REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
31-07-2023	Final Report	30-Sep-2022 - 29-Jun-2023

4. TITLE AND SUBTITLE

Final Report: Information Systems: Computing Sciences: Knowledge Systems: Intel-Miner: Semantic Machine for Robust Interpretation of Noisy Intelligence

5a. CONTRACT NUMBER
W911NF-22-1-0280

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES

Auburn University
310 Samford Hall

Auburn, AL 36849 -5131

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
79475-MI-II.4

12. DISTRIBUTION AVAILIBILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Shubhra Karmaker
UU	UU	UU	UU		19b. TELEPHONE NUMBER: 334-844-4330

RPPR Final Report
as of 03-Aug-2023

Agency Code: 21XD

Proposal Number: 79475MIII **Agreement Number: W911NF-22-1-0280**
INVESTIGATOR(S):

Name: Ph.D. Shubhra Kanti Karmaker
Email: sks0086@auburn.edu
Phone Number: 3348444330
Principal: Y

Organization: **Auburn University**
Address: 310 Samford Hall, Auburn, AL 368495131
Country: USA
DUNS Number: 066470972 EIN: 636000724
Report Date: 29-Sep-2023 Date Received: 31-Jul-2023
Final Report for Period Beginning 30-Sep-2022 and Ending 29-Jun-2023
Title: Information Systems: Computing Sciences: Knowledge Systems: Intel-Miner: Semantic Machine for Robust Interpretation of Noisy Intelligence
Begin Performance Period: 30-Sep-2022 **End Performance Period:** 29-Jun-2023
Report Term: 0-Other
Submitted By: Ph.D. Shubhra Karmaker Email: sks0086@auburn.edu
Phone: (334) 844-4330
Distribution Statement: 1-Approved for public release; distribution is unlimited.

STEM Degrees: 1 **STEM Participants:** 3

Major Goals: Multi-Perspective Narratives (MPN) are ubiquitous, useful tools for verifying facts from different alternative narratives; thus, MPNs facilitate more informed decisions by offering a concise overall picture of the current situation. Assimilation and digestion of such MPNs at a large scale pose significant challenges for users, making it extremely arduous to filter out reliable information from these MPNs. Despite great progress made in natural language processing (NLP), computers are still far from being able to accurately analyze multi-perspective narratives and effectively summarize the common information/overlapping content they provide. In this project, we will design and develop solutions for this crucial yet relatively unexplored NLP task, which we formally refer to as Semantic Overlap Summarization (SOS). SOS entails generating a single summary from multiple alternative narratives that can convey the common information provided by those narratives. Next, we will design and implement deep self-supervised sequence-to-sequence learning techniques to solve the SOS task. The reason for choosing a such an approach is that models can be trained in a mostly unsupervised fashion without requiring a large amount of labeled training data. For rigorous evaluation, we will create and annotate a benchmark data set with at least 150 narrative pairs and simultaneously develop appropriate evaluation metrics for the SOS task by levering the semantically powerful sentence encoders.

Major Goal 1: Benchmark Dataset Creation

Although the proposed SOS task is closely related to multi-document summarization (MDS), it is different from traditional MDS tasks in that the goal is to summarize content with an additional constraint: the overlap criteria (i.e., the output should only contain the common information from both input narratives). Because (1) there is no existing dataset that we can readily use to evaluate the SOS task and (2) Multi-document summarization datasets cannot be utilized in this scenario since their reference summaries do not follow the semantic overlap constraint, our first objective is to create a benchmark dataset with gold reference summaries} to be able to evaluate this task rigorously. Without loss of generality, we will consider exactly two narratives as inputs for the SOS task. Goal 1 will embrace three sub-tasks:

RPPR Final Report
as of 03-Aug-2023

1. Create/Collect Multi-Perspective Narrative Data-set

2. Annotate Narrative Pairs with Human-Written Overlap Summaries

3. Conduct Meta-Evaluation of Overlap Summaries against Human Judgements

Major Goal 2: Overlap Summary Generation

Semantic Overlap Summarization (SOS) is a novel and relatively under-explored sequence-to-sequence task which entails summarizing common information from multiple alternate narratives. One of the major challenges for solving this task is the lack of existing datasets for supervised training. To address this challenge, we propose a novel data augmentation technique, which allows us to create large amount of synthetic data for training a sequence-to-sequence model that can perform the SOS task. This will create an artificial corpus that will facilitate self-supervised training.

Major Goal 3: Design Appropriate Evaluation Metric

When it comes to evaluation, we propose a new Semantic-F1 (SEM-F1) metric for accurate evaluation, which computes sentence-level information overlap to be eventually aggregated. The motivation for proposing this new evaluation metric is the findings in our preliminary study that (1) the popular ROUGE metric is unreliable for evaluation of the SOS task, while (2) our initial results with sentence-level overlap labels---A (Absent), PP (Partially Present), or P (Present)---yield higher inter-annotator agreement. Therefore, we propose the SEM-F1 metric which can infer these overlap labels automatically.

The proposed SEM-F1 is a precision-recall-style evaluation metric based on significant-overlap/partial-overlap/no-overlap between a pair of sentences in terms of their semantic meaning. To capture this, we define three levels: Present (P), Partial-Present (PP), and Absent (A) based on a threshold on the semantic similarity score, for each sentence in the generated narrative with respect to the entire reference narrative (to compute precision) and vice versa (to compute recall). Next, we will assign an overlap reward as follows: 1 for Present (P), 0.5 for Partial-Present (PP), and 0 for Absent (A). Finally, these reward scores will be averaged to deliver the precision/recall scores as well as the corresponding SEM-F1 score (simple harmonic mean) for the particular generated narrative. This process will continue for all the testing samples and be further averaged to compute the overall SEM-F1 score.

Evaluation Plan:

Evaluation of Goal 1 will be mainly focused on the meta-evaluation of annotated overlap summaries against human judgments using inter-rater agreement computed by Pearson's Correlation Coefficient. A high inter-rater agreement measure (>0.7) is considered a high-quality annotation, and the created dataset will be regarded as a good benchmark for evaluation. On the other hand, Goal 2 (Overlap Summary Generation) evaluation will be performed using the SEM-F1 (proposed by the PI, ROUGE and SARI metrics).

Accomplishments: In this project, we studied an important yet relatively unexplored NLP task called Semantic Overlap Summarization (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.

Next, we exclusively focused on the automated evaluation of the SOS task using the benchmark dataset. As our experiments discovered that ROUGE is not suitable for this novel task, therefore, we proposed a new sentence-level precision-recall style automated evaluation metric, called SEM-F1 (Semantic F1). It is inspired by the benefits of the sentence-wise annotation technique using overlap labels reported by the previous work. Our results showed that the proposed automated SEM-F1 metric yields a higher correlation with human judgment and higher inter-rater agreement compared to the ROUGE metric.

We realized that one of the major challenges associated with implementing a sequence-to-sequence model which can perform the SOS task is the lack of readily available training data for supervised learning. One may manually create a training corpus for a particular domain (e.g., news, health etc.) by spending a significant amount of time and money, yet it is unclear how much it will generalize for other domains. Therefore, an unsupervised approach is desired to address this problem. Therefore, we designed a new unsupervised data generation technique which can generate an arbitrarily large number of synthetic training examples for the SOS task. More specifically, given an arbitrary text corpus from a particular domain, our data generation algorithm can produce an infinite number of SOS examples of the form {D_A, D_B, (D_A \cap D_B)} where, D_A and D_B are two narratives (in text) and (D_A \cap D_B) is the desired reference summary of semantic overlap. Although the reference overlap summaries in our synthetic examples are noisy and do not ensure the high quality of human-written summaries, they can at least help us train an SOS model in a weakly supervised fashion and allow us to leverage the powerful yet data-hungry sequence-to-sequence deep learning architectures.

Noteworthy, our main focus was to propose an intelligent way to create a synthetic dataset for training existing sequence-to-sequence models for the SOS task rather than proposing a new model specifically customized for it. Therefore, finding the best model to solve the SOS task is an orthogonal goal to our work and hence, out of scope for this paper. Rather, the goal of this work is to leverage existing pre-trained sequence-to-sequence summarization models as an approximation of the overlap summary generator and create artificial examples to further fine-tune such sequence-to-sequence models. As such, it is important to validate whether fine-tuning with artificial examples are indeed useful for improving the accuracy of the sequence-to-sequence models. To achieve this, we used the human annotated and verified dataset from our previous work to show the efficacy of sequence-to-sequence models fine-tuned on our synthetic examples as compared to the pre-trained baseline models with no fine-tuning.

Through extensive experiments using narratives from the news domain, we showed that the models fine-tuned using our synthetic dataset provide significant performance improvements over the pre-trained-only baselines and are close to the models fine-tuned on the golden training data; which essentially demonstrates the effectiveness of the proposed data augmentation technique.

Training Opportunities: This award funded one Ph.D. student partially to work with the PI on everything described in the proposal, including the design, implementation, and validation of the proposed techniques, applications, and evaluation strategies.

RPPR Final Report
as of 03-Aug-2023

Results Dissemination: The following publications were presented at premier NLP conferenced during the reporting period:

1. Naman Bansal, Mousumi Akter, and Shubhra Kanti Karmaker Santu. 2022. Semantic Overlap Summarization among Multiple Alternative Narratives: An Exploratory Study. In Proceedings of the 29th International Conference on Computational Linguistics, pages 6195–6207, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

2. Naman Bansal, Mousumi Akter, and Shubhra Kanti Karmaker Santu. 2022. SEM-F1: an Automatic Way for Semantic Evaluation of Multi-Narrative Overlap Summaries at Scale. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 780–792, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

3. Naman Bansal, Mousumi Akter, and Shubhra Kanti Karmaker Santu. 2022. Learning to Generate Overlap Summaries through Noisy Synthetic Data. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11765–11777, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Honors and Awards: 1. Two of the PI's students, Naman Bansal and Mousumi Akter received Auburn University's Outstanding Doctoral student award.

2. Two of the PI's students, Souvika Sakar and Mousumi Akter were named 100+ Women Strong Outstanding Graduate Student Award recipients.

3. A team consisting of the PI and two of his students became Champion in the "Food for Thought" National NLP challenge hosted by Coleridge Initiative in collaboration with the US Department of Agriculture.

4. The PI was appointed as the Communication Chair of ACL Rolling Reviews (ARR) Initiative.

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Shubhra Kanti Karmaker Santu
Person Months Worked: 2.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Mousumi Akter
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Naman Bansal
Person Months Worked: 3.00 **Funding Support:**
Project Contribution:
National Academy Member: N

RPPR Final Report
as of 03-Aug-2023

CONFERENCE PAPERS:

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing
Date Received: 28-Jul-2023 Conference Date: 07-Dec-2022 Date Published: 08-Dec-2022
Conference Location: Abu Dhabi, United Arab Emirates
Paper Title: Learning to Generate Overlap Summaries through Noisy Synthetic Data
Authors: Bansal, Naman; Akter, Mousumi; Karmaker Santu, Shubhra Kanti
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing
Date Received: 28-Jul-2023 Conference Date: 07-Dec-2022 Date Published: 07-Dec-2022
Conference Location: Abu Dhabi, United Arab Emirates
Paper Title: SEM-F1: an Automatic Way for Semantic Evaluation of Multi-Narrative Overlap Summaries at Scale
Authors: Bansal, Naman; Akter, Mousumi; Karmaker Santu, Shubhra Kanti
Acknowledged Federal Support: **Y**

Publication Type: Conference Paper or Presentation **Publication Status:** 1-Published
Conference Name: Proceedings of the 29th International Conference on Computational Linguistics
Date Received: 28-Jul-2023 Conference Date: 12-Oct-2022 Date Published: 12-Oct-2022
Conference Location: Gyeongju, Republic of Korea
Paper Title: Semantic Overlap Summarization among Multiple Alternative Narratives: An Exploratory Study
Authors: Bansal, Naman; Akter, Mousumi; Karmaker Santu, Shubhra Kanti
Acknowledged Federal Support: **N**

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Shubhra Kanti Karmaker Santu
Signature Date: 7/31/23 4:57PM

Project Report

PI: Shubhra Kanti Karmaker ("Santu")
Big Data Intelligence (BDI) Lab
Department of Computer Science and Software Engineering
College of Engineering, Auburn University
sks0086@auburn.edu

July 31, 2023

Abstract

Multi-Perspective Narratives (MPN) are ubiquitous, useful tools for verifying facts from different alternative narratives; thus, MPNs facilitate more informed decisions by offering a concise overall picture of the current situation. Assimilation and digestion of such MPNs at a large scale pose significant challenges for users, making it extremely arduous to filter out reliable information from these MPNs. Despite great progress made in natural language processing (NLP), computers are still far from being able to accurately analyze multi-perspective narratives and effectively summarize the common information/overlapping content they provide. In this project, we have designed and developed solutions for this crucial yet relatively unexplored NLP task, which we formally refer to as Semantic Overlap Summarization (SOS). SOS entails generating a single summary from multiple alternative narratives that can convey the common information provided by those narratives. Next, we designed and implemented deep self-supervised sequence-to-sequence learning techniques to solve the SOS task. The reason for choosing such an approach is that models can be trained in a mostly unsupervised fashion without requiring a large amount of labeled training data. For rigorous evaluation, we will create and annotate a benchmark data set with at least 150 narrative pairs and simultaneously develop appropriate evaluation metrics for the SOS task by levering the semantically powerful sentence encoders.

1 Objectives of this Project

1.1 Problem Definition

For this project, 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 evaluate this task accurately. *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?

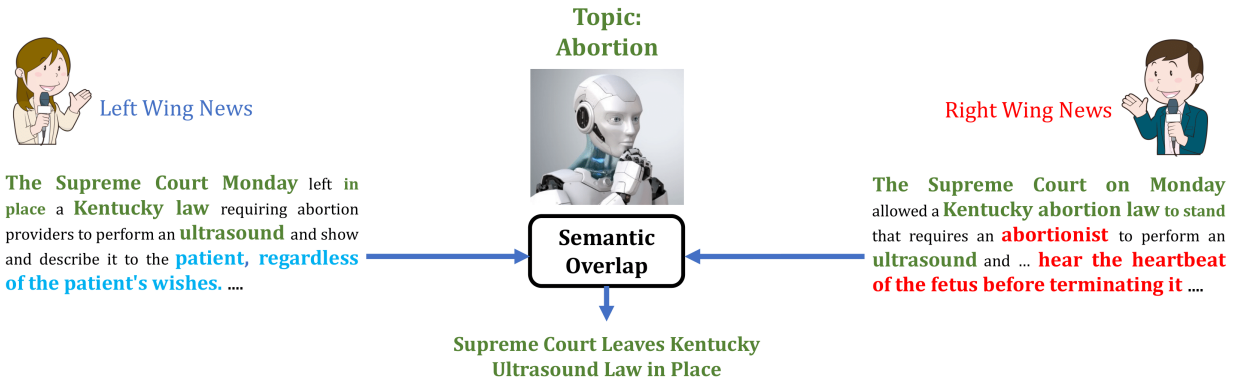

Figure 1: A toy example of *Semantic Overlap Summarization* (**SOS**) Task (from multiple alternative narratives). Here, an abortion issue-related event has been reported by two news media (left-wing and right-wing). "Green" Text denotes the common information from both news media, while "Blue" and "Red"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

Multiple alternative narratives appear frequently in a variety of domains, including education [1], the health sector [2], businesses intelligence [3], content analysis [4, 5] and privacy [6]. 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 evaluate this task properly. 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 convey 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 to create a benchmark dataset for this task?* To address these challenges, we aim to achieve the following goals and objectives in this project.

1.2 Objective 1: Benchmark Dataset Creation

Although the proposed SOS task is closely related to multi-document summarization (MDS) [7,8], it is different from traditional MDS tasks in that the goal is to summarize content with an additional constraint: the **overlap** criteria (i.e., the output should only contain the common information from both input narratives). Because (1) there is no existing dataset that we can readily use to evaluate the *SOS* task and (2) Multi-document summarization datasets cannot be utilized in this scenario since their reference summaries do not follow the semantic overlap constraint, our first objective is to **create a benchmark dataset with gold reference summaries** to be able to evaluate this task rigorously. Without loss of generality, we will consider exactly two narratives as inputs for the *SOS* task. Goal 1 will embrace three sub-tasks:

1. Create/Collect Multi-Perspective Narrative Data-set.
2. Annotate Narrative Pairs with Human-Written Overlap Summaries.
3. Conduct a Meta-Evaluation of Overlap Summaries against Human Judgements.

1.3 Objective 2: Design Appropriate Evaluation Metric

When it comes to evaluation, we propose a new Semantic-F1 (SEM-F1) metric for accurate evaluation, which computes sentence-level information overlap to be eventually aggregated. The motivation for proposing this new evaluation metric is the findings in our preliminary study that (1) the popular ROUGE metric is unreliable for evaluation of the SOS task, while (2) our initial results with sentence-level overlap labels—A (Absent), PP (Partially Present), or P (Present)—yield a higher inter-annotator agreement. Therefore, we propose the SEM-F1 metric, which can infer these overlap labels automatically.

The proposed SEM-F1 is a precision-recall-style evaluation metric based on significant overlap/partial overlap/no overlap between a pair of sentences in terms of their semantic meaning. To capture this, we define three levels: Present (P), Partial-Present (PP), and Absent (A) based on a threshold on the semantic similarity score for each sentence in the generated narrative with respect to the entire reference narrative (to compute precision) and vice versa (to compute recall). Next, we will assign an overlap reward as follows: 1 for Present (P), 0.5 for Partial-Present (PP), and 0 for Absent (A). Finally, these reward scores will be averaged to deliver the precision/recall scores as well as the corresponding SEM-F1 score (simple harmonic mean) for the particular generated narrative. This process will continue for all the testing samples and be further averaged to compute the overall SEM-F1 score.

1.4 Objective 3: Overlap Summary Generation

Semantic Overlap Summarization (SOS) is a novel and relatively under-explored sequence-to-sequence task which entails summarizing common information from multiple alternate narratives. One of the major challenges in solving this task is the lack of existing datasets for supervised training. To address this challenge, we propose a novel data augmentation technique, which allows us to create a large amount of synthetic data for training a sequence-to-sequence model that can perform the SOS task. This will create an artificial corpus that will facilitate self-supervised training.

1.5 Evaluation Goal

Evaluation of Goal 1 will be mainly focused on the meta-evaluation of annotated overlap summaries against human judgments using inter-rater agreement computed by Pearson's Correlation Coefficient. A high inter-rater agreement measure (>0.7) is considered a high-quality annotation, and the created dataset will be regarded as a good benchmark for evaluation. On the other hand, Goal 3 (Overlap Summary Generation) evaluation will be performed using the SEM-F1 (proposed by the PI, ROUGE, and SARI metrics).

2 Findings of this Project

2.1 Findings under Objective 1

2.1.1 Benchmark Dataset Creation

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, labeled 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.

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 $3 \times 137 = 411$ references in total.

Testing Dataset and Human Annotations[1]: We engaged human volunteers to thoroughly annotate our testing samples (narrative pairs) in order to create multiple reference overlap summaries

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.

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 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.

2.1.2 Evaluating SOS Task using ROUGE

As *ROUGE* [9] 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 pre-trained abstractive summarization models as *naive baselines* for *Semantic-Overlap Summarizer (SOS)*. These models are: 1) **BART** [10], fine-tuned on CNN and multi-English Wiki news datasets, 2) **Pegasus** [11], fine-tuned on CNN and Daily Mail dataset, and 3) **T5** [12], 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 [9] 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 3).

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 3: 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 4 shows the summarization models and the number of parameters used in our experiments.

Model	#Parameters
BART	~ 406 M
T5	~ 223 M
Pegasus	~ 571 M

Table 4: 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 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 5. 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*.

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	**0.62**	—		**0.65**	—		**0.69**	—	
I_3	**0.3**	**0.38**	—	**0.27**	**0.37**	—	**0.27**	**0.44**	—
I_4	**0.17**	**0.34**	**0.34**	0.14	**0.33**	**0.21**	**0.18**	**0.35**	**0.33**
Average		**0.36**			**0.33**			**0.38**	

Table 5: 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.

2.1.3 Creating a Benchmark Dataset with More Accurate References

As the ROUGE metric is unstable across multiple reference overlap summaries, 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.

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[2]. 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 sum-

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

maries. 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 compute 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 6. As we notice, there is definitely some improvement in agreement among annotators compared to the one-number annotation. 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 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 6: 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.

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.

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 7: 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 ($p<0.05$).

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 7 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 8: 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.

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 9: 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.

Reward-based Inter-Rater-Agreement: Alternatively, we defined a reward matrix (Table 8) 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 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 9. 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 [13].

2.2 Findings under Objective 2

2.2.1 Semantic-F1: an Automated Metric

Human evaluation is costly and time-consuming. Thus, one needs an automatic evaluation metric for large-scale experiments. But, how can we devise an automated metric to perform the sentence-wise precision-recall style evaluation discussed in the previous section? To achieve this, we propose a new evaluation metric called **SEM-F$_1$**. The details of our **SEM-F$_1$** metric are described in algorithm 1 and the respective notations are mentioned in table 10. F_1 scores are computed by the harmonic mean of the precision (pV) and recall (rV) values. Algorithm 1 assumes only one reference summary but can be trivially extended for multiple references. As mentioned previously, in the case of multiple references, we concatenate them for precision score computation. Recall scores are computed individually for each reference summary and later, an average recall is computed across references.

The basic intuition behind **SEM-F$_1$** is to compute the sentence-wise similarity (e.g., cosine similarity between two sentence embeddings) to infer the semantic overlap between a system-generated sentence and a reference sentence from both precision and recall perspectives and then, combine them into the F_1 score.

Notations	Description
S_G	Machines generated summary
S_R	Reference summary
$T \coloneqq (t_l, t_u)$	Tuple representing the lower and upper threshold values (between 0 and 1).
M_E	Sentence embedding model
pV, rV	Precision, Recall value for (S_G, S_R) pair

Table 10: Table of notations for algorithm 1

Reliability Testing The SEM-F_1 metric computes cosine similarity scores between sentence pairs from both precision and recall perspectives. To verify whether the SEM-F_1 metric correlates with human judgement, we further converted the sentence-wise cosine similarity scores into *Presence* (P), *Partial Presence* (PP) and *Absence* (A) labels using user-defined thresholds as described in algorithm 2. This helped us to directly compare the SEM-F_1 inferred labels against the human annotated labels.

We leveraged state-of-the-art sentence embedding models to encode sentences from both the model-generated summaries and the human-written reference summaries. To be more specific,

Algorithm 1 Semantic-F_1 Metric

Given S_G, S_R, M_E
$raw_{pV}, raw_{rV} \leftarrow$ COSINESIM(S_G, S_R, M_E) ▷ Sentence-wise precision and recall values
$pV \leftarrow$ MEAN(raw_{pV})
$rV \leftarrow$ MEAN(raw_{rV})
$f_1 \leftarrow \dfrac{2 * pV * rV}{pV + rV}$
return (f_1, pV, rV)

procedure COSINESIM(S_G, S_R, M_E)
    $l_G \leftarrow$No. of sentences in S_G
    $l_R \leftarrow$No. of sentences in S_R
    init: $cosSs \leftarrow zeros[l_G, l_R]$; $i \leftarrow 0$
    for each sentence sG in S_G **do**
        $E_{sG} \leftarrow M_E(sG)$;$j \leftarrow 0$
        for each sentence sR in S_R **do**
            $E_{sR} \leftarrow M_E(sR)$
            $cosSs[i, j] \leftarrow Cos(E_{sG}, E_{sR})$
        end for
    end for
    $\boldsymbol{x} \leftarrow$ Row-wise-max($cosSs$)
    $\boldsymbol{y} \leftarrow$ Column-wise-max($cosSs$)
    return $(\boldsymbol{x}, \boldsymbol{y})$
end procedure

Algorithm 2 Threshold Function

procedure THRESHOLD$(rawSs, T)$
    initialize $Labels \leftarrow []$
    for each element e in $rawSs$ **do**
        if $e \geq t_u\%$ **then**
            $Labels.append(P)$
        else if $t_l\% \leq e \leq t_u\%$ **then**
            $Labels.append(PP)$
        else
            $Labels.append(A)$
        end if
    end for
    return $Labels$
end procedure

we experimented with 3 sentence encoder models: Paraphrase-distilroberta-base-v1 (*P-v1*) [14], stsb-roberta-large (*STSB*) [14] and universal-sentence-encoder (*USE*) [15]. Along with the various embedding models, we also experimented with multiple threshold values used to infer the

		Machine-Human Agreement in terms of Kendall Rank Correlation						
		$\mathbf{T = (25, 75)}$	$\mathbf{T = (35, 65)}$	$\mathbf{T = (45, 75)}$	$\mathbf{T = (55, 65)}$	$\mathbf{T = (55, 75)}$	$\mathbf{T = (55, 80)}$	$\mathbf{T = (60, 80)}$
Sentence Embedding: P-v1								
Precision Reward	L_1	0.55	0.6	0.58	0.59	0.57	0.56	0.54
	L_2	0.61	0.67	0.63	0.67	0.64	0.67	0.68
	L_3	0.54	0.62	0.56	0.64	0.6	0.56	0.52
Recall Reward	L_1	0.53	0.64	0.66	0.62	0.61	0.62	0.59
	L_2	0.55	0.64	0.67	0.63	0.63	0.64	0.61
	L_3	0.54	0.65	0.64	0.66	0.65	0.65	0.61
Sentence Embedding: STSB								
Precision Reward	L_1	0.57	0.67	0.58	0.66	0.6	0.57	0.58
	L_2	0.66	0.63	0.65	0.63	0.7	0.63	0.6
	L_3	0.56	0.57	0.58	0.56	0.59	0.57	0.56
Recall Reward	L_1	0.55	0.65	0.64	0.62	0.62	0.61	0.59
	L_2	0.56	0.65	0.65	0.63	0.63	0.64	0.63
	L_3	0.54	0.59	0.61	0.57	0.58	0.57	0.54
Sentence Embedding: USE								
Precision Reward	L_1	0.58	0.62	0.6	0.61	0.59	0.62	0.65
	L_2	0.68	0.7	0.68	0.68	0.68	0.7	0.73
	L_3	0.66	0.67	0.65	0.64	0.63	0.53	0.56
Recall Reward	L_1	0.53	0.59	0.56	0.61	0.62	0.61	0.6
	L_2	0.54	0.6	0.61	0.62	0.64	0.64	0.62
	L_3	0.52	0.6	0.58	0.61	0.61	0.6	0.6

Table 11: Average Precision and Recall Kendall Tau between label-annotators (L_i) and automatically inferred labels using SEM-F_1. The results are shown for different embedding models and multiple threshold levels $T = (t_l, t_u)$. For all the annotators L_i ($i \in \{1, 2, 3\}$), correlation numbers are quite high (≥ 0.50). Moreover, the reward values are consistent/stable across all 5 embedding models and threshold values. All values are statistically significant at p-value<0.05.

sentence-wise overlap labels: *presence (P)*, *partial presence (PP)* and *absence (A)*, in order to simulate different user preferences and accordingly, report the sensitivity of the metric with respect to different thresholds. These thresholds are: $(25, 75)$, $(35, 65)$, $(45, 75)$, $(55, 65)$, $(55, 75)$, $(55, 80)$, $(60, 80)$. For example, the threshold range $(45, 75)$ means that if the similarity score $<$ 45%, infer the label "absent", else if the similarity score $\geq$ 75%, infer the label "present" and else, infer the label "partially-present". Next, we computed the average precision and recall rewards for 50 samples annotated by label-annotators (L_i) and the labels inferred by SEM-F_1 metric. For this, we repeated the same procedure as in Table 9, but this time compared human labels against "SEM-F_1" inferred labels. The corresponding results are shown in 12. As we can notice, the average reward values are consistently high (≥ 0.50) for all the 3 label-annotators (L_i). Moreover, the reward values are stable across all the 3 embedding models and threshold values, signifying that SEM-F_1 is indeed robust across various sentence embeddings and thresholds used.

Following the procedure in Table 7, we also compute Kendall's Tau between human label annotators and automatically inferred labels using SEM-F_1. Our results in table Table 11 are consistent with both reward-based inter-rater-agreement (Table 9) and Kendall rank correlation -

		Machine-Human Agreement in terms of Reward Function						
		$\mathbf{T = (25, 75)}$	$\mathbf{T = (35, 65)}$	$\mathbf{T = (45, 75)}$	$\mathbf{T = (55, 65)}$	$\mathbf{T = (55, 75)}$	$\mathbf{T = (55, 80)}$	$\mathbf{T = (60, 80)}$
Sentence Embedding: P-v1								
Precision Reward	L_1	0.73 ± 0.27	0.81 ± 0.25	0.77 ± 0.26	0.85 ± 0.23	0.80 ± 0.24	0.77 ± 0.24	0.77 ± 0.26
	L_2	0.72 ± 0.30	0.73 ± 0.29	0.73 ± 0.30	0.78 ± 0.27	0.79 ± 0.27	0.75 ± 0.26	0.73 ± 0.29
	L_3	0.81 ± 0.23	0.86 ± 0.21	0.79 ± 0.24	0.78 ± 0.28	0.74 ± 0.28	0.69 ± 0.28	0.69 ± 0.27
Recall Reward	L_1	0.66 ± 0.19	0.79 ± 0.16	0.75 ± 0.16	0.76 ± 0.18	0.71 ± 0.17	0.66 ± 0.17	0.61 ± 0.18
	L_2	0.67 ± 0.19	0.78 ± 0.16	0.76 ± 0.15	0.73 ± 0.19	0.72 ± 0.18	0.70 ± 0.18	0.65 ± 0.21
	L_3	0.66 ± 0.15	0.72 ± 0.17	0.68 ± 0.17	0.68 ± 0.22	0.64 ± 0.20	0.59 ± 0.19	0.57 ± 0.20
Sentence Embedding: STSB								
Precision Reward	L_1	0.75 ± 0.29	0.75 ± 0.29	0.75 ± 0.29	0.75 ± 0.29	0.75 ± 0.29	0.75 ± 0.30	0.75 ± 0.23
	L_2	0.63 ± 0.32	0.63 ± 0.31	0.63 ± 0.32	0.63 ± 0.31	0.63 ± 0.32	0.64 ± 0.32	0.64 ± 0.32
	L_3	0.81 ± 0.23	0.82 ± 0.23	0.81 ± 0.23	0.82 ± 0.23	0.81 ± 0.23	0.81 ± 0.22	0.81 ± 0.22
Recall Reward	L_1	0.66 ± 0.21	0.67 ± 0.21	0.66 ± 0.21	0.68 ± 0.21	0.67 ± 0.21	0.65 ± 0.21	0.66 ± 0.21
	L_2	0.57 ± 0.20	0.58 ± 0.21	0.57 ± 0.20	0.59 ± 0.20	0.59 ± 0.20	0.58 ± 0.20	0.58 ± 0.21
	L_3	0.67 ± 0.19	0.67 ± 0.20	0.67 ± 0.19	0.68 ± 0.20	0.68 ± 0.19	0.67 ± 0.18	0.68 ± 0.18
Sentence Embedding: USE								
Precision Reward	L_1	0.76 ± 0.29	0.77 ± 0.30	0.78 ± 0.27	0.80 ± 0.28	0.80 ± 0.27	0.77 ± 0.27	0.80 ± 0.27
	L_2	0.69 ± 0.32	0.66 ± 0.32	0.71 ± 0.30	0.68 ± 0.30	0.72 ± 0.30	0.76 ± 0.29	0.78 ± 0.29
	L_3	0.82 ± 0.24	0.85 ± 0.22	0.85 ± 0.23	0.86 ± 0.21	0.85 ± 0.23	0.82 ± 0.23	0.78 ± 0.25
Recall Reward	L_1	0.64 ± 0.19	0.67 ± 0.19	0.68 ± 0.19	0.70 ± 0.21	0.69 ± 0.22	0.64 ± 0.20	0.65 ± 0.21
	L_2	0.62 ± 0.19	0.63 ± 0.20	0.66 ± 0.18	0.66 ± 0.21	0.68 ± 0.20	0.68 ± 0.19	0.69 ± 0.21
	L_3	0.64 ± 0.16	0.68 ± 0.19	0.66 ± 0.16	0.69 ± 0.20	0.65 ± 0.19	0.60 ± 0.17	0.60 ± 0.18

Table 12: Average Precision and Recall reward/correlation (mean $\pm$ std) between label-annotators (L_i) and automatically inferred labels using SEM-F_1. The results are shown for different embedding models and multiple threshold levels $T = (t_l, t_u)$. For all the annotators L_i ($i \in \{1, 2, 3\}$), correlation numbers are quite high (≥ 0.50). Moreover, the reward values are consistent/stable across all 5 embedding models and threshold values.

	Random Reference SEM-F_1 Scores			**Random Output SEM-F_1 Scores**			**Actual SEM-F_1 Scores**		
	P-V1	**STSB**	**USE**	**P-V1**	**STSB**	**USE**	**P-V1**	**STSB**	**USE**
BART	0.16	0.21	0.22	0.21	0.27	0.27	0.65	0.67	0.67
T5	0.17	0.21	0.23	0.20	0.26	0.26	0.58	0.60	0.60
Pegasus	0.15	0.20	0.22	0.19	0.26	0.26	0.59	0.60	0.62
Average	0.16	0.21	0.22	0.20	0.26	0.26	0.61	0.62	0.63

Table 13: SEM-F_1 Scores and Random Baselines

based inter-rater-agreement (Table 7); the correlation values are ≥ 0.50 with little variation along various thresholds for both precision and recall.

SEM-F1 Scores and Distinguishability: Here, we present the actual SEM-F_1 scores for the three models (BART, T5 and Pegasus) along with scores for two random baselines: 1) Random Refer-

Pearson's Correlation Coefficients									
	P-V1			STSB			USE		
	I_1	I_2	I_3	I_1	I_2	I_3	I_1	I_2	I_3
I_2	**0.69**	—		**0.65**	—		**0.71**	—	
I_3	**0.40**	**0.50**	—	**0.50**	**0.52**	—	**0.51**	**0.54**	—
I_4	**0.33**	**0.44**	**0.60**	**0.33**	**0.36**	**0.56**	**0.37**	**0.42**	**0.66**
Average		**0.49**			**0.49**			**0.54**	

Table 14: Max (across 3 models) Pearson's correlation between the SEM-F_1 scores corresponding to different annotators. Here I_i refers to the i^{th} annotator where $i \in \{1, 2, 3, 4\}$ and "Average" row represents average correlation of the max values across annotators. All values are statistically significant at p-value < 0.05.

ence, 2) Random Output.

Random Reference: Here, the model-generated summary is compared against a random reference to compute SEM-F_1 scores. The random selection is done by sampling a reference summary from the pool of remaining $136 \times 4 = 544$ references.

Random Output: In this case, a randomly generated output is compared against actual human-written reference summaries to compute SEM-F_1 scores. The random selection is done by sampling a machine-generated output from the pool of remaining 136 machine-generated outputs.

As reported in table 13, abstractive summarization models achieve approximately 40-45 percent improvement over the random baseline scores suggesting SEM-F_1 can indeed distinguish the *"good"* from the *"bad"*.

Pearson Correlation for SEM-F1: Following the case-study based on ROUGE, we computed the Pearson's correlation coefficients between each pair of raw SEM-F_1 scores obtained using each of the 4 reference summaries. The corresponding correlations are shown in Table 14. For each annotator pair, we report the maximum (across 3 models) correlation value. The average correlation value across annotators is 0.49, 0.49 and 0.54 for P-V1, STSB, USE embeddings, respectively, suggesting a clear improvement over ROUGE.

2.3 Findings under Objective 3

2.3.1 Synthetic Training Data Generation

Our basic idea is to divide a given document D into two parts D_1 and D_2 such that there is a non-empty overlap between D_1 and D_2 in terms of the sentences they contain, i.e., $D_1 \cap D_2 = D_O (\neq \phi)$ and $D_1 \cup D_2 = D$ (constraint I). Here the $\cap$ and $\cup$ operators are classical set operators, i.e., they mean intersection and union in terms of the set of sentences and should not be confused with *SOS* output ($\cap_O$). Now, consider $\{\{D_1, D_2\}, D_O\}$ as our training sample where the unordered pair $\{D_1, D_2\}$ is the input to our *SOS* model and D_O is the target overlap summary. If we naively train a model on such samples, it will simply learn to copy the repeated sentences (D_O) and would fail terribly in a real testing scenario. Also, identifying repeated sentences is a trivial task and training a seq-to-seq model for this has no practical value. Indeed, true semantic overlap should be

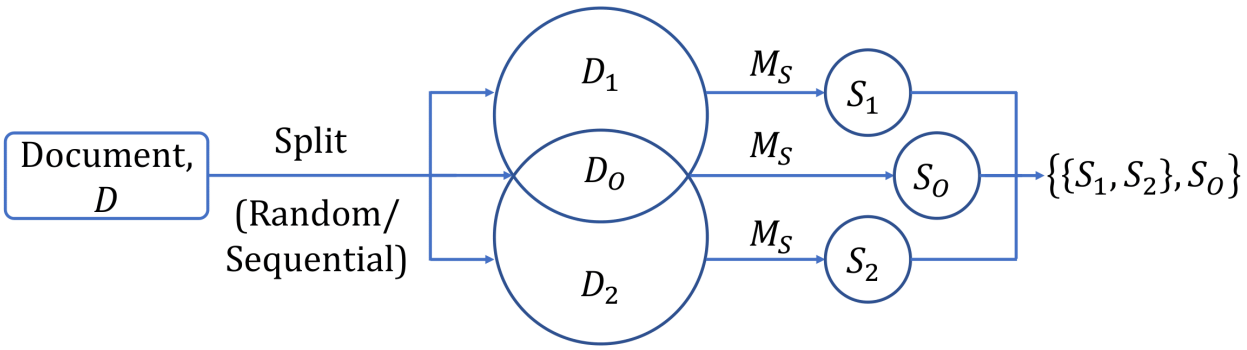

Figure 2: Synthetic Training Data Generation for Semantic Overlap

written in an abstract fashion, which is a much harder computational task than identifying repeated sentences.

Now, assume we have a *perfect abstractive summarizer* M_S and using it, we generate summaries for each of the documents D_1, D_2 and D_O. More specifically, we generate summaries S_1, S_2 and S_O which would contain the core information content of the original documents D_1, D_2 and D_O, respectively. Although D_1, D_2 and D_O have some repeated sentences between them by definition, assuming a perfect abstractive summarizer, one can expect that S_1, S_2 and S_O will most likely have no repeated sentence as they have been transformed through an abstractive summarizer. The assumption of a perfect abstractive summarizer also means that S_O will only have the common information present in both S_1 and S_2. In other words, S_O can be regarded as a true semantic overlap of S_1 and S_2 and at the same time, S_1, S_2 and S_O will have minimal lexical overlap. Thus, having $\{\{S_1, S_2\}, S_O\}$ as our synthetic sample will be perfect for training a seq-to-seq model with $\{S_1, S_2\}$ being the input and S_O as the target semantic overlap.

However, in the absence of a perfect summarizer, we hypothesize that a reasonable abstractive summarizer pre-trained on a particular domain will be able to generate a large number of noisy synthetic training examples in the form of $\{\{S_1, S_2\}, S_O\}$ and subsequently, fine-tuning a seq-to-seq model using such noisy data will still help us in learning to generate overlap summaries. One nice benefit of our data generation technique is that a large number of synthetic samples can easily be generated from a domain-specific corpus of documents. By partitioning a single document into two overlapping segments and then introducing non-linearity through an abstractive summarization model, we propose a simple yet effective synthetic data generation technique for training the SOS task.

This process is described in algorithm 3 and visually presented in Figure 2. We use two basic heuristics methods to split the document into two halves, namely SEQUENTIALSPLIT and RANDOMSPLIT such that the constraint I holds. In the Sequential Split, we simply divide document D into two halves (D_1 and D_2) while keeping some common sentences among both of them. For example, for a common percentage value p (say 50), we choose the first 75 percent of the sentences as D_1 and the last 75 percent as document D_2. On the other hand, in Random Split, we randomly select some common sentences (C_S) and randomly divide the remaining sentences into two halves, say H_1 and H_2. To generate D_1, we combine/concatenate C_S and H_1 while keeping the original order of sentences in D intact. Similarly, for D_2, we combine C_S and H_2 while maintaining the original order.

Algorithm 3 Generate Synthetic Data.

given Document D, Abstractive Summarization Model M_S, Overlap Percentage p, Split Type spt
$D_1, D_2, D_O \leftarrow$ SPLIT(spt, D, p)
$S_1, S_2, S_O \leftarrow M_S(D_1), M_S(D_2), M_S(D_O)$
return $\{S_1, S_2\}, S_O$

---

procedure SPLIT(spt, Doc, p)
    if spt is sequential **then**
        return $\leftarrow$ SEQUENTIALSPLIT(Doc, p)
    else if spt is random **then**
        return $\leftarrow$ RANDOMSPLIT(Doc, p)
    end if
end procedure

---

procedure SEQUENTIALSPLIT(D, p)
    $d_1 \leftarrow$ First $\frac{100+p}{2}\%$ of sentences in D
    $d_2 \leftarrow$ Last $\frac{100+p}{2}\%$ of sentences in D
    $d_O \leftarrow$ Middle $p\%$ of sentences in D
    return (d_1, d_2, d_O)
end procedure

---

procedure RANDOMSPLIT(Doc, p)
    $d_{int} \leftarrow$ Pop $p\%$ of random sentences from D
    h_1, $h_2 \leftarrow$ Randomly partition $D - d_{int}$ in two halves
    $d_1 \leftarrow$ CONCAT(h_1, d_O) w.r.t. original order
    $d_2 \leftarrow$ CONCAT(h_2, d_O) w.r.t. original order
    return (d_1, d_2, d_O)
end procedure

2.3.2 Initial Qualitative Inspection

We started with a simple text dataset, i.e., the WikiHow dataset [16], to test whether our synthetic data generation process for semantic overlap is indeed going to work. To generate the synthetic reference summaries, we used the PEGASUS model [11], a state-of-the-art abstractive summarization model. Row 15.1 from Table 15 shows that the sentences in turquoise and yellow colour have indeed been summarized in the orange sentences in the output summary (S_O).

Next, we switched to the CNN-DailyMail dataset [17] since it is more in line with our AllSides testing dataset. We used the same process to generate synthetic samples as before, but this time, we observed an issue with the default settings of the PEGASUS model. Specifically, we found fabricated information in the S_O output summary which is not at all present in inputs S_1 and S_2 (red sentences in table row 15.2). This mainly happened because D_O, the input document to the PEGASUS model, was too small and we were simply expecting larger summaries from short input

Table 15: Qualitative analysis of generated synthetic samples. Turquoise, yellow and orange color shows the common information among S_1, S_2 and S_O respectively. The red colour marks some of the issues described in 2.3.2. (. . .) denotes the sentences which for not shown for brevity.

	S_1	S_2	S_O
	WikiHow Sample		
15.1	. . . Make a list of all of your artistic connections and contacts.<n> Keep track of all of your business expenses.<n>Calculate the cost of each piece you make.<n>Stay up to date on the art market in your area.<n>Devote time to your art.	Keep all of your receipts and expenses organized.<n>Calculate the cost of each piece you make.<n>Research the market for your work.<n>Price your work carefully.<n>	Keep track of all of your expenses.<n>Calculate the cost of each piece you make.<n>Keep up with the market.<n>Remember that time is money.
	CNN DailyMail Dataset: Fabricated Information		
15.2	Dr. Anthony Moschetto is charged in what authorities say was a failed scheme to have another physician hurt or killed.<n>Moschetto,54, pleaded not guilty to all charges Wednesday..He was released after posting $2 million bond and surrendering his passport.<n>Two other men - identified as James Chmela, 43, and James Kalamaras, 41 - - were named as accomplices.	Two other men - - identified as James Chmela, 43, and James Kalamaras, 41 - - were named as accomplices.<n>Police officers allegedly discovered approximately 100 weapons at Moschetto's home.<n>Moschetto allegedly told officers during one buy that he needed dynamite to "blow up a building"	The investigation began back in December, when undercover officers began buying heroin and oxycodone pills from Moschetto in what was initially a routine investigation into the sale of prescription drugs, officials said.<n>During the course of the undercover operation, however, Moschetto also sold the officers two semiautomatic assault weapons as well as ammunition, prosecutors said.<n> Police officers allegedly discovered approximately 100 weapons at . . .
	CNN DailyMail Dataset: Sample generated by controlling the length of output summaries. This helps in controlling the information fabrication issue		
15.3	. . . Al-Saeedni is the leader of a group that may have been inspired by al Qaeda, an Italian activist says.<n>The activist was also a freelance journalist.<n>Arrigoni was from the northern Italian region of Lombardy.<n> He was working in Gaza as a humanitarian activist.<n>. . . Arrigoni was also working as a freelance journalist.<n>He was from the northern Italian region of Lombardy.<n> WARNING GRAPHIC IMAGES.<n> The video was posted on YouTube on Thursday night.<n> . . .	. . . A video was posted on YouTube showing a man identified by his colleagues as Arrigoni.<n>Arrigoni was from the northern Italian region of Lombardy.<n> . . . He was also working as a freelance journalist.<n> . . . "Vittorio Arrigoni is a hero of Palestine," said a statement released by a Palestinian human rights official.<n> . . . Al-Saeedni is the leader of a group that may have been inspired by al Qaeda, an official said.<n>The video was posted hours after a man identified by his colleagues as Arrigoni was seen.<n>The grisly outcome came hours after a video was posted on YouTube showing a man identified by his colleagues as Arrigoni.<n> . . .	The abductors may have been inspired by al Qaeda, an Italian activist says.<n>Arrigoni was from the northern Italian region of Lombardy.<n>He was working as a freelance journalist.<n> The Palestinian Centre for Human Rights calls him a hero of Palestine.<n> A video of Arrigoni was posted on YouTube.<n> The activist's fate was unknown until his colleagues saw a video of him.<n>The video was posted hours after a man identified as Arrigoni was taken.

documents.

To mitigate this issue, we tried to control the length of the generated summaries (S_1, S_2 and S_O) so that the chances of information fabrication in the output (overlap) summary are low. The

samples produced from this approach can be seen in Table row 15.3 with length parameters set as follows: 200-300 words for S_1, S_2 and 50-100 words for S_O. Based on manual inspection, we found that the generated synthetic samples are satisfactory and thus, we stick with these settings for all the future experiments in the paper.

2.3.3 Quantitative Analysis

After the initial qualitative evaluation, we performed a quantitative evaluation of our synthetic data. First, we generated 4 variations of the synthetic dataset, which we call Rand35, Rand50, Seq35, Seq50 for the respective split-type (Sequential or Random) and overlap-percentage (35% or 50%) values.

Next, we computed the semantic similarity between synthetic summary pairs, i.e., the similarity between $\{S_1, S_2\}$, $\{S_1, S_O\}$ and $\{S_2, S_O\}$. The aim is to understand the impact of split-type and overlap-percentage parameters on the generation process. For semantic similarity, we utilized three sentence embedding models namely, Paraphrase-distilroberta-base-v1 (*P-v1*) [14], stsb-roberta-large (*STSB*) [14] and universal-sentence-encoder (*USE*) [15] and computed cosine similarity between the sentences of the two documents. The similarity between the two documents is computed as follows -

$$\frac{\frac{1}{n}\sum_j max_i\{cosine(A_i, B_j)\} + \frac{1}{m}\sum_i max_j\{cosine(A_i, B_j)\}}{2}$$

where A_i and B_j are the vectors corresponding to the i^{th} and j^{th} sentence in documents A and B with m and n sentences respectively. As we notice in Table 16, there is indeed enough overlap between synthetic summary pairs with 50% of variants showing higher overlap on the expected lines.

2.3.4 Further Validation by Humans

Following [18,19], we further involved human judges to evaluate the quality of generated synthetic samples. Humans evaluated the synthetic overlap summaries along the four dimensions: *Coherence*, *Consistency*, *Fluency*, *Relevance*; as done by [18,20,21]. We slightly modified the definition of *Consistency* and *Relevance* to fit our *SOS* task. *Coherence* and *Fluency* evaluate the quality of a document on its own, whereas, *Consistency* and *Relevance* evaluate the overlap summary given the input document pairs and is analogous to precision and recall, respectively. More details about them are provided below.

Coherence: It represents the collective quality of all sentences. This dimension aligns with the DUC quality question [22] of structure and coherence whereby the generated summary/document should be well-structured and well-organized. It should not just be a heap of related information but should build from sentence to sentence to a coherent body of information about a topic.

Fluency: It represents the quality of individual sentences. Again following DUC quality guidelines, the sentences in the generated summary should have no formatting problems, capitalization errors or ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.

Relevance: It checks whether the important overlapping content from the source documents has been selected and is similar to recall. The overlap summary should include only important infor-

USE	Rand-35	Seq-35	Rand-50	Seq-50
$\{S_1, S_2\}$	0.55	0.51	0.59	0.56
$\{S_1, S_O\}$	0.59	0.6	0.63	0.61
$\{S_2, S_O\}$	0.59	0.56	0.62	0.61
Average	0.57	0.56	0.61	0.59
P-V1	**Rand-35**	**Seq-35**	**Rand-50**	**Seq-50**
$\{S_1, S_2\}$	0.56	0.53	0.61	0.57
$\{S_1, S_O\}$	0.6	0.61	0.64	0.63
$\{S_2, S_O\}$	0.6	0.57	0.63	0.63
Average	0.59	0.57	0.63	0.61
STSB	**Rand-35**	**Seq-35**	**Rand-50**	**Seq-50**
$\{S_1, S_2\}$	0.58	0.55	0.62	0.59
$\{S_1, S_O\}$	0.61	0.63	0.65	0.64
$\{S_2, S_O\}$	0.61	0.59	0.65	0.64
Average	0.6	0.59	0.64	0.62

Table 16: Sentence-wise Similarity scores between document pairs for various synthetic datasets.

mation[3] from the source documents. Annotators were told to penalize overlap summaries which contain redundancies and excess information.

Consistency: A factually consistent overlap summary should contain only statements that are there in both the input documents. Annotators were told to penalize the overlap summary that contains hallucinated facts. It is similar to precision.

In summary, Coherence and Fluency evaluate a given document individually whereas Consistency and Relevance evaluate the overlap summary given the input documents pair.

	$\mathbf{S_1}$	$\mathbf{S_2}$	$\mathbf{S_O}$
Coherence	4.13	4.23	4.28
Fluency	4.48	4.47	4.65
Consistency	-	-	3.93
Relevance	-	-	3.87

Table 17: Average (across 20 samples and 3 annotators) human evaluation scores (on a scale of 1-5) of the synthetic samples across 4 dimensions.

We asked 3 humans[4] to rate the summaries on a Likert scale from 1 to 5 (higher better) for 20 synthetic samples across the above specified 4 dimensions. For a given sample, a human would

[3]The reason to say important information is that our task is constrained summarization. So we are not expecting the overlap summary to have all the common facts from input documents pair.

[4]All graduate students with research experience in NLP.

assign 2 labels (*Coherence* and *Fluency*) for 3 documents (S_1, S_2, S_O) and another two labels (*Consistency* and *Relevance*) for S_O given the input documents pair $\{S_1, S_2\}$, i.e. 8 numbers or labels per sample. In total, we had $20 \times 8 \times 3 = 480$ labels annotated by humans. As we notice in Table 17, the generated samples are on average rated a score ≥ 4 across *Coherence* and *Fluency* and ~ 4 across *Consistency* and *Relevance*. These numbers are consistent with the prior results for the Pegasus model as reported by [18].

2.3.5 Experiments and Results

We aim to show the efficacy of our synthetic data generation technique rather than proposing a new specialized solution for the SOS task. Thus, we leverage off-the-shelf abstractive summarization models as a proxy for *SOS* models and simply, fine-tune them using our synthetic examples.

Baseline Models: We experimented with multiple SoTA pre-trained abstractive summarization models. These models are 1) **DistilBart** [23], distilled version of BART [10], 2) **Distill-PEGASUS** [23], distilled version of PEGASUS [11], and 3) **T5** [12] fine-tuned on multi-english Wiki news dataset. To generate the target overlap summary, we concatenate the two input documents $A \oplus B$ and $B \oplus A$ (where $\oplus$ represents concatenation operation) and feed them as two separate examples to the model.

Along with single document summarizers, we also experimented with a multi-document summarizer, **Hi-MAP** [24] (with default settings), since this is the only model trained in a supervised fashion compared to other available models [25, 26].

Implementation Details: In the case of Single Document Summarizers (SDS), we froze all the encoder layers and positional embeddings and only fine-tuned the decoder layers. All the 3 models were trained for 4 epochs and other hyper-parameters were set to their default values following the HuggingFace repo. On the other hand, Multi-Document Summarizer (MDS) was trained for 10,000 steps with default parameter settings following the official repository. The AS_T (AllSides training data) was used as the validation set to avoid over-fitting.

For testing, we report the average ROUGE-F_1 score [9] (from 137 samples) by comparing the machine-generated overlap summaries against the four human-written reference summaries.

Fine-Tune with CNN Synthetic Data: We took 1000 documents from the CNN/DailyMail dataset and created 4 versions of synthetic datasets as described in section 2.3.3. We also created one more synthetic dataset by using 10K sample from CNN/DailyMail, we call it Rand50-10K (random split with 50% overlap). As a whole, we call these datasets, *CNN Synthetic Datasets*. Initially, we only experimented with the DistilBart model. We observed that none of the models trained on *CNN Synthetic Datasets* shows any improvement over the baseline performance (Raw scores for each model are presented below in the appendix, table 18).

Fine-tune with AllSides Synthetic Data: Due to the lack of success with CNN dataset, we hypothesized that the reason for this is the difference in data distribution, i.e., AllSides testing data is different from CNN the DailyMail dataset. To test this, we created the synthetic dataset using the AllSides training set (AS_T). More specifically, we only took individual articles into consideration and used our synthetic data generation algorithm to create synthetic samples (random split with 50% overlap). To be very clear, we never look at the ground truth overlap summary or "theme-description".

DistilBart	R1	R2	RL
Baseline	0.44992	0.28450	0.36472
Seq35	0.45236	0.28433	0.36315
Rand35	0.44927	0.28124	0.36181
Seq50	0.45276	0.28500	0.36374
Rand50	0.44977	0.28136	0.36167
Rand50-10K	0.44977	0.28136	0.36167

Table 18: ROUGE Scores for baseline DistilBart compared to the one fine-tuned on *CNN Synthetic Datasets*.

		R1	R2	RL
Distil-Bart	**B**	0.45	0.28	0.36
	FT	**0.48**	**0.34**	**0.41**
Distill-Pegasus	**B**	0.46	0.30	0.38
	FT	**0.47**	**0.33**	**0.40**
T5	**B**	0.39	0.26	0.28
	FT	**0.47**	**0.32**	**0.38**
Hi-Map	**B**	0.39	0.24	0.26
	FT	**0.47**	**0.32**	**0.39**

Table 19: ROUGE Scores using *AllSides Synthetic Dataset*. **B** is the model and **FT** is the fine-tuned model. All **FT** models perform better than **B** models across all the 3 ROUGE metrics with statistically significant performance improvements (p-value < 0.05).

The model performance in the test set is reported in table 19 for all the representative models. All the 4 models fine-tuned using *AllSides Synthetic Dataset* outperform their baseline variants across all the 3 ROUGE metrics (p-value < 0.05). This shows that our synthetic data generation can indeed help in learning to generate *Overlap Summaries*.

Fine-tune with Golden Training Data: Next, we wanted to quantify how bad is training with noisy synthetic data compared to training with high-quality golden data for our SOS task. Fortunately, we do have ~2750 training samples (AS_T dataset) from AllSides. Therefore, we selected 2000 samples for training/fine-tuning the 4 models and the remaining samples are used for validation. Then we conducted training on this golden dataset to report the upper bound of ROUGE scores. As we notice in table 20, models trained on the synthetic dataset suffer little accuracy loss compared to the models trained on the gold dataset. More surprisingly, for Distil-Pegasus and Hi-Map, our synthetic data significantly outperformed training with golden data, demonstrating the effectiveness of noisy synthetic examples for training an SOS model.

Fine-tune with Augmented Data: We also tested the performance of models fine-tuned on the

		R1	R2	RL
Distil-Bart	**FT-S**	0.48	0.34	0.41
	FT-G	0.54	0.38	0.47
	FT-A	0.51	0.36	0.44
Distill-Pegasus	**FT-S**	**0.47**	**0.33**	**0.40**
	FT-G	0.47	0.31	0.39
	FT-A	0.48	0.34	0.41
T5	**FT-S**	0.47	0.32	0.38
	FT-G	0.53	0.36	0.46
	FT-A	0.48	0.33	0.41
Hi-Map	**FT-S**	**0.47**	**0.32**	**0.39**
	FT-G	0.39	0.20	0.32
	FT-A	0.50	0.35	0.44

Table 20: Comparison of ROUGE Scores for models fine-tuned on AllSides Gold data (**FT-G**) VS AllSides Synthetic Data (**FT-S**) VS Augmented Data (**FT-A**).

augmented data, i.e., gold + synthetic data, by combining the 2K all sides gold samples from the previous experiment with all synthetic data. This new augmented data is used to fine-tune all 4 models and their respective rouge scores are reported in Table 20. As expected, **FT-A** models consistently perform better than **FT-S** models across all 3 rouge metrics. However, when compared with **FT-G** models, **FT-A** models perform just like **FT-S** models. More specifically, for Distill-Pegasus and Hi-Map models, **FT-A** performed better than **FT-G** models. We believe this phenomenon occurs because our augmented data contains a lot more (noisy) synthetic samples compared to gold samples ($> 50\%$).

3 Final Words

In 2018, the PI introduced the vision of SOFSAT (Set-like Operator based Framework for Semantic Analysis of Text) [27], where he imagined a novel framework that can support set-like operators (TextIntersect/Overlap, TextUnion, and TextDifference) for semantic analysis of MPNs. After joining Auburn as a tenure-track assistant professor, the PI started the actual design and development of the original SOFSAT idea; as a first step, he focused on the TextIntersect/Overlap operator by framing it as a constrained summary generation task, where the constraint is the commonality criteria between two narratives.

As no benchmark dataset was readily available for the *Overlap* generation task, we created one by collecting $2,925$ alternative narrative pairs from the web and then went through the tedious pro-

cess of manually creating 411 different reference summaries by engaging human annotators [28]. For evaluating the performance of *Overlap summary generation* task, we proposed a new precision-recall style evaluation metric, called SEM-F_1 (Semantic F_1). Experimental results show that the proposed SEM-F_1 metric yields a higher correlation with human judgment as well as higher inter-rater-agreement compared to the traditional ROUGE metric, which has been recently accepted by EMNLP 2022 [29]. One of the major challenges in generating *Overlap* summaries is the lack of existing datasets for supervised training. To address this challenge, we proposed a novel data augmentation technique, which allows us to create a large amount of synthetic data for training a seq-to-seq model that can perform the semantic overlap generation task. Through extensive experiments using narratives from the news domain, we showed that the models fine-tuned using the synthetic dataset provide significant performance improvements over the pre-trained vanilla summarization techniques and are close to the models fine-tuned on the golden training data, which essentially demonstrates the effectiveness of our proposed data augmentation technique for training seq-to-seq models. This work has already been accepted to EMNLP 2022 [30].

4 Acknowledgements

This work has been partially supported by Army Research Office (ARO) Grant Award #W911NF22-1-0280 (ARO Proposal No. 79475-MI-II). We thank Auburn University College of Engineering and the Department of CSSE for their continuous support through Student Fellowships and Faculty Startup Grants.

References

[1] S. Somasundaran, M. Flor, M. Chodorow, H. Molloy, B. Gyawali, and L. McCulla, "Towards evaluating narrative quality in student writing," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 91–106, 2018.

[2] B. S. Bijoy, S. J. Saba, S. Sarkar, M. S. Islam, S. R. Islam, M. R. Amin, and S. K. Karmaker Santu, "Covid19α: Interactive spatio-temporal visualization of covid-19 symptoms through tweet analysis," in *26th International Conference on Intelligent User Interfaces-Companion*, 2021, pp. 28–30.

[3] O. Azeroual and H. Theel, "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*, vol. 4, no. 3, pp. 30–40, 2018.

[4] N. Hassan, A. Poudel, J. Hale, C. Hubacek, K. T. Huq, S. K. K. Santu, and S. I. Ahmed, "Towards automated sexual violence report tracking," in *Proceedings of the international AAAI conference on web and social media*, vol. 14, 2020, pp. 250–259.

[5] S. K. Karmaker Santu, L. Li, Y. Chang, and C. Zhai, "Jim: Joint influence modeling for collective search behavior," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 637–646.

[6] S. Wilson, F. Schaub, F. Liu, K. M. Sathyendra, D. Smullen, S. Zimmeck, R. Ramanath, P. Story, F. Liu, N. Sadeh *et al.*, "Analyzing privacy policies at scale: From crowdsourcing to automated annotations," *ACM Transactions on the Web (TWEB)*, vol. 13, no. 1, pp. 1–29, 2018.

[7] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, p. 113679, 2020.

[8] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," *Artificial Intelligence Review*, vol. 47, no. 1, pp. 1–66, 2017.

[9] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[10] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[11] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," *arXiv preprint arXiv:1912.08777*, 2019.

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *arXiv preprint arXiv:1910.10683*, 2019.

[13] D. Harman and P. Over, "The effects of human variation in DUC summarization evaluation," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 10–17. [Online]. Available: https://aclanthology.org/W04-1003

[14] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[15] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[16] M. Koupaee and W. Y. Wang, "Wikihow: A large scale text summarization dataset," *arXiv preprint arXiv:1810.09305*, 2018.

[17] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," *arXiv preprint arXiv:1704.04368*, 2017.

[18] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 2021.

[19] P. West, C. Bhagavatula, J. Hessel, J. D. Hwang, L. Jiang, R. L. Bras, X. Lu, S. Welleck, and Y. Choi, "Symbolic knowledge distillation: from general language models to commonsense models," *arXiv preprint arXiv:2110.07178*, 2021.

[20] S. Gehrmann, Y. Deng, and A. M. Rush, "Bottom-up abstractive summarization," *arXiv preprint arXiv:1808.10792*, 2018.

[21] W. Kryściński, N. S. Keskar, B. McCann, C. Xiong, and R. Socher, "Neural text summarization: A critical evaluation," *arXiv preprint arXiv:1908.08960*, 2019.

[22] H. T. Dang, "Overview of duc 2005," in *Proceedings of the document understanding conference*, vol. 2005, 2005, pp. 1–12.

[23] S. Shleifer and A. M. Rush, "Pre-trained summarization distillation," *arXiv preprint arXiv:2010.13002*, 2020.

[24] A. R. Fabbri, I. Li, T. She, S. Li, and D. R. Radev, "Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model," *arXiv preprint arXiv:1906.01749*, 2019.

[25] J. Zhao, M. Liu, L. Gao, Y. Jin, L. Du, H. Zhao, H. Zhang, and G. Haffari, "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*, 2020, pp. 1949–1952.

[26] L. Lebanoff, K. Song, and F. Liu, "Adapting the neural encoder-decoder framework from single to multi-document summarization," *arXiv preprint arXiv:1808.06218*, 2018.

[27] S. K. Karmaker Santu, C. Geigle, D. Ferguson, W. Cope, M. Kalantzis, D. Searsmith, and C. Zhai, "Sofsat: Towards a setlike operator based framework for semantic analysis of text," *ACM SIGKDD Explorations Newsletter*, vol. 20, no. 2, pp. 21–30, 2018.

[28] N. Bansal, M. Akter, and S. K. K. Santu, "Semantic overlap summarization among multiple alternative narratives: An exploratory study," in *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, N. Calzolari, C. Huang, H. Kim, J. Pustejovsky, L. Wanner, K. Choi, P. Ryu, H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S. Na, Eds. International Committee on Computational Linguistics, 2022, pp. 6195–6207. [Online]. Available: https://aclanthology.org/2022.coling-1.541

[29] N. Bansal, M. Akter, and **Karmaker Santu, Shubhra Kanti**, "Sem-f1: an automated metric for evaluating multi-narrative overlap summaries." in *EMNLP 2022*. Abu Dhabi, UAE: Association for Computational Linguistics, December 2022.

[30] N. Bansal, M. Akter, and S. K. Karmaker Santu, "Learning to generate overlap summaries through noisy synthetic data," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates:

Association for Computational Linguistics, Dec. 2022, pp. 11 765–11 777. [Online]. Available: https://aclanthology.org/2022.emnlp-main.807