Semantic Overlap Summarization among Multiple Alternative Narratives: an Exploratory Study

Naman Bansal, Mousumi Akter and **Shubhra Kanti Karmaker ("Santu")**
Big Data Intelligence (BDI) Lab
Department of Computer Science and Software Engineering
College of Engineering, Auburn University
{nbansal, mza0170, sks0086}@auburn.edu

Abstract

In this paper, we introduce an important yet relatively unexplored NLP task called **S**emantic **O**verlap **S**ummarization (**SOS**), which entails generating a single summary from multiple alternative narratives which can convey the *common information* provided by those narratives. As no benchmark dataset is readily available for this task, we created one by collecting $2,925$ alternative narrative pairs from the web and then, went through the tedious process of manually creating 411 different reference summaries by engaging human annotators. As a way to evaluate this novel task, we first conducted a systematic study by borrowing the popular *ROUGE* metric from text-summarization literature and discovered that *ROUGE* is not suitable for our task. Subsequently, we conducted further human annotations to create 200 document-level and $1,518$ sentence-level ground-truth *overlap labels*. Our experiments show that the sentence-wise annotation technique with three overlap labels, i.e., {Absent (A), Partially-Present (PP), and Present (P)}, yields a higher correlation with human judgment and higher inter-rater agreement compared to the ROUGE metric.

1 Introduction

In this paper, we look deeper into the challenging yet relatively under-explored area of automatic summarization of multiple alternative narratives with different perspectives. To be more specific, we formally introduce a new NLP task called **S**emantic **O**verlap **S**ummarization (**SOS**) from multiple alternative narratives and conduct a systematic study of this task by creating a benchmark dataset as well as exploring how to accurately evaluate this task. *SOS* essentially means the task of *summarizing the overlapping information* present in multiple alternate narratives by cross-verifying their information contents against each other. Computationally, our research question is the following:

Given two distinct narratives N_1 and N_2 of an event e, how can we automatically generate a single summary about e which conveys the common information provided by both N_1 and N_2?

Multiple alternative narratives appear frequently in a variety of domains, including education (Somasundaran et al., 2018), the health sector (Bijoy et al., 2021), businesses intelligence (Azeroual and Theel, 2018), content analysis (Hassan et al., 2020; Karmaker Santu et al., 2018b) and privacy (Wilson et al., 2018). Therefore, automatic summarization of multiple-perspective narratives has become a pressing need in this information explosion era and can be highly useful for digesting such multi-narratives at scale and speed.

Figure 1 shows a toy example of the *SOS* task, where both articles cover the same event related to "abortion". However, they report from different political perspectives, i.e., one from the *left* wing and the other from the *right* wing. For greater visibility, "Left" and "Right" wing reporting biases are represented by *blue* and *red* text, respectively. *Green* text denotes the common information in both news articles. The goal of the *SOS* task is to generate a summary that conveys the common/overlapping information provided by the *green* text.

At first glance, the *SOS* task may appear similar to a traditional multi-document summarization task where the goal is to provide an overall summary of the (multiple) input documents. However, the difference is that, for *SOS*, the goal is to provide summarized content with an additional constraint, i.e., the commonality criteria. There is no current baseline method or an existing dataset that exactly matches our task; more importantly, it is unclear which one is the right evaluation metric to properly evaluate this task. As a starting point, we frame *SOS* as a constrained seq-to-seq task where the goal is to generate a summary from two input documents that conveys the overlapping information present in both input text documents. However, the bigger challenge we need to address first is the following: *1) How can we evaluate this task?* and *2) How*

Proceedings of the 29th International Conference on Computational Linguistics, pages 6195–6207
October 12–17, 2022.

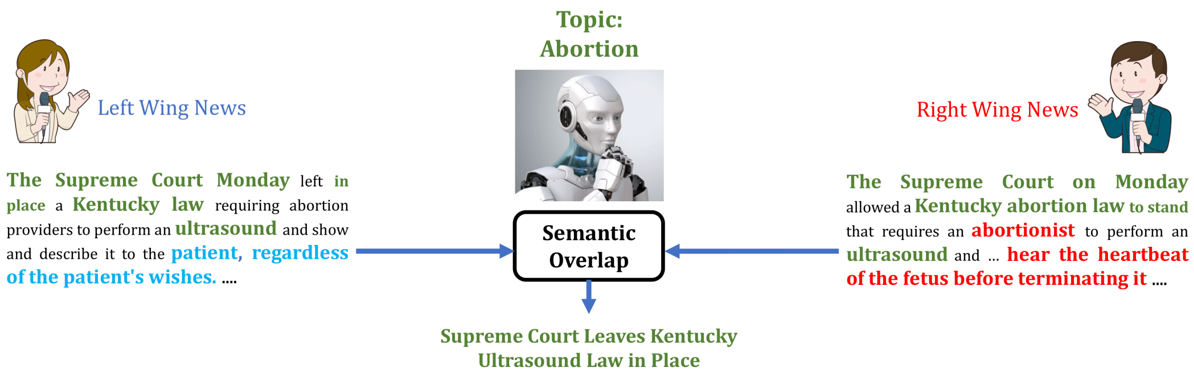

Figure 1: A toy example of *Semantic Overlap Summarization* (**SOS**) Task (from multiple alternative narratives). Here, an abortion issue-related event has been reported by two news media (left-wing and right-wing). "Green" Text denotes the common information from both news media, while "Blue" and "Red" text denotes the unique perspectives of *left* and *right* wing. Some real examples from the benchmark dataset are provided in the Table 3.

to create a benchmark dataset for this task? To address these challenges, we make the following contributions in this paper.

1. We formally introduce *Semantic Overlap Summarization* (**SOS**) (from multiple alternative narratives) as a new NLP task and conduct a systematic study by formulating it as a constrained summarization problem.
2. We create and release the first benchmark dataset consisting of $2,925$ alternative narrative pairs for facilitating research on the *SOS* task. Also, we went through the tedious process of manually creating 411 different ground-truth reference summaries and conducted further human annotations to create 200 document-level and $1,518$ sentence-level ground-truth *overlap labels* to construct the benchmark dataset.
3. As a starting point, we experiment with *ROUGE*, a widely popular metric for evaluating text summarization tasks, and demonstrate that *ROUGE* is NOT suitable for the evaluation of *SOS* task.
4. We do further human experiments, which show that sentence-level evaluation is the proper way to evaluate the *SOS* task. It also improves the inter-rater agreement compared to the traditional *ROUGE* metric and shows a higher correlation with human judgments.

2 Related Works

As *SOS* can be viewed as a multi-document summarization task with additional commonality constraints, text summarization literature is the most relevant to our work. Over the years, many paradigms for document summarization have been explored (Zhong et al., 2019). The two most popular among them are *extractive* approaches (Cao et al., 2018; Narayan et al., 2018; Wu and Hu, 2018; Zhong et al., 2020) and *abstractive* approaches (Bae et al., 2019; Hsu et al., 2018; Liu et al., 2017; Nallapati et al., 2016). Some researchers have also tried combining extractive and abstractive approaches (Chen and Bansal, 2018; Hsu et al., 2018; Zhang et al., 2019).

Recently, encoder-decoder-based neural models have become really popular for abstractive summarization (Rush et al., 2015; Chopra et al., 2016; Zhou et al., 2017; Paulus et al., 2017). It has become even more prevalent to train a general language model on a huge corpus of data and then transfer/fine-tune it for the summarization task (Radford et al., 2019; Devlin et al., 2019; Lewis et al., 2019; Xiao et al., 2020; Yan et al., 2020; Zhang et al., 2019; Raffel et al., 2019). Summary length control for abstractive summarization has also been studied (Kikuchi et al., 2016; Fan et al., 2017; Liu et al., 2018; Fevry and Phang, 2018; Schumann, 2018; Makino et al., 2019). In general, multiple document summarization (Goldstein et al., 2000; Yasunaga et al., 2017; Zhao et al., 2020; Ma et al., 2020; Meena et al., 2014) is more challenging than single document summarization. However, the *SOS* task is different from traditional multi-document summarization tasks in that the goal here is to summarize content with an *overlap* constraint, i.e., the output should only contain the common information from both input narratives.

Alternatively, one could aim to recover verb-predicate alignment structure (Roth and Frank, 2012; Xie et al., 2008; Wolfe et al., 2013) from a sentence and further, use this structure to compute the overlapping information (Wang and Zhang, 2009; Shibata and Kurohashi, 2012). Sentence Fusion is another related area that aims to combine the information from two given sentences with some additional constraints (Barzilay et al., 1999; Marsi and Krahmer, 2005; Krahmer et al., 2008; Thadani and McKeown, 2011). A related but simpler task is to retrieve parallel sentences (Cardon and Grabar, 2019; Nie et al., 1999; Murdock and Croft, 2005) without performing an actual overlap summary generation. However, these approaches are more targeted toward individual sentences and do not directly translate to arbitrarily long documents. Thus, the *SOS* task is still an open problem and there is no existing dataset, method, or evaluation metric that has been systematically studied.

Along the evaluation dimension, *ROUGE* (Lin, 2004) is perhaps the most commonly used metric today for evaluating automated summarization techniques; due to its simplicity and automation. However, *ROUGE* has been criticized a lot for primarily relying on lexical overlap (Nenkova, 2006; Zhou et al., 2006; Cohan and Goharian, 2016; Akter et al., 2022) of n-grams. As of today, around 192 variants of *ROUGE* are available (Graham, 2015) including *ROUGE* with word embedding (Ng and Abrecht, 2015) and synonym (Ganesan, 2018), graph-based lexical measurement (ShafieiBavani et al., 2018), Vanilla *ROUGE* (Yang et al., 2018) and highlight-based *ROUGE* (Hardy et al., 2019). However, there has been no study yet as to whether the *ROUGE* metric is appropriate for evaluating the *SOS* task, which is one of the central goals of our work.

3 Motivation and Applications

Multiple alternative narratives are frequent in a variety of domains and, therefore, have direct implications in technical areas such as Information Retrieval/Search Engines, Question Answering, Machine Translation, etc. Below are a few examples of use cases.

Peer-Reviewing: Given two peer-review narratives for an article, the *SOS* task can generate a summary of portions of the narratives that agree with each other, which can help prepare a meta-review quickly.

Security and Privacy: *SOS* task can enable real-world users to quickly conduct a comparative analysis of multiple privacy policies by mining and summarizing overlapping clauses from those policies. Thus, it can help users make informed decisions while choosing from various alternative web services.

Health Sector: *SOS* can be used to compare clinical notes in patient records to perform a comparative analysis of patients with the same diagnosis/treatment. For example, *SOS* can be applied to the clinical notes of two (or more) patients who went through the same treatments to assess the effectiveness of the treatment.

Military Intelligence: If A and B are two intelligence reports related to a mission coming from two human agents, the *SOS* task can help cross-verify the claims in each report w.r.t. the other, thus, *SOS* can be used as an automated claim-verification tool.

Computational Social Science and Journalism: Assume that two news agencies (with different political biases) are reporting the same real-world event and their bias is somewhat reflected in the articles they write. If A and B are two such news articles, then the *SOS* output will likely surface the facts (common information) about the event.

Below are some of the potential applications of the *SOS* task in various technical areas.

Information Retrieval/Search Engines: One could summarize the common information in the multiple results fetched by a search engine for a given query and show it in a separate box to the user. This would immensely help them to quickly parse the desired information rather than going through each article individually. If they wish, they could further explore the specific articles for more details.

Question Answering: One could apply *SOS* to summarize the common information/answer from multiple documents pertinent to a given question. This will help formulate a more accurate answer by consulting multiple sources.

Robust Translation: Suppose you have multiple machine translator models which translate a given document from language A to language B. One could further apply the *SOS* to different translated documents and get a more robust translation by summarizing their semantic overlap.

In general, SOS task could be employed in any setting where we require comparative text analysis.

4 Problem Formulation

What is *Semantic Overlap Summarization*? This is indeed an open question and there is no single

AllSides Dataset: Statistics				
Split	#words (per docs)	#sents (per docs)	#words (per reference)	#sents (per reference)
Train	1613.69	66.70	67.30	2.82
Test	959.80	44.73	65.46/38.06/21.72/32.82	3.65/2.15/1.39/1.52

Table 1: Statistics for the Training and Testing dataset. Two input narratives are concatenated to compute the statistics. Four numbers for reference (#words/#sents) in the Test split correspond to the 4 reference overlap summaries. Our test dataset contains 137 samples, wherein each sample has 4 ground truth references. Out of these 4 references, *one* summary is provided by AllSides, and 3 of them were manually written by 3 human annotators. Thus, we generated 3*137 = 411 references in total.

correct answer. To simplify notations, let us stick to having only two documents D_A and D_B as our input since it can easily be generalized in case of more documents using *SOS* repeatedly. Also, let us define the output as $D_O \leftarrow D_A \cap_O D_B$. A human would mostly express the output in the form of natural language, and this is why we frame the *SOS* task as a constrained multi-seq-to-seq (text generation) task, where the output text only contains information that is present in both the input documents. We also argue that brevity (minimal repetition) is a desired property of *Semantic Overlap Summarization*. For example, if a particular piece of information or quote is repeated twice in both the documents, we don't necessarily want it to be present in output overlap summary two times. The output can either be extractive summary or abstractive summary or a mixture of both, as per the use case. This task is inspired by the set-like intersection operator as envisioned by (Karmaker Santu et al., 2018a) and the aim of this work is to summarize the overlapping information in an abstract fashion. Additionally, *SOS* should follow the *commutative* property, i.e. $D_A \cap_O D_B = D_B \cap_O D_A$.

5 The Benchmark Dataset

As mentioned in section 1, there is no existing dataset that we could readily use to evaluate the *SOS* task[1]. To address this challenge, we collected data from AllSides.com. AllSides is a third-party online news forum that exposes people to news and information from all sides of the political spectrum so that the general people can get an "unbiased" view of the world. To achieve this, AllSides displays each day's top news stories from news media widely-known to be affiliated with different sides of the political spectrum including "Left" (e.g., New York Times, NBC News), and "Right" (e.g., Townhall, Fox News) wing media. AllSides also provides its own *factual* description of the reading material, labelled as "Theme" so that readers can see the so-called "neutral" point-of-view. Table 1 gives an overview of the dataset statistics created by crawling from AllSides.com, which consists of news articles (from at least one "Left" and one "Right" wing media) covering 2,925 events in total and also having a minimum length of "theme-description" to be 15 words. Given two narratives ("Left" and "Right"), we used the theme description as a proxy for ground-truth reference summaries. We divided this dataset into testing data (described next) and training data (remaining samples) [see Table 1]. Table 2 shows the different attributes of the same AllSides dataset.

Feature	Description
theme	headlines by AllSides
theme-description	news description by AllSides
right/left head	right/left news headline
right/left context	right/left news description

Table 2: Overview of dataset scraped from AllSides. AllSides is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Testing Dataset and Human Annotations[2]: We engaged human volunteers to thoroughly annotate our testing samples (narrative pairs) in order to create multiple reference overlap summaries for each pair. This helped in creating a comprehensive testing benchmark for more rigorous evaluation. Specifically, we randomly sampled 150 narrative pairs describing 150 unique events (each pair consists of one narrative from the "Left" wing and one

[1]Multi-document summarization datasets can not be utilized in this scenario as their reference summaries do not follow the semantic overlap constraint.

[2]Dataset and manual annotations can be found at: https://karmake2.github.io/publications/

Narrative Pair Example # 1

Narrative 1: N_1	Narrative 2: N_2
WASHINGTON – U.S. intelligence and law enforcement agencies have confirmed that President Donald Trump's campaign aides and associates had constant contact with Russian intelligence officials before the election, directly contradicting public statements made by top administration officials. On Jan. 15, shortly before Trump took office, Vice President Mike Pence repeatedly said on television that there were zero contacts between the campaign and Russian officials. . . . Pence also answered "of course not" when asked a similar question that day by "Fox News Sunday" host Chris Wallace . . . Trump himself also denied these interactions . . . "There's nothing that would conclude me that anything different has changed with respect to that time period," Spicer said. . . .	President Trump said Wednesday that new reports saying his associates had contact with Russian officials during last year's campaign are "non-sense" and accused the U.S. intelligence community of illegally leaking information to news outlets. "This Russian connection non-sense is merely an attempt to cover-up the many mistakes made in Hillary Clinton's losing campaign," Mr. Trump tweeted. . . . Among those supposedly communicating with Russian nationals was former Trump campaign chairman Paul Manafort, the report said. Mr. Manafort denied that he ever knowingly talked to any intelligence official "or anyone

Reference Overlap Summaries

A_1	A_2	A_3	AllSides
President Trump and the Trump administration deny allegations that advisers close to Trump were in constant communication during the campaign with Russians known to US intelligence.	Trump denied climas that advisers close to him were in "constant communication during the campaign with Russians known to US intelligence.	Donald Trump and his group claimed that there is no contact with Russian officials during his last year's campaign.	Russian intelligence officials made repeated contact with members of President Trump's campaign staff, according to new reports that cite anonymous U.S. officials. American agencies were concerned about the contacts but haven't seen proof of collusion between the campaign and the Russian security apparatus.

Narrative Pair Example # 2

Narrative 1: N_1	Narrative 2: N_2
John McCain is out of McConnell's clutches for a week or two. While Sen. John McCain remains in Arizona recovering from Friday's craniotomy, surgery to remove a 5 cm blood clot from above his left eye, business will not go on as usual in Washington. Majority Leader Mitch McConnell, who has to have every Republican senator voting to have a prayer of passing Trumpcare, has postponed the vote for the week or two (more likely two) that McCain's recovery will take. That means there's more time for opponents to fight this thing, from the side of all of us trying to keep 22 million people from losing insurance and from the other side. . . . With both Paul and Sen. Susan Collins (R-ME) solid "no" votes on the bill, opponents only need one more out of the eight or so who've expressed reservations about the bill and the secretive, exclusive process McConnell	WASHINGTON - The Republican effort to repeal and replace Obamacare faces a major setback as Sen. John McCain, R-Ariz., left the nation's capital for surgery on his eye. Over the weekend, Senate Majority Leader Mitch McConnell, R-Ky., announced the scheduled Better Care Act vote would be delayed indefinitely because of McCain's absence. Subsequently, the Congressional Budget Office (CBO) also delayed its analysis of the bill. With two Republican senators opposed to the measure, McConnell needs as least 50 "yes" votes to pass it. Sen. Rand Paul, R-Ky., says the bill, which keeps taxes on investments and other pieces of Obamacare, doesn't go far enough. Moderate Sen. Susan Collins, R-Maine, is also withholding her support because it would slow the rate of growth in spending on Medicaid. . . .

Reference Overlap Summaries

A_1	A_2	A_3	AllSides
Sen. John McCain remains in Arizona recovering from eye surgery. Senate Majority Leader Mitch McConnell postponed the vote due to McCain's absence. Two Republican senators opposed to the bill. Possibility of bill failing.	Sen. John McCain remains unavailable because of the surgery on his eye. Senate Majority Leader Mitch McConnell delayed the vote in his absence. Sen. Rand Paul and Sen. Susan Collins said "no" votes on the bill.	Senate Majority Leader Mitch McConnell, R-Ky., announced the scheduled health care vote would be delayed indefinitely because of McCain's absence.	Senate Majority Leader Mitch McConnell, R-Ky., announced the scheduled Better Care Act vote would be delayed indefinitely because of McCain's absence.

Table 3: Some examples of *SOS* references from 3 human annotators (**A_i**) and the AllSides "theme-description" for a given narrative pair $\{N_1, N_2\}$. $(\ldots)$ denotes some sentences which for not shown for brevity. More examples can be found over here. Having multiple human annotators is critical to perform robust evaluation, but it is laborious and time-consuming on humans' part. This also shows that the lack of available datasets is a huge challenge for the SOS task.

from the "Right" wing, thus 300 narratives in total) and then asked 3 humans to write a summary of common information present in both narratives describing each of the 150 events.

After the first round of annotations, we immediately observed a discrepancy among the three annotators in terms of the *real* definition of "common/overlapping information". For example, one annotator argued that the reference summary should be non-empty as long as there is an overlap between two narratives along at least one of the *5W1H* facets (Who, What, When, Where, Why, and How), while another annotator argued that overlap in only one facet is not enough to decide whether there is indeed enough semantic overlap between the two narratives and reference summary should be left empty in such cases. As an example, one of the annotators wrote only "Donald Trump" as the reference summary for a couple of cases where the actual narratives were substantially different except for "Donald Trump" being the only common entity, while others had those cases marked as "*empty*".

To mitigate this issue, we only retained the narrative pairs where at least two of the annotators wrote a minimum of 15 words as their reference summaries, assuming that a human-written summary will contain 15 words or more only in cases where there is indeed a *significant* overlap between the two original narratives. This filtering step gave us a test set with 137 narrative pairs where each sample had 4 reference summaries, *one* from AllSides and *three* from human annotators, resulting in a total of 548 reference summaries. A couple of sample narrative pairs are shown in Table 3 along with the human annotations.

6 Evaluating SOS Task using ROUGE

As *ROUGE* (Lin, 2004) is the most popular metric used today for evaluating summarization tasks, we first conducted a case study with *ROUGE* as the evaluation metric for the *SOS* task. For methods, we experimented with multiple SoTA pretrained abstractive summarization models as *naive baselines* for *Semantic-Overlap Summarizer (SOS)*. These models are: 1) **BART** (Lewis et al., 2019), fine-tuned on CNN and multi-English Wiki news datasets, 2) **Pegasus** (Zhang et al., 2019), fine-tuned on CNN and Daily Mail dataset, and 3) **T5** (Raffel et al., 2019), fine-tuned on multi-English Wiki news dataset. As our primary goal is to construct a benchmark dataset for the *SOS* task and explore how to accurately evaluate this task, experimenting with only 3 abstractive summarization models is not a barrier to our work. Proposing a custom method fine-tuned for the *Semantic-Overlap* task is an orthogonal goal to this work and we leave it as future work. Also, we shall use the phrases "summary" and "overlap-summary" interchangeably from here. To generate the summary, we concatenate a narrative pair and feed it directly to the model.

For evaluation, we first evaluated the machine-generated overlap summaries for the 137 manually annotated testing samples using the ROUGE metric by following the procedure mentioned in Lin (2004) to compute the ROUGE-F_1 scores against multiple reference summaries. More precisely, since we have 4 reference summaries, we got 4 precision, recall pairs which are used to compute the corresponding F_1 scores. For each sample, we took the max of these four F_1 scores and averaged them out across the test dataset (see Table 4).

Model	R1	R2	RL
BART	40.73	25.97	29.95
T5	38.50	24.63	27.73
Pegasus	46.36	29.12	37.41

Table 4: Average ROUGE-F_1 Scores for all the test models across test dataset. For a particular sample, we take the maximum value out of the 4 F_1 scores corresponding to the 4 reference summaries.

Implementation Details: For generating summaries, we used off-the-shelf models in our experiments with default settings for summarization task following the Huggingface repo. Apart from this, we set the min and max length parameters to 10 and 300, respectively, based on our dataset. All the models are publicly available with details of the source. For ROUGE computation, we followed the implementation from the HuggingFace repo with the following parameters: $\{use\_stemmer = True, bootstrap\_aggregation = False\}$. Apart from this, we just used a sentence tokenizer from nltk library with English to create the input tokens. So, most of the method and ROUGE implementations are already publicly available. As such, there was no training involved in our experiments, but we still made use of the GPU (NVIDIA Quadro RTX 5000 with 16 GB of memory) to generate summaries using these models. Table 5 shows the summarization models and the number of parameters used in our experiments.

Model	#Parameters
BART	$\sim$ 406 M
T5	$\sim$ 223 M
Pegasus	$\sim$ 571 M

Table 5: Models and their corresponding number of parameters used in our experiments.

Results and Findings: We computed Pearson's correlation coefficients (using the scipy package) between each pair of ROUGE-F_1 scores obtained using all of the 4 reference overlap summaries (3

Pearson's Correlation Coefficients									
	R1			**R2**			**RL**		
	I_1	I_2	I_3	I_1	I_2	I_3	I_1	I_2	I_3
I_2	$\mathbf{0.62}$	—		$\mathbf{0.65}$	—		$\mathbf{0.69}$	—	
I_3	$\mathbf{0.3}$	$\mathbf{0.38}$	—	$\mathbf{0.27}$	$\mathbf{0.37}$	—	$\mathbf{0.27}$	$\mathbf{0.44}$	—
I_4	$\mathbf{0.17}$	$\mathbf{0.34}$	$\mathbf{0.34}$	0.14	$\mathbf{0.33}$	$\mathbf{0.21}$	$\mathbf{0.18}$	$\mathbf{0.35}$	$\mathbf{0.33}$
Average		**0.36**			**0.33**			**0.38**	

Table 6: Max (across 3 models) Pearson's correlation between the F_1 ROUGE scores corresponding to different annotators. Here I_i refers to the i^{th} annotator where $i \in \{1, 2, 3, 4\}$ and the "Average" row represents the average correlation of the max values across annotators. Boldface values are statistically significant at p-value < 0.05. For 5 out of 6 annotator pairs, the correlation values are quite small (≤ 0.50), thus, implying the poor inter-rated agreement with regards to the ROUGE metric.

human written summaries and 1 AllSides theme description) to test the robustness of *ROUGE* metric for evaluating the *SOS* task. The corresponding correlations are shown in table 6. For each annotator pair, we report their maximum (across 3 models) correlation value. The average correlation value across annotators is 0.36, 0.33 and 0.38 for R1, R2 and RL, respectively, suggesting that the ROUGE metric demonstrates high variance across multiple human-written overlap-summaries and thus, *unreliable*.

7 Can We Do Better than ROUGE?

Section 6 shows that the ROUGE metric is unstable across multiple reference overlap-summaries. Therefore, an immediate question is: Can we come up with a better metric than ROUGE? To investigate this question, we started by manually assessing the machine-generated overlap summaries to check first whether humans agree among themselves or not, i.e., whether human annotators can reach a consensus or not.

7.1 Different trials of Human Judgement

Assigning a Single Numeric Score: As an initial trial, we decided to first label 25 testing samples using two human annotators (we refer to them as label annotators, L_1 and L_2). Both label annotators read each of the 25 narrative pairs as well as the corresponding system-generated overlap summary (generated by fine-tuned BART) and assigned a numeric score between 1-10 (inclusive). This number reflects their judgment/confidence about how accurately the system-generated summary captures the *actual* overlap of the two input narratives. Note that, *the reference overlap summaries were not included in this label annotation process and the label-annotators judged the system-generated summary exclusively with respect to the input narratives*. To quantify the agreement between human scores, we computed the Kendall rank correlation coefficient (or Kendall's Tau) between two annotator labels since these are ordinal values. We used an open-source scipy package for computing Kendall's Tau correlation. However, to our disappointment, the correlation value was 0.20 with the p-value being 0.22[3]. This shows that even human annotators are disagreeing among themselves and we need to come up with a better labelling guideline to reach a reasonable agreement among the human annotators.

On further discussions among annotators, we realized that one annotator only focused on the *precision* of the output overlap summaries, whereas the other annotator took both *precision* and *recall* into consideration. Therefore, subsequently, we decided to assign two separate scores for precision and recall.

Precision-Recall Inspired Double Scoring: This time, three label annotators (L_1, L_2 and L_3) assigned two numeric scores between 1-10 (inclusive) for the same set of 25 system-generated summaries. These numbers represented their belief about how precise the system-generated summaries were (the precision score) and how much of the actual ground-truth overlap information was covered by the same (the recall score). Also, note that *labels were assigned exclusively with respect to the input narratives only*. As the assigned numbers represent ordinal values (i.e. can't be directly used to com-

[3]The higher p-value means that the correlation value is insignificant because of the small number of samples.

Human agreement in terms of Kendall's Tau for Double Scoring				
	Precision		Recall	
	L_1	L_2	L_1	L_2
L_2	0.52	—	0.37	—
L_3	0.18	0.29	0.31	0.54
Average	**0.33**		**0.41**	

Table 7: Kendall's rank correlation coefficients among the precision and recall scores for pairs of human annotators (25 samples). L_i refers to the i^{th} label annotator.

pute the F_1 score), we computed Kendall's rank correlation coefficient among the precision scores and recall scores separately for all the annotator pairs. The corresponding correlation values can be seen in table 7. As we notice, there is definitely some improvement in agreement among annotators compared to the one-number annotation in section 7.1. However, the average correlation is still 0.33 and 0.41 for precision and recall, respectively, much lower than 0.5 (the random baseline).

Human agreement in terms of Kendall's Tau Sentence-wise Scoring				
	Precision		Recall	
	L_1	L_2	L_1	L_2
L_2	0.68	—	0.75	—
L_3	0.59	0.64	0.69	0.71
Average	**0.64**		**0.72**	

Table 8: Average precision and recall Kendall rank correlation coefficients between sentence-wise annotation for different annotators. L_i refers to the i^{th} label annotator. All values are statistically significant (p<0.05).

7.2 Sentence-wise Scoring

From the previous trials, we realized the downsides of assigning one/two numeric scores to judge an entire system-generated overlap summary. Therefore, as a next step, we decided to assign *overlap labels* (defined below) to each sentence within the system-generated overlap summary and use those labels to compute the overall precision and recall.

Overlap Labels: Label annotators (L_1, L_2 and L_3) were asked to look at each machine-generated sentence separately and determine if the core information conveyed by it is absent (A), partially present (PP) or present (P) in any of the four reference summaries (provided by I_1, I_2, I_3 and I_4) and respectively, assign the label *A*, *PP* or *P*. More precisely, annotators were provided with the following instructions: if the human feels that there is more than 75% overlap (between each system-generated sentence and any reference-summary sentence), assign label *P*, else if the human feels there is less than 25% overlap, assign label *A*, otherwise, assign label *PP*. This sentence-wise labelling was done for 50 different samples (with 506 sentences in total for system and reference summary), which resulted in a total of $3 \times 506 = 1,518$ sentence-level ground-truth labels.

To create the overlap labels (*A*, *PP* or *P*) for precision, we concatenated all 4 reference summaries to make one big reference summary and asked label-annotators (L_1, L_2, and L_3) to use it as a single reference for assigning the overlap labels to each sentence within machine generated summary. We argue that if the system could generate a sentence conveying information that is present in any of the references, it should be considered a hit. For recall, label-annotators were asked to assign labels to each sentence in each of the 4 reference summaries separately (provided by (I_1, I_2, I_3 and I_4)), with respect to the machine summary.

Inter-Rater-Agreement: After annotating each system-generated sentence (for precision) and reference sentence (for recall) with the labels (*A*, *PP* or *P*), we used the Kendall rank correlation coefficient to compute the pairwise annotator agreements among these ordinal labels. Table 8 shows that the correlations for both precision and recall are ≥ 0.50, signifying higher inter-annotator agreement.

Label from Annotator B		P	PP	A
Label from Annotator A	P	1	0.5	0
	PP	0.5	1	0
	A	0	0	1

Table 9: Reward matrix used to compare the labels assigned by two label annotators for a given sentence and helps to compute the agreement between the annotator pairs.

Reward-based Inter-Rater-Agreement: Alternatively, we defined a reward matrix (Table 9) which is used to compare the label of one annotator (say annotator A) against the label of another annotator (say annotator B) for a given sentence. This reward matrix acts as a form of correlation between two

Human agreement in terms of Reward function				
	Precision		**Recall**	
	L_1	L_2	L_1	L_2
L_2	0.81 ± 0.26	—	0.85 ± 0.11	—
L_3	0.79 ± 0.26	0.70 ± 0.31	0.80 ± 0.16	0.77 ± 0.17
Average	**0.77**		**0.81**	

Table 10: Average precision and recall reward scores (mean $\pm$ std) between sentence-wise annotation for different annotators. L_i refers to the i^{th} label-annotator.

annotators. Once the reward has been computed for each sentence, one can compute the average precision and recall rewards for a given sample and accordingly, for the entire test dataset. The corresponding reward scores can be seen in Table 10. Both precision and recall reward scores are high (≥ 0.70) for all the different annotator pairs, thus signifying, a high inter-label-annotator agreement.

We believe, one of the reasons for higher reward/Kendall scores could be that sentence-wise labelling puts a lesser cognitive load on the human mind allowing them to be more consistent in contrast to the single or double score(s) for the entire overlap summary and, therefore, shows high agreement in terms of human interpretation. A similar observation was noted in Harman and Over (2004).

8 Limitations and Future Work

One particular limitation of this work is that we have used pre-trained abstractive summarization models as *naive baselines* / proxies for semantic overlap summarizer and did not attempt to develop a custom method that optimizes for the *overlap* constraint. However, the primary focus of this paper is to define the *SOS* task, as well as establish the first benchmark dataset and a suitable evaluation approach for the task. Therefore, the design and optimization of methods is an orthogonal task to this paper, which we will pursue as our immediate future work.

Another limitation of our work is that the test set is not big (~ 150 examples), which makes it difficult to do a rigorous evaluation. However, while the number 150 may initially appear to be small; cleaning and annotating the dataset required significant time and resources. To elaborate further, our test dataset contains 137 samples, where each sample consists of two alternative narratives along with 4 ground truth references. Out of these 4 references, 3 of them were manually written by 3 human annotators. Thus, we manually created $3 * 137 = 411$ reference summaries in total. Additionally, for each sample (narrative pair), each annotator first had to carefully read through two alternative narratives several times, digest the semantic overlap between them and then summarize the overlap in their own words. This process took on average 40 minutes per annotator per sample, which means we spent around $411 * 40 = 16,440$ minutes of human efforts in one round of the annotation process. Next, we had to resolve conflicts among annotators by going through each of their annotated summaries (a very painstaking process) and figuring out the reasons for the conflicts. Based on follow-up discussions, we revised the guidelines for annotation and went through the entire annotation process again. In total, we needed 4 iterations ($16,440 * 4 =\sim 65,760$ minutes) to resolve most of the conflicts. The whole process took more than 8 months for us. Finally, we agree that having more samples in the test dataset would definitely help. But this is both time and money-consuming. We are working towards it and would like to increase the sample size in the future.

9 Conclusion

In this work, we introduced a new NLP task, called Semantic Overlap Summarization (*SOS*) and created a benchmark dataset through meticulous human efforts to initiate a new research direction. As a starting point, we framed the problem as a constrained summarization task and showed that *ROUGE* is not a reliable evaluation metric for this task. Further human experiments show that sentence-wise evaluation leads to higher agreement with human judgment, therefore, an evaluation metric that aggregates sentence-wise overlap labels should be used while evaluating the SOS task.

10 Acknowledgements

We thank Auburn University College of Engineering and the Department of CSSE for their continuous support through Student Fellowships and Faculty Startup Grants.

References

Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker. 2022. Revisiting automatic evaluation of extractive summarization task: Can we do better than rouge? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1547–1560.

Otmane Azeroual and Horst Theel. 2018. The effects of using business intelligence systems on an excellence management and decision-making process by start-up companies: A case study. *International Journal of Management Science and Business Administration*, 4(3):30–40.

Sanghwan Bae, Taeuk Kim, Jihoon Kim, and Sang-goo Lee. 2019. Summary level training of sentence rewriting for abstractive summarization. *arXiv preprint arXiv:1909.08752*.

Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 550–557.

Biddut Sarker Bijoy, Syeda Jannatus Saba, Souvika Sarkar, Md Saiful Islam, Sheikh Rabiul Islam, Mohammad Ruhul Amin, and Shubhra Kanti Karmaker Santu. 2021. Covid19α: Interactive spatio-temporal visualization of covid-19 symptoms through tweet analysis. In *26th International Conference on Intelligent User Interfaces-Companion*, pages 28–30.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–161, Melbourne, Australia. Association for Computational Linguistics.

Rémi Cardon and Natalia Grabar. 2019. Parallel sentence retrieval from comparable corpora for biomedical text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 168–177.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

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.

Angela Fan, David Grangier, and Michael Auli. 2017. Controllable abstractive summarization. *arXiv preprint arXiv:1711.05217*.

Thibault Fevry and Jason Phang. 2018. Unsupervised sentence compression using denoising auto-encoders. *arXiv preprint arXiv:1809.02669*.

Kavita Ganesan. 2018. ROUGE 2.0: Updated and improved measures for evaluation of summarization tasks. *CoRR*, abs/1803.01937.

Jade Goldstein, Vibhu O Mittal, Jaime G Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*.

Yvette Graham. 2015. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 128–137. The Association for Computational Linguistics.

Hardy, Shashi Narayan, and Andreas Vlachos. 2019. Highres: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3381–3392. Association for Computational Linguistics.

Donna Harman and Paul Over. 2004. The effects of human variation in DUC summarization evaluation. In *Text Summarization Branches Out*, pages 10–17, Barcelona, Spain. Association for Computational Linguistics.

Naeemul Hassan, Amrit Poudel, Jason Hale, Claire Hubacek, Khandaker Tasnim Huq, Shubhra Kanti Karmaker Santu, and Syed Ishtiaque Ahmed. 2020. Towards automated sexual violence report tracking. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 250–259.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. *arXiv preprint arXiv:1805.06266*.

Shubhra Kanti Karmaker Santu, Chase Geigle, Duncan Ferguson, William Cope, Mary Kalantzis, Duane Searsmith, and Chengxiang Zhai. 2018a. Sofsat: Towards a setlike operator based framework for semantic analysis of text. *ACM SIGKDD Explorations Newsletter*, 20(2):21–30.

Shubhra Kanti Karmaker Santu, Liangda Li, Yi Chang, and ChengXiang Zhai. 2018b. Jim: Joint influence modeling for collective search behavior. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 637–646.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. *arXiv preprint arXiv:1609.09552*.

Emiel Krahmer, Erwin Marsi, and Paul van Pelt. 2008. Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion. In *Proceedings of ACL-08: HLT, Short Papers*, pages 193–196.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, and Hongyan Li. 2017. Generative adversarial network for abstractive text summarization. *arXiv preprint arXiv:1711.09357*.

Yizhu Liu, Zhiyi Luo, and Kenny Zhu. 2018. Controlling length in abstractive summarization using a convolutional neural network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4110–4119.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2020. Multi-document summarization via deep learning techniques: A survey. *arXiv preprint arXiv:2011.04843*.

Takuya Makino, Tomoya Iwakura, Hiroya Takamura, and Manabu Okumura. 2019. Global optimization under length constraint for neural text summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1039–1048.

Erwin Marsi and Emiel Krahmer. 2005. Explorations in sentence fusion. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.

Yogesh Kumar Meena, Ashish Jain, and Dinesh Gopalani. 2014. Survey on graph and cluster based approaches in multi-document text summarization. In *International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014)*, pages 1–5. IEEE.

Vanessa Murdock and W Bruce Croft. 2005. A translation model for sentence retrieval. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 684–691.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. *arXiv preprint arXiv:1802.08636*.

Ani Nenkova. 2006. Summarization evaluation for text and speech: issues and approaches. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing*. ISCA.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1925–1930. The Association for Computational Linguistics.

Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. 1999. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 74–81.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. 2019. Better language models and their implications. *OpenAI Blog https://openai.com/blog/better-language-models*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Michael Roth and Anette Frank. 2012. Aligning predicate argument structures in monolingual comparable texts: A new corpus for a new task. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 218–227.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Raphael Schumann. 2018. Unsupervised abstractive sentence summarization using length controlled variational autoencoder. *arXiv preprint arXiv:1809.05233*.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond K. Wong, and Fang Chen. 2018. A graph-theoretic summary evaluation for rouge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 762–767. Association for Computational Linguistics.

Tomohide Shibata and Sadao Kurohashi. 2012. Predicate-argument structure-based textual entailment recognition system exploiting wide-coverage lexical knowledge. *ACM Transactions on Asian Language Information Processing (TALIP)*, 11(4):1–23.

Swapna Somasundaran, Michael Flor, Martin Chodorow, Hillary Molloy, Binod Gyawali, and Laura McCulla. 2018. Towards evaluating narrative quality in student writing. *Transactions of the Association for Computational Linguistics*, 6:91–106.

Kapil Thadani and Kathleen McKeown. 2011. Towards strict sentence intersection: decoding and evaluation strategies. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 43–53.

Rui Wang and Yi Zhang. 2009. Recognizing textual relatedness with predicate-argument structures. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 784–792.

Shomir Wilson, Florian Schaub, Frederick Liu, Kanthashree Mysore Sathyendra, Daniel Smullen, Sebastian Zimmeck, Rohan Ramanath, Peter Story, Fei Liu, Norman Sadeh, et al. 2018. Analyzing privacy policies at scale: From crowdsourcing to automated annotations. *ACM Transactions on the Web (TWEB)*, 13(1):1–29.

Travis Wolfe, Benjamin Van Durme, Mark Dredze, Nicholas Andrews, Charley Beller, Chris Callison-Burch, Jay DeYoung, Justin Snyder, Jonathan Weese, Tan Xu, et al. 2013. Parma: A predicate argument aligner. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 63–68.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. *arXiv preprint arXiv:1804.07036*.

Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. *arXiv preprint arXiv:2001.11314*.

Lexing Xie, Hari Sundaram, and Murray Campbell. 2008. Event mining in multimedia streams. *Proceedings of the IEEE*, 96(4):623–647.

Yu Yan, Weizhen Qi, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*.

An Yang, Kai Liu, Jing Liu, Yajuan Lyu, and Sujian Li. 2018. Adaptations of ROUGE and BLEU to better evaluate machine reading comprehension task. In *Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, Melbourne, Australia, July 19, 2018*, pages 98–104. Association for Computational Linguistics.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. *arXiv preprint arXiv:1706.06681*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. Summpip: Unsupervised multi-document summarization with sentence graph compression. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1949–1952.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. *arXiv preprint arXiv:1907.03491*.

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard H. Hovy. 2006. Paraeval: Using paraphrases to evaluate summaries automatically. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. The Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. *arXiv preprint arXiv:1704.07073*.