Which Information Matters? Dissecting Human-written Multi-document Summaries with Partial Information Decomposition

Laura Mascarell and **Yan L'Homme** and **Majed El Helou**
ETH Zurich
lmascarell@inf.ethz.ch, lhommey@ethz.ch, majed.elhelou@inf.ethz.ch

Abstract

Understanding the nature of high-quality summaries is crucial to further improve the performance of multi-document summarization. We propose an approach to characterize human-written summaries using partial information decomposition, which decomposes the mutual information provided by all source documents into union, redundancy, synergy, and unique information. Our empirical analysis on different MDS datasets shows that there is a direct dependency between the number of sources and their contribution to the summary.

1 Introduction

Multi-document Summarization (MDS) consists of providing an abridged version of multiple documents. While some abstractive summarization approaches concatenate all documents into a single input (Johner et al., 2021; Xiao et al., 2022), the large size of input text represents a major challenge in MDS. Therefore, most methods implement two-stage approaches that extract salient text spans based on different heuristics (Lebanoff et al., 2018; Liu et al., 2018; Liu and Lapata, 2019; Zhu et al., 2021). The text is then fed into a summarization model under the assumption that a high-quality summary is based on such information. While earlier work quantifies the properties of human-written MDS based on n-gram matching (Banko and Vanderwende, 2004), we argue that there is a lack of in-depth analyses. Without an understanding of the nature of summaries, improving the quality of MDS remains vague and without clear interpretability.

This work sheds light on what information constitutes a high-quality multi-document summary. In particular, we propose to categorise the summary information into information provided by at least one source (**union**) or by a **unique** source, **redundant** information from all source documents, and even new information derived from considering them jointly (**synergy**). In information theory, Partial Information Decomposition (PID) decomposes information in the same way to assess how information about a target is distributed among multiple source variables (Williams and Beer, 2010).

We therefore implement PID in MDS and present SPIDer; a novel approach to quantify the degree to which the PID components—such as redundancy or unique information—contribute to a summary.[1] We then perform an empirical analysis on human-written summaries from different MDS datasets using our approach. Our results demonstrate that the number of sources has a direct dependence on how they contribute to the summary. We also show that, surprisingly, the order of the source documents matters, and the first three documents are frequently considered as the main source of unique information for any number of sources.

To the best of our knowledge, we present the first fine-grained information analysis in human-written MDS. We open-source SPIDer[2] and hope that it helps to enhance the performance of future MDS methods. We suggest that our PID approach could also be used to automatically build MDS datasets that align with human quality in future work.

2 Information Theory Background

Mutual Information (MI) (Shannon, 1948) has been widely used in NLP tasks to quantify the information that a source provides about an output (Li et al., 2016; Li and Jurafsky, 2016; Takayama and Arase, 2019; Padmakumar and He, 2021; Mascarell et al., 2023). However, mutual information is insufficient in the context of MDS, as it can only be applied to pairs of random variables.

Partial Information Decomposition (PID) tackles the multivariate problem and decomposes the information that a set of sources conveys about a target

[1] Our implementation also works for single inputs.
[2] GitHub: mediatechnologycenter/SPIDer

Findings of the Association for Computational Linguistics: ACL 2024, pages 5333–5338

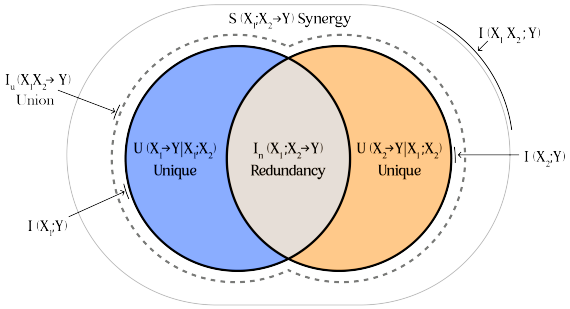

Figure 1: Relationship between the PID components *union*, *redundancy*, *synergy*, and *unique* information that two sources X_1 and X_2 provide about a target Y. $I(X_1, X_2; Y)$ represents the information that both sources provide jointly about Y, whereas $I(X_1; Y)$ and $I(X_2; Y)$ represent the information that each source provides individually.

into union, redundancy, unique, and synergistic information (Williams and Beer, 2010). To date, PID has only been applied in NLP to measure morphological fusion in Socolof et al. (2022). Figure 1 illustrates the relationships between the different PID components for two sources, X_1 and X_2, and a target Y. See a textual example in Table 1.

In this work, we consider the PID approach proposed in Kolchinsky (2022), which is based on the definitions of union and intersection from set theory. In contrast to previous work, Kolchinsky (2022)'s approach can be applied to any number of sources, a crucial requirement in our MDS setting.

Formally, given a set of source random variables $\mathcal{X} = \{X_1, ..., X_n\}$ and a target Y, *redundancy* $I_\cap(\mathcal{X} \to Y)$ is the intersection of information among the sources about Y. This redundancy is the maximum information Q we can obtain about Y that is less informative than any of the sources:

$$I_\cap(\mathcal{X} \to Y) := \sup_Q I(Y; Q) \mid \forall i, Q \sqsubset X_i \quad (1)$$

where $\sqsubset$ is an ordering relation that determines when X_i is more informative than Q. Conversely, *union* $I_\cup(\mathcal{X} \to Y)$ is the minimum information Q we can obtain about the target Y that is more informative than any of the sources:

$$I_\cup(\mathcal{X} \to Y) := \inf_Q I(Q; Y) \mid \forall i, X_i \sqsubset Q \quad (2)$$

Finally, *unique* and *synergistic* information are derived from Eq. (1) and Eq. (2), respectively.

3 Decomposing Information in MDS

We adopt the PID approach in Kolchinsky (2022) to MDS, considering sentences as units of information. Formally, let $\mathcal{X} = \{D_1, ..., D_n\}$ be a set of n source documents, where each document is a collection of sentences $D_i = \{d_i^1, ..., d_i^{|D_i|}\}$, and a multi-document summary of m sentences $S = \{s^1, ..., s^m\}$. Using Eq. (1) and (2), we define *redundancy* and *union* in MDS as follows:

$$I_\cap^{\text{MDS}}(\mathcal{X} \to S) := \sup_{D \in \mathcal{D}} I(S; D) | \forall i, D \sqsubset D_i \quad (3)$$

$$I_\cup^{\text{MDS}}(\mathcal{X} \to S) := \inf_{D \in \mathcal{D}} I(S; D) | \forall i, D_i \sqsubset D \quad (4)$$

where $D = \{d^1, ..., d^{|D|}\}$ is a collection of document sentences from all possible sentences $\mathcal{D}$,[3] $I(S; D)$ represents the MI between the summary sentences S and D, and $\sqsubset$ is an ordering relation. We formally define the ordering $\sqsubset$ and MI in our MDS setting later in this section.

We then define *unique* information and *synergy* using Eq. (3) and (4) as in Kolchinsky (2022)

$$U^{\text{MDS}}(D_i \to S | \mathcal{X}) = I(S; D_i) - I_\cap^{\text{MDS}}, \quad (5)$$

$$S^{\text{MDS}}(\mathcal{X} \to S) = I(S; \mathcal{X}) - I_\cup^{\text{MDS}}, \quad (6)$$

where $I(S; D_i)$ and $I(S; \mathcal{X})$ are the MI between the sentences of a summary and a specific document D_i or all source documents $\mathcal{X}$, respectively.

Pairwise Mutual Information $I(S; D)$ quantifies the mutual information between summary sentences S and a collection of source sentences D. To compute $I(S; D)$, we measure pairwise mutual information on all summary-source sentence pairs $pmi(s; d) = log \frac{p(s;d)}{p(s)p(d)}$ as in Padmakumar and He (2021), using a language model to estimate probabilities.[4] To compute $p(s; d)$ the sentences s and d

Source X_1	Kimchi is fermented cabbage.
Source X_2	Fermented foods are rich in probiotics.
Target Y	Fermented foods, such as Kimchi, are rich in probiotics.
Unique X_1	The nature and preparation of kimchi.
Unique X_2	A general characteristic of fermented foods.
Redundancy	Information about *fermented*.
Synergy	Inferring that Kimchi is a fermented food.

Table 1: Example with two sources, X_1 and X_2, and the PID information that they provide about Y.

[3] Since optimizing over all collections of sentences is computationally intractable, we implement two optimization strategies. First, we assume that a summary should only contain information from the sources. Therefore, we restrict the set of all possible sentences to the sentences from the sources $\mathcal{D} := \mathcal{P}(\cup_{i=1}^n D_i)$, which results in a finite optimization problem with set size $|\mathcal{D}| = 2^{\sum_i |D_i|}$. Second, since the growth of $|\mathcal{D}|$ is still exponential, we implement beam search to find an approximation of the optimal collection of sentences for union and redundancy.

[4] GPT-2 checkpoint: openai-community/gpt2-large

Dataset	Source	Summary	#Sources
MultiNews	32.7	13.8	from 2 to 10
WikiNewsSum	21.7	17.5	from 2 to 10
DUC2004	25.5	6.9	10
WCEP	20.0	3.2	10

Table 2: Overview of the analyzed data. Average sentence count per summary and per source document, and the number of sources considered in each dataset.

are concatenated and their joint probability is estimated with the language model. We define $I(S;D)$ as the expected value of all $pmi(s;d)$, which is the sum of all values weighted by their likelihood,

$$I(S;D) = \mathop{\mathbb{E}}_{s,d\sim S,D}[pmi(s;d)] \tag{7}$$

Ordering Relation $D \sqsubset D'$ specifies that a collection of source sentences D' is more informative than a different collection D. That is, the information that D provides to a summary S is contained within the information that D' provides to S. Hence, we define our ordering relation as:

$$\begin{aligned} D \sqsubset D' \iff & I(S;D) \leq I(S;D') \\ & \wedge \forall s \in S : I(S;D)_s \leq I(S;D')_s \end{aligned} \tag{8}$$

where $I(S,D)_s = \mathbb{E}_{d\sim D}[pmi(s;d)]$. Note that the second component of Eq. (8) guarantees that all the information provided by the individual sentences in D is also contained in D'.

4 PID of Human-written Summaries

We analyze the information components of human-written summaries from multiple MDS datasets using our framework. Specifically, we consider a random sample of 100 instances per dataset and number of sources, if available (see Table 2).[5]

4.1 Datasets

MultiNews MDS dataset of news articles from the website newser.com, where summaries are written by professional editors. The sources are the cited articles in each summary (Fabbri et al., 2019).

WikiNewsSum The summaries are full articles from the collaborative news platform Wikinews.org and the sources are the cited references. Sources are at least 1.5 times larger than its summary on a character basis (Calizzano et al., 2022).

[5]Samples with less than 100 instances: MultiNews with 9 (89 instances) and 10 sources (33); WikiNewsSum with 8 (89 instances), 9 (77), and 10 sources (51).

Dataset	Union	Synergy	Red.	Unique
MultiNews	1.0 (±0.0)	0.0 (±0.0)	0.48 (±0.2)	0.30 (±0.1)
WikiSum	1.0 (±0.0)	0.0 (±0.0)	0.48 (±0.2)	0.30 (±0.1)
DUC2004	1.0 (±0.0)	0.0 (±0.0)	0.43 (±0.2)	0.35 (±0.1)
WCEP	1.0 (±0.0)	0.0 (±0.0)	0.29 (±0.2)	0.41 (±0.2)

Table 3: Overall SPIDer scores across datasets (mean and standard deviation).

Dataset	2	3	4	5	6	7	8	9	10
MNews	3.9	4.1	4.9	4.8	4.8	5.0	5.3	5.1	5.6
WSum	3.8	4.3	4.6	5.1	4.7	4.8	5.1	5.1	5.1
DUC	-	-	-	-	-	-	-	-	4.5
WCEP	-	-	-	-	-	-	-	-	6.5

Table 4: Variance of unique information (%) across datasets and different numbers of sources.

DUC2004 This dataset consists of 50 sets of 10 articles from the Associated Press and New York Times newswires, each paired with four handwritten summaries. Humans were required to summarize each source independently before writing a multi-document summary of up to 665 characters.[6]

WCEP The dataset comprises summaries of news events from the Wikipedia Current Events Portal and the corresponding cited sources (only 1.2 on average). The set of sources are complemented with related articles from the Common Crawl archive (Gholipour Ghalandari et al., 2020), which results in a rather synthetic MDS dataset. Summaries are short, up to 30-40 words; hence, we only consider a subset consisting of multi-sentence summaries of 10 sources each.

4.2 Results

We compute SPIDer scores on all samples and compare the results among datasets and number of sources. Due to length differences in both summary and sources across datasets, we normalise all values by the total mutual information between the summary and the source documents $I(S;D_1,...,D_n)$.

We observe in Table 3 that *synergy* is negligible in all datasets, and therefore, *union* represents the total mutual information (see Eq. (6)). Figure 2 compares *redundancy* and *unique* information across different numbers of sources. The results show that redundancy decreases with the number of sources, while the unique information increases. That is, the more sources, the more they contribute individually to the summary. Additionally, DUC2004 scores are comparable to MultiNews and WikiNewsSum with 10 sources, whereas WCEP

[6]https://duc.nist.gov/duc2004/

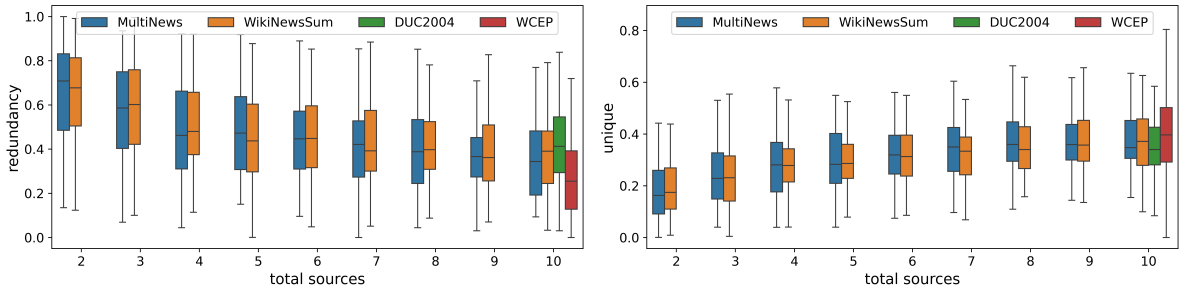

Figure 2: Redundancy (left) and unique (right) information scores across datasets and number of sources. The more sources, the less redundancy and the more unique information contributes to the summary. WCEP scores differ the most from the other datasets. Note that WCEP is extended with additional sources not considered in the summaries.

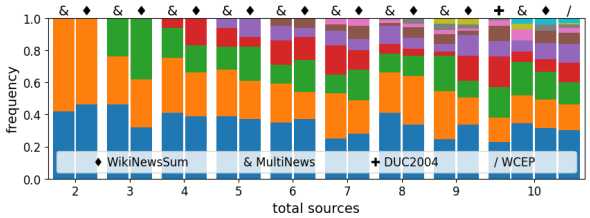

Figure 3: Frequency of each source contributing the most to the summary with their unique information across datasets and total number of sources. The first three sources (blue, orange, and green) contribute the most for any number of sources in all datasets.

shows significantly lower redundancy and higher unique information. WCEP is extended with Common Crawl articles, so it is unclear how these additional articles should contribute to the summary in a MDS task. The difference in SPIDer scores with the other datasets highlights the importance of using real MDS datasets for this task.

To get better insights into the individual contributions of the sources, we analyze their variance, where a value of 0 indicates that all sources contribute equally to the summary. Table 4 demonstrates that variance, and hence, the variability of their individual contributions, increases with the number of sources. Furthermore, we compute the frequency at which each source contributes the most with their unique information (Figure 3). Interestingly, we observe that the first three sources consistently contribute the most in all datasets, regardless of the number of sources. Similarly, Wolhandler et al. (2022) report that the information in multi-document summaries is often covered by a single source document. Our results also highlight a strong bias towards the order of the sources when summarizing multiple documents, consistently observed across the datasets.

Answer	Union	Synergy	Red.	Unique
Unrelated	0.04 (±0.3)	0.24 (±0.5)	0.01 (±0.2)	0.03 (±0.1)
Incorrect	0.05 (±0.3)	0.26 (±0.5)	0.02 (±0.2)	0.03 (±0.1)
Correct	0.05 (±0.3)	0.28 (±0.5)	0.02 (±0.2)	0.04 (±0.1)

Table 5: SPIDer scores on MultiRC as MDS data (mean and standard deviation).

5 Measuring Synergistic Information

Since synergy is negligible in the analyzed MDS summaries (see Section 4.2), we perform an additional experiment to assess whether our approach can measure synergistic information using the MultiRC dataset (Khashabi et al., 2018). Specifically, MultiRC is a reading comprehension dataset consisting of multi-sentence paragraphs and multiple-choice questions, which also specifies the sentences that are required to answer each question. Most importantly, the dataset ensures that correct answers can only be derived by considering multiple sentences jointly. That is, synergistic information should be prominent in correct answers.

To apply our PID approach, we transform MultiRC into a MDS dataset, where each instance comprises the set of sentences (sources) required to answer a question, and a question-answer pair concatenated into a single sentence (summary). Although both correct and incorrect answers share the same question, the former should result in higher synergy than the latter. We also generate unrelated instances for each set of source sentences, where the question-answer pairs are randomly sampled from a different paragraph (see Figure 6).

Table 5 confirms that synergistic information is the dominant information component and correct answers achieve the highest synergy. However, synergy is also present in unrelated instances. Given that synergy represents new information, this raises the question whether it could also be an indicator

Sources		
The story revolves around an upright and principled Police Officer, A.C.P. Ramakant Chaudhary whose eldest son Vikas is killed in a pre-planned accident.		
The day comes when Vishal confronts Baba Khan and Manna Shetty which leads to tension and gory situation for the A.C.P., as the ganglords threaten to eliminate the A.C.P. as well as his wife Revati and son Vishal.		
Summary	Correct Answer	What is the name of Revati's husband? Ramakant Chaudhary
Summary	Incorrect Answer	What is the name of Revati's husband? Baba Khan
Summary	Unrelated Q&A	How many people come to comfort the baby? 2

Table 6: Examples of sources-summary instances of MultiRC as a MDS dataset. Each sentence represents an independent source and the summary is the concatenation of the question and answer. For each question in the dataset, we generate an unrelated instance consisting of a question and answer from a different paragraph.

of hallucination. This is an interesting research direction for future work, since more experiments are needed to support this hypothesis.

6 Conclusion

We characterize the information present in human-written multi-document summaries to get insights into what information comprises a high-quality summary. In particular, we suggest to decompose the mutual information that the source documents provide about a summary and propose SPIDer, a novel approach to quantify such information using partial information decomposition. We then analyze human-written summaries from widely used MDS datasets. The results reveal that redundancy decreases, whereas unique information increases with the number of sources. Furthermore, the order of the documents has an impact on the summarization process, as the first three documents contribute the most in terms of unique information.

Limitations

Some limitations of our work can be traced to the beam-search-based approximation caused by the intractable sentence search space, or due to using sentences as our base unit of information. We also note that despite the remarkable recent advances, large language models' probability distributions possibly diverge from the true underlying distribution. Their approximation is, however, continuously improving and future models can directly be substituted into our method.

Ethics Statement

From an ethical perspective, it is important to underline the importance of transparency when using language models, as they are becoming nearly indistinguishable from human writers. We advocate for clear transparency when they are used. Our work promotes interpretability in the space of multi-document summarization, and we hope both interpretability and transparency will be cornerstones for future work in the field.

In all our experiments, we rigorously follow the ACL Code of Ethics, using pre-existing open-source benchmark datasets where privacy concerns were already addressed by the respective authors.

Acknowledgements

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

References

Michele Banko and Lucy Vanderwende. 2004. Using n-grams to understand the nature of summaries. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 1–4, Boston, Massachusetts, USA. Association for Computational Linguistics.

Rémi Calizzano, Malte Ostendorff, Qian Ruan, and Georg Rehm. 2022. Generating extended and multilingual summaries with pre-trained transformers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1640–1650, Marseille, France. European Language Resources Association.

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.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics.

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.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. 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 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Artemy Kolchinsky. 2022. A novel approach to the partial information decomposition. *Entropy*, 24(3).

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.

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.

Jiwei Li and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372*.

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. 2019. 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.

Laura Mascarell, Ribin Chalumattu, and Julien Heitmann. 2023. Entropy-based sampling for abstractive multi-document summarization in low-resource settings. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 123–133, Prague, Czechia. Association for Computational Linguistics.

Vishakh Padmakumar and He He. 2021. Unsupervised extractive summarization using pointwise mutual information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2505–2512, Online. Association for Computational Linguistics.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Michaela Socolof, Jacob Louis Hoover, Richard Futrell, Alessandro Sordoni, and Timothy J. O'Donnell. 2022. Measuring morphological fusion using partial information decomposition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 44–54, Gyeongju, Republic of Korea. International Committee on 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.

Paul L Williams and Randall D Beer. 2010. Non-negative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Ruben Wolhandler, Arie Cattan, Ori Ernst, and Ido Dagan. 2022. How "multi" is multi-document summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5761–5769, 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.

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.