANEWS, when treated as a SDS task, will benefit extractive methods. On the other hand, MIRANEWS(S-A), i.e. the n-grams novelty of the summaries with respect to their assisting documents $\mathbf{A}$, is much higher, comparable with XSum. this shows that assisting documents in MIRANEWS are not redundant to the main documents. The level comparable to XSum suggests that they indeed describe the same news event, i.e., are relevant to the summaries.

LEAD and EO. We further evaluate two well established extractive methods on MIRANEWS and other benchmarks. LEAD is often used as a strong lower bound for summarization (Nenkova, 2005) and creates a summary by selecting the first few sentences or words in the document. For

[4] The minimum and maximum number of assisting documents in each example is 1 and 4. We keep the four assisting documents at most for each example.

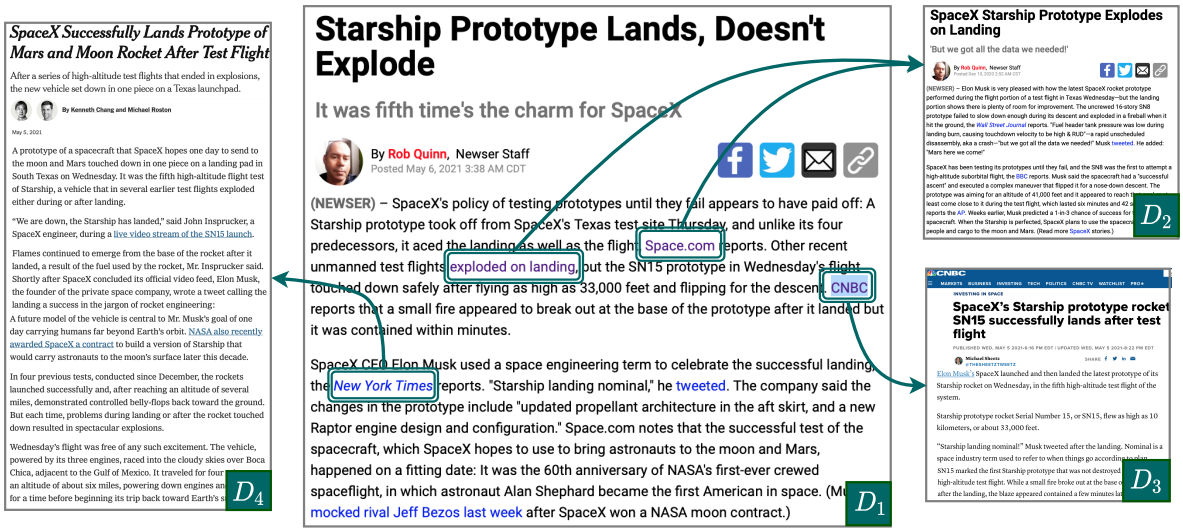

Figure 3: Example of a page on *newser.com*: a `newser.com` article is a news event including editor-picked links to relevant news articles from other news websites. This example shows the webpage `https://www.newser.com/story/305823/starship-prototype-lands-doesnt-explode.html`. In the webpage (D_1), three extra news pieces (D_2, D_3, D_4) from *nytimes*, *newser*, and *CNBC* are linked. All of these four news articles report on the same event of starship prototype landing.

Datasets	# examples			avg. doc len		avg. summ len		vocabulary size	
	train	valid	test	words	sents	words	sents	document	summary
CNN	90,266	1,220	1,093	760.50	33.98	45.70	3.59	343,516	89,051
DailyMail	196,961	12,148	10,397	653.33	29.33	54.65	3.86	563,663	179,966
NY Times	589,284	32,736	32,739	800.04	35.55	45.54	2.44	1,399,358	294,011
XSum	204,045	11,332	11,334	431.07	19.77	23.26	1.00	399,147	81,092
Newsroom	995,041	105,760	105,760	658.60	—	26.70	—	6,925,712	
MiRAnews	119,150	13,018	15,670	690.20	32.82	33.24	1.81	736,496	136,304

Table 1: Comparison of summarization datasets: size of training, validation, and test set, average document (source) and summary (target) length (in terms of words and sentences), and vocabulary size for both source and target. The numbers for CNN DailyMail, NY Times, and XSum are reported in Narayan et al. (2018a). The numbers for Newsroom are reported in Grusky et al. (2018). All tokens in MiRAnews vocabulary are lowercased.

MiRAnews(S-D), we select the first three sentences in the main document, and report ROUGE scores (Lin and Hovy, 2003) with respect to the gold summary. For MiRAnews(S-A), we select the first three sentences in each of the assisting documents and calculate ROUGE with respect to the gold summary individually; the reported ROUGE is then averaged over the individual documents. Furthermore, we use the *extractive oracle* (EO), which is often used as an upper bound for extractive models (Nallapati et al., 2017; Narayan et al., 2018b). It creates an oracle summary by selecting the best possible set of sentences in the document that gives the highest ROUGE score with respect to the gold summary.[5] For MiRAnews(S-D), we select the best three sentences in the main document as the summary, while for MiRAnews(S-A), we choose the best three sentences from all assisting documents as the summary. All selected summaries are evaluated using ROUGE against gold summaries. Higher ROUGE scores intuitively correspond to more extractive summaries.

The middle and right sections in Table 2 show the Lead and EO results, respectively. Both reach high scores on MiRAnews(S-D), while EO shows that improved content selection helps more. Although both methods achieve a much worse performance on MiRAnews(S-A) compared to MiRAnews(S-D), ROUGE scores are comparable to the ones reached on XSum. This confirms the conclusions we draw from the n-grams novelty metric.

3.2 Informativeness of Assisting Documents

Next, we evaluate the informativeness of the assisting documents with the following four metrics: We use n-gram novelty and Ext-Oracle from

[5] We use the greedy method from `https://github.com/pltrdy/extoracle_summarization`.

Dataset	% of novel n-grams in gold summary				LEAD			EO		
	1-gram	2-gram	3-gram	4-gram	R1	R2	RL	R1	R2	RL
CNN	16.75	54.33	72.42	80.37	29.15	11.13	25.95	50.38	28.55	46.58
DailyMail	17.03	53.78	72.14	80.28	40.68	18.36	37.25	55.12	30.55	51.24
NY Times	22.64	55.59	71.93	80.16	31.85	15.86	23.75	52.08	31.59	46.72
XSum	35.76	83.45	95.50	98.49	16.30	1.61	11.95	29.79	8.81	22.65
MiRA(S-D)	16.31	35.43	42.72	45.75	38.38	28.78	34.24	59.38	47.71	53.18
MiRA(S-A)	32.11	75.90	90.62	94.96	18.32	4.10	12.35	34.42	12.76	23.33
MiRA(S-D&A)	10.29	30.36	40.01	44.04	—	—	—	61.36	49.18	54.47

Table 2: Corpus bias towards extractive methods in popular dataset and MIRANEWS. We show the proportion of novel n-grams in gold summaries. We also report ROUGE scores for the LEAD baseline and the extractive oracle system EXT-ORACLE. Results are computed on the test set. The numbers for CNN, DailyMail, NY Times and XSum are reported by Narayan et al. (2018a). For MIRANEWS, S-D, S-A and S-D&A represent summary-document, summary-assisting document and summary-document & assisting document, respectively.

the previous section for measuring extractive token overlap. We also introduce two new metrics based on semantic similarity, which abstracts away from the actual tokens and is thus better suited for abstractive summarization.

• **N-gram novelty** MIRANEWS(S-D&A) in Table 2 reports the n-gram novelty of the summaries with respect to their main and assisting documents, which is substantially lower than MIRANEWS(S-D). Introducing the assisting documents contributes new information to support the summary better.

• **EO** MIRANEWS(S-D&A) in Table 2 contains the best three sentences from the main and assisting documents against the summary. The higher ROUGE scores on MIRANEWS(S-D&A), as compared to MIRANEWS(S-D), indicate that assisting documents A contribute additional information to the summaries, which is absent from the main document D.

• **Summary Fact-weights** evaluate the semantic correspondence between a document and its summary using a representation based on "*facts*". We follow Xu et al. (2020) and represent facts in a sentence by adapting Semantic Role Labelling (Palmer et al., 2005), which roughly captures "who did what to whom" in terms of predicates and their arguments. The facts in the document and summary are represented as $\{F_1^D, F_2^D, \cdots F_I^D\}$ and $\{F_1^S, F_2^S, \cdots F_J^S\}$, respectively. We apply automatic content weighting as defined in (Xu et al., 2020) and weight each fact F_j in the summary using its maximum semantic similarity to the facts in the document $w_j^f = \max_{i \in I} d_{ij}^f$, where d_{ij}^f is the semantic similarity based on BERT embeddings (Devlin et al., 2019). The *Summary Fact-weights* score is then defined as the average weights over all facts in the summary:

$$\textit{SFweights} = \text{avg}_{j=1\cdots J} w_j^f \in [-1, 1] \quad (1)$$

A high *SFweights* score indicates that the facts in the summaries are well supported by the facts mentioned in the documents.

The top section in Table 3 shows *SFweights* scores reported on MIRANEWS(S-D), MIRANEWS(S-A) and MIRANEWS(S-D&A), which weight facts in the summaries using facts in the main document, assisting documents, and both, respectively. As expected, *SFweights* on MIRANEWS(S-D) is higher than on MIRANEWS(S-A), indicating that the summary mainly contains facts from the main document D and can't be generated from assisting documents alone. However, *SFweights* on MIRANEWS(S-D&A) is higher than on MIRANEWS(S-D), which indicates that the assisting documents provide additional information beyond the main document and still preserve the facts in the summary.

• **Assist Rate** extends *SFweights* by first weighting the facts in the summary using the main document $\left[w_1^{fc}, w_2^{fc}, \cdots w_J^{fc}\right]$, and the assisting document $\left[w_1^{fa}, w_2^{fa}, \cdots w_J^{fa}\right]$. It is then defined as:

$$\textit{AsstRate} = \frac{\sum_{j=1}^{J} f\left(w_j^{fc}, w_j^{fa}\right)}{J} \quad (2)$$

$$f\left(w_j^{fc}, w_j^{fa}\right) = \begin{cases} 1, & \text{if } w_j^{fa} > w_j^{fc}. \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

where J is the number of facts in the summary. *AsstRate* represents the percentage of the facts in the summary that are *better represented* in the assisting documents than in the main document.[6] We also extend the fact-level *AsstRate* to the summary level, where we report the proportion of summaries in the entire corpus whose fact-level *AsstRate* is

[6]While the main document might contain the facts, their structure is more accurately covered in assisting documents.

Metrics	Results
SFweights MiRA(S-D)	0.633
SFweights MiRA(S-A)	0.584
SFweights MiRA(S-D&A)	0.658
AsstRate [fact level] (%)	27.67
AsstRate [summary level] (%)	30.20

Table 3: Summary Fact-weights (*SFweights*) and Assist Rate (*AsstRate*) show that the assisting documents provide additional information beyond the main document to the summary.

over 0. The bottom section in Table 3 shows that more than 27% of facts existing in 30% of summaries are better grounded on assisting documents.

4 Benchmarks

4.1 Baselines

After establishing the lower and upper bounds for extractive summarization models (see Section 3.1), we mainly focus on abstractive approaches in our experiments. Many existing powerful sequence to sequence models, e.g. BART (Lewis et al., 2020), target conditional text generation tasks including summarization. Specific instances of Transformer-based (Vaswani et al., 2017) models, such as Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), PEGASUS (Zhang et al., 2020a), HEPOS (Huang et al., 2021) and Hierarchical Transformer (HT) (Liu and Lapata, 2019a), are designed for encoding long sequences.

In order to measure the effect that transfer learning has on MIRANEWS, we try BART-large[7] (Lewis et al., 2020) which is pre-trained and can take 1024 words as input, and HT[8] (Liu and Lapata, 2019a) which is trained from scratch and can handle a longer input of up to 2000 words. We test four different variants for both models:

• **Single (-S):** We only consider the main document as the input for generating the summary, replicating the SDS setup.

• **Concatenation (-C):** We simply append the assisting documents at the end of the main document. Since each document contains around 700 words on average (see Table 1), we truncate the main document to half the size of the model capacity, i.e. 500 words for BART-large and 1000 words for HT, respectively. To include information from all assisting documents, we truncate each of them to fill the remaining half of the model capacity evenly.

• **Pipeline (-P):** Previous approaches T-DMCA (Liu et al., 2018), TLM (Pilault et al., 2020) and SEAL (Zhao et al., 2020a) show that long input settings for abstractive summarization benefit from a content extraction preprocessing step. We thus introduce a simple weakly supervised content extraction method for the assisting documents, and concatenate the selected content to the end of the main document on the input. Note that the content selection in MIRANEWS is conditioned on the main document, which is different from content selection in both SDS and MDS that select sentences without additional conditioning.

In particular, we first compute a contextual embedding for each sentence in both main and assisting documents using BERT (Devlin et al., 2019), represented as $D^{emb} = \{e_1^D, e_1^D, \cdots e_N^D\}$ and $A^{emb} = \{e_1^A, e_1^A, \cdots e_K^A\}$. Then we calculate the semantic relevance for each sentence in the assisting documents with respect to each sentence in the main document, as the cosine distance between their sentence embeddings. In turn, we select the sentence k in the assisting documents if:

$$\alpha_1 < \text{avg}_{n=1:N}\text{cosdist}\left(e_n^D, e_k^A\right) < \beta_1, \text{and}$$
$$\alpha_2 < \max_{n=1:N}\text{cosdist}\left(e_n^D, e_k^A\right) < \beta_2, \text{and}$$
$$\alpha_3 < \min_{n=1:N}\text{cosdist}\left(e_n^D, e_k^A\right) < \beta_3.$$

All thresholds are calculated on the training set using the gold content selection introduced in the following variant.[9]

• **Gold (-G):** We introduce a "heuristic" upper bound baseline by replacing the weakly supervised procedure above with gold content selection, following a procedure introduced by (Pilault et al., 2020; Nallapati et al., 2017) We select top sentences s_D from both main and assisting documents based on their extraction scores computed against sentences s_S from the ground-truth summary S: $\text{SCORE}_{ext}(s_D) = \frac{1}{3}\sum_{r\in 1,2,L}\text{ROUGE}_r\left(s_D, s_S\right)$, where $s_D \in D \cup \mathbf{A}; s_S \in S$. We clean up the sentences that are selected multiple times.

[7] Implementation used: https://huggingface.co/transformers/model_doc/bart.html.

[8] We use the implementation from https://github.com/nlpyang/hiersumm.

[9] We calculate the avg. cosdist(), max. cosdist() and min. cosdist() for each sentence in the gold content selection with respect to the corresponding main document. Then we calculate the distribution of the scores in each of these three category in terms of mean μ and variance σ. The lower and upper bound thresholds in each category are $(\mu - \sigma)$ and $(\mu + \sigma)$. Hence we get α_1=0.73, β_1=0.83, α_2=0.81, β_2=0.91, α_3=0.59, β_3=0.75.

Systems	ROUGE			BertScore		
	R1	R2	RL	P	R	F1
BART-S	46.07	34.19	42.14	.701	.674	.684
BART-C	45.44	33.70	41.56	.701	.666	.679
BART-P	**46.32**	**34.31**	**42.29**	.701	**.677**	**.685**
HT-S	46.76	**36.18**	**43.22**	.685	.682	.680
HT-C	46.77	36.06	43.11	**.690**	.682	**.682**
HT-P	**46.83**	36.08	43.13	.684	**.686**	.681
BART-G	60.09	46.72	55.39	.769	.745	.755
HT-G	55.16	43.15	51.02	.716	.731	.721

Table 4: Evaluation on ROUGE and BertScore.

4.2 Evaluation Metrics

We evaluate the approaches described in Section 4.1 from four perspectives:

• **Similarity to Reference** focuses on evaluating the generated summary with respect to its similarity to a human-authored ground-truth reference summary. We adopt the exact-matching metric ***ROUGE*** (Lin and Hovy, 2003) and the soft-matching metric ***BertScore*** (Zhang et al., 2020b).

• **Extractiveness level** aims at the bias of each system towards generating extractive summaries. We introduce the ***n-grams coverage***, which equals to 1 − n-gram novelty (see Section 3), to measure the percentage of n-grams in the generated summary that appear in the main and assisting documents. Higher n-gram coverage scores indicate that the system is more extractive.

• **Support from Assisting Documents** measures the proportion of information appearing in the generated summary that originates from assisting documents only. We propose the ***n-grams coverage*** over n-grams in the generated summary with respect to the n-grams that appear **only** in the assisting documents (i.e, not in the main document).

• **Extrinsic Hallucination** aims at evaluating how much the facts in the generated summary are grounded in the main and the assisting documents. We adopt the ***SFweights*** introduced in Section 3.2. A high SFweights score indicates that the facts in the generated summary are unlikely to be a result of extrinsic hallucination.

5 Experiment results

Similarity to Reference. The results of reference-based automatic metrics are shown in Table 4. The performance of BART and HT are comparable in most of the variants, which indicates that systems trained from scratch on MIRANEWS are able to achieve similar performance to the systems fine-tuned on the pre-trained checkpoints.

On most metrics, the concatenation variants (-C) of the models perform worse than the pipeline approaches (-P) and SDS-trained systems (-S). On the other hand, both -P outperform the -S systems in most cases. The gold systems (-G) achieve the best performance with a large margin. The performance of BART-G is even comparable with the upper bound of the extractive models (EO generated from MIRANEWS(S-D&A)). Hence, we conclude that (1) introducing assisting documents benefits the abstractive summary generation; (2) better content selection improves the performance of the abstractive models; (3) the margin between the gold upper baseline and the rest is notable, which suggests that there is room for improvement for content selection.

Extractiveness Level. The results are shown in the left section of Table 5. N-grams coverage scores for HT are much higher than BART's, with 4-grams over 90%. This indicates that HT tends to generate very extractive outputs. For each of the two models, the concatenation systems are more extractive than single-document and pipeline systems. For the BART variants, the gold system leans to generate more abstractive summaries compared to the remaining variants; for HT, the gold system is as extractive as all other variants.

Support from Assisting Documents. The middle section of Table 5 shows the amount of information each system learns from the assisting documents alone. In both models, the gold, concatenate and pipeline variants include substantially more expressions occurring in the assisting documents compared to the single-document systems.

Extrinsic Hallucination. The results in the right section of Table 5 show that HT achieves a higher SFweights score, i.e. lower level of extrinsic hallucination, than BART – probably due to the high extractiveness of HT. In other words, extractive summaries that copy sentences directly from the document tend to maintain higher SFweights scores. On the other hand, BART systems demonstrate a much higher level of abstractiveness, while preserving a similar SFweights score with HT. Thus, the BART systems do not introduce more hallucinations while generating abstractive summaries.

Within each of the two models, summaries generated by each variant preserves a roughly similar level of extractiveness. In both models, concatenation and pipeline systems achieve a lower

Systems	Extractiveness level (%)				Support from Assisting Documents (%)				SFweights
	1-gram	2-gram	3-gram	4-gram	1-gram	2-gram	3-gram	4-gram	
BART-S	87.24	72.94	63.85	57.61	1.76	1.32	0.55	0.24	.814
BART-C	88.37	75.74	66.98	60.71	2.99	3.22	2.24	1.62	**.860**
BART-P	87.79	74.16	65.19	59.00	2.57	2.37	1.39	0.90	.850
HT-S	98.14	95.70	93.98	92.82	0.51	0.38	0.16	0.08	.840
HT-C	99.48	98.46	97.58	96.86	1.53	2.37	2.33	2.19	**.881**
HT-P	99.20	97.86	96.83	96.00	0.92	1.18	1.01	0.87	.860
BART-G	87.14	71.42	60.94	53.91	4.22	4.96	3.59	2.65	.817
HT-G	98.88	96.57	94.58	93.10	2.82	4.56	4.48	4.17	.845

Table 5: Evaluation of extractiveness level using **n-gram coverage** (left), support from assisting documents (middle) calculated by **n-gram coverage** with respect to the n-grams appearing in the assisting documents only, and extrinsic hallucination (right) evaluated using SFweights.

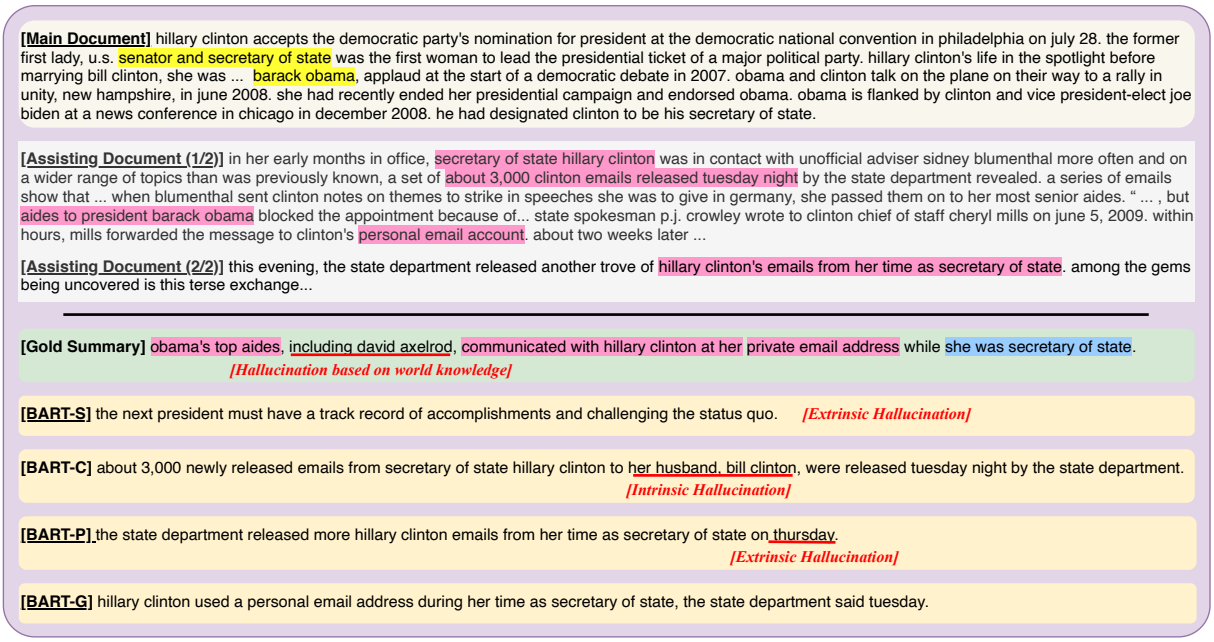

Figure 4: An example from MIRANEWS, where the key information in the gold summary and summaries generated by systems conditioning on the main document (BART-S) or both on the main and assisting documents (rest variants) were only mentioned in the assisting documents. Facts in the gold summary supported by the assisting documents only are highlighted in pink. Information grounded in both main and assisting documents is highlighted in blue. Other error type examples, including *Extrinsic Hallucination*, *World Knowledge-based Extrinsic Hallucination* and *Intrinsic Hallucination* in summaries are ***[labeled in red]***.

level of extrinsic hallucination compared to the single-document systems. SFweights for BART-G is lower than most other setups, probably due to a high level of abstractiveness in this system. To better understand the relation between introducing assisting documents and reducing extrinsic hallucinations, we conduct an example-based analysis in the next section.

6 Hallucination Analysis

We manually identify 4 types of hallucinations from a small random sample (30 main/assisting documents and summaries) from the development set of MIRANEWS, as summarised in Table 6. In particular, we examined claims in the summaries that were not mentioned in the main or assisting documents and were (1) erroneous (*Extrinsic Hallucinations*), (2) factual possibly due to pretraining (*World knowledge*), (3) only mentioned in the assisting document correctly (*Grounded Asst.*), or (4) mentioned in the main document in a different way (*Intrinsic*). We omit the HT variants from our analysis as their output is more extractive, and therefore less prone to hallucinations. The SDS variant of BART (BART-S) has the highest percentages of extrinsic (7) and intrinsic (4) hallucinations and a number of claims that are based on world knowledge (3). On the other hand, the inclusion of assisting documents sees an overall reduction

Systems	Extr.	World	Asst.	Intr.
GOLD	1	**10**	11	0
BART-S	**7**	3	0	**4**
BART-C	0	0	6	2
BART-P	3	1	3	1
BART-G	3	0	**11**	2

Table 6: Manual analysis of types of hallucinations (counterfactual extrinsic [**Extr.**], factual extrinsic based on world knowledge [**World**], grounded exclusively on assisting documents [**Asst.**], intrinsic [**Intr.**]) on a sample of 30 summaries from MIRANEWS.

in both types with up to 55% on extrinsic hallucinations when using the assisting documents for training efficiently (BART-G). At the same time, we observe 'extrinsic hallucinations' that are correctly grounded only on the assisting documents (11), and rarely *guessed* based on pre-training (only 1 fact based on world knowledge). Interestingly, we also observed a number of facts (10) in the gold summary that are grounded exclusively on the assisting documents, further supporting the value of our approach. An example of outputs from variants of BART is shown in Figure 4.

7 Related Work

Single Document Summarization aims to compress a single textual document while keeping salient information. SDS includes two directions: extractive summarization (Nallapati et al., 2017) which aims at extracting salient sentences from the input document, and abstractive summarization (See et al., 2017; Narayan et al., 2018a; Yang et al., 2019; Liu and Lapata, 2019b; Liu et al., 2020; Rothe et al., 2020; Raffel et al., 2020) which generates a novel short representation of the input.

Multi-Document Summarization aims to compress multiple textual documents to a shorter summary (Fabbri et al., 2019). Approaches mainly focus on increasing the capacity of the encoder to process longer inputs (Liu and Lapata, 2019a; Beltagy et al., 2020; Zaheer et al., 2020; Zhang et al., 2020a; Huang et al., 2021), leveraging knowledge graphs (Fan et al., 2019; Li et al., 2020; Jin et al., 2020), and including content selection steps (Nayeem et al., 2018; Wang et al., 2020; Xu and Lapata, 2020; Grenander et al., 2019; Liu et al., 2018).

Hallucinations in Summarization are a well established problem (Maynez et al., 2020; Cao et al., 2018; Falke et al., 2019). Previous research aimed to reduce hallucination by adapting model architectures, training and decoding, e.g. Cao et al. (2018); Zhang et al. (2020c); Falke et al. (2019); Zhao et al. (2020b). However, we are the first research aiming to reduce the hallucinations by adapting the dataset.

8 Conclusions and Future Work

In this work, we found that up to 36% facts in the ground truth summaries in traditional SDS datasets are not faithful to the source article. In other words, the ground truth summaries also contain 'extrinsic hallucinations'. Summarization models trained on such data will be prone to extrinsic hallucinations. To tackle this problem, we introduce a new task, Multi-Resource-Assisted News summarization, which produces a summary based on the events present in the main article while reaching to a set of assisting documents for complementary background. We release the MIRANEWSdataset, which includes multiple assisting news articles from different news resources for each document-summary pair. Our newly introduced evaluation metrics confirm that introducing assisting documents offers better grounding to more than 27% facts in the reference summaries. We report benchmark results on MIRANEWS. We also show that the model trained with assisting documents produces 55% less counterfactual hallucinations than a model trained only with main documents.

In future work, we plan to explore a retrieval-based approaches (Azzopardi and Staff, 2012; Bouras and Tsogkas, 2012) that are able to search and filter relevant assisting documents for a given news event, without the help of human-edited resources such as `newser.com`. In the paper, we demonstrated that the assisting documents contain useful facts to support the summarization of the main news event. Thus, efficient content selection that eliminates noise and grounds in the relevant facts appearing in either main or assisting documents will also be explored in our future work.

Acknowledgments

This research received funding from the EPSRC project AISec (EP/T026952/1), Charles University project PRIMUS/19/SCI/10, a Royal Society research grant (RGS/R1/201482), a Carnegie Trust incentive grant (RIG009861). This research also received funding from Apple to support research at Heriot-Watt University and Charles University. We thank the anonymous reviewers and the area chair for their helpful comments and hard work.

References

Joel Azzopardi and Christopher Staff. 2012. Incremental clustering of news reports. *Algorithms*, 5(3):364–378.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Christos Bouras and Vassilis Tsogkas. 2012. A clustering technique for news articles using WordNet. *Knowledge-Based Systems*, 36:115–128.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *AAAI Conference on Artificial Intelligence*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

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.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

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.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. 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 6019–6024, Hong Kong, China. Association for Computational Linguistics.

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.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, page 1693–1701, Cambridge, MA, USA. MIT Press.

Luyang Huang, Shuyang Cao, Nikolaus Nova Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. *CoRR*, abs/2104.02112.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6244–6254, Online. 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.

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.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

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, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 3075–3081. AAAI Press.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. 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*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. 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 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Mir Tafseer Nayeem, Tanvir Ahmed Fuad, and Yllias Chali. 2018. Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1191–1204, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ani Nenkova. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *Proceedings of the 20th National Conference on Artificial Intelligence - Volume 3*, page 1436–1441. AAAI Press.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Jonathan Pilault, Raymond Li, Sandeep Suramanian, and Chris Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 8051–8067.

Abigail See, Peter Liu, and Christopher Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Kexiang Wang, Baobao Chang, and Zhifang Sui. 2020. A spectral method for unsupervised multi-document summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 435–445, Online. Association for Computational Linguistics.

Xinnuo Xu, Ondřej Dušek, Jingyi Li, Verena Rieser, and Ioannis Konstas. 2020. Fact-based content weighting for evaluating abstractive summarisation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5071–5081, Online. Association for Computational Linguistics.

Yumo Xu and Mirella Lapata. 2020. Coarse-to-fine query focused multi-document summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3632–3645, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Manzil Zaheer, Guru Prashanth Guruganesh, Avi Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Minh Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Mahmoud El Houssieny Ahmed. 2020. Big bird: Transformers for longer sequences.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020c. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, Online. Association for Computational Linguistics.

Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. SEAL: segment-wise extractive-abstractive long-form text summarization. *CoRR*, abs/2006.10213.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020b. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics.