Learning to Generate Overlap Summaries through Noisy Synthetic Data

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

Semantic Overlap Summarization (SOS) is a novel and relatively under-explored seq-to-seq 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 seq-to-seq model that can perform the SOS task. Through extensive experiments using narratives from the news domain, we show 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 out proposed data augmentation technique for training seq-to-seq models on the SOS task.

1 Introduction

Semantic Overlap Summarization (SOS) is a novel and relatively under-explored seq-to-seq task. It refers to the process of generating a summary of *common information* from multiple alternative narratives originating from various sources. Figure 1 depicts a toy use-case of *SOS* task on two alternative narratives (Bansal et al., 2022). Here, both articles cover the same event related to a supreme court ruling on *abortion* in Kentucky, but from two opposite political viewpoints; one from the left wing and the other from the right wing (color coded). The goal of the SOS task is to generate the overlapping information (in green text). In this sense, the SOS task can be defined as a multi-seq-to-seq task with an additional overlap constraint.

Multiple alternative narratives are frequent in a variety of domains, including education, health, and privacy, and summarizing the common information from these alternative narratives can be highly useful for digesting those multi-narratives at scale and speed (Karmaker Santu et al., 2018). However, one of the major challenges associated with implementing a seq-to-seq 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.

In this paper, we propose 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 $\{\{D_A, D_B\}, (D_A \cap_O D_B)\}$, where, D_A and D_B are two narratives (in text) and $D_A \cap_O D_B$ 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 seq-to-seq deep learning architectures.

Noteworthy, our main focus in this paper is to propose an intelligent way to create a synthetic dataset for training existing seq-to-seq 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 seq-to-seq summarization models as an approximation of the overlap summary generator and create artificial examples to further fine-tune such seq-to-seq models. As such, it is

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11765–11777
December 7-11, 2022

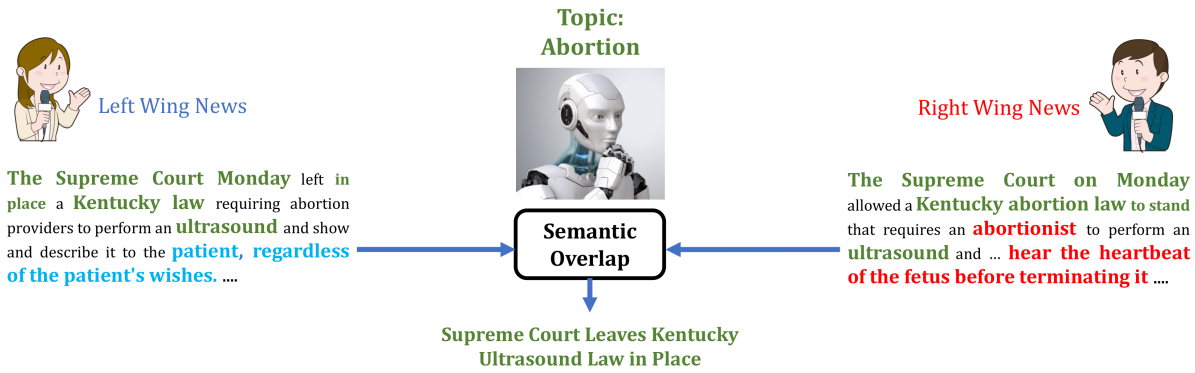

Figure 1: A toy use-case for Semantic Overlap Task (*TextOverlap*). A news on topic abortion has been presented by two news media (left-wing and right-wing). "Green" Text denotes the overlapping information from both news media, while "Blue" and "Red" text denotes the respective biases of *left* and *right* wing.

important to validate whether fine-tuning with artificial examples are indeed useful for improving the accuracy of the seq-to-seq models. To achieve this, we use the human annotated and verified dataset (Bansal et al., 2022) to show the efficacy of seq-to-seq 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 show 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. Overall, we make the following contributions in this paper.

1. We conduct a systematic study of the novel *Semantic Overlap Summarization* (SOS) task by formulating it as a constrained multi-seq-to-seq supervised learning problem.

2. We propose a new unsupervised synthetic data generation technique in the absence of any training dataset available for the SOS task. We conduct qualitative analysis and further, human experiments to show that generated synthetic samples are of high quality across 4 dimensions (section 5).

3. We conduct experiments using 3 single document summarizers and 1 multi-document summarizer to show that our synthetic data generation approach can indeed help in learning to generate *Overlap Summaries* (section 6).

2 Related Works

Text Summarization: Technically, *Semantic Overlap Summarization* can be viewed as a multi-document summarization task, i.e., multi-seq-to-seq task, with an extra commonality constraint (Bansal et al., 2022). Over the past two decades, many document summarizing approaches have been investigated (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; 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).

Pre-training/ Fine-Tuning Paradigm: Encoder-decoder-based neural models have recently gained a lot of attraction, especially for abstractive summarization tasks, (Rush et al., 2015; Chopra et al., 2016; Zhou et al., 2017; Paulus et al., 2017). Training a generic language model on a large corpus of data and then transferring/fine-tuning it for the summarization job has become a standard approach (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). 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; Lebanoff et al., 2018; Fabbri et al., 2019) is more challenging than single document summarization. However, unlike typical multi-document summarizing tasks, the *SOS* task aims to summarize multiple alternative narratives with an extra overlapping constraint, i.e., the output should only contain the com-

mon information from all the input narratives (Karmaker Santu et al., 2018).

Data Augmentation (DA): Given that the main focus of this work is on synthetic data generation, data augmentation literature is quite relevant. DA techniques aim to automatically augment or generate training samples without directly collecting more data. This is done either by directly modifying the existing samples or by creating new synthetic samples. On the text modification front, random noise is added to the input data (Xie et al., 2017) or hidden states (Le et al., 2015) to make the models more resistant to slight perturbations. Other basic approaches for data augmentation, such as word insertion, deletion, random synonym substitution, word order exchange, and so on, have also been investigated (Wei and Zou, 2019). Alternatively, tf-idf-based unwanted word removal has also been proposed (Xie et al., 2019).

On the text generation front, DA approaches employ generative models and sample synthetic samples from them. These approaches could be rule/ template-based (Leppänen et al., 2017), LDA-based (Ming et al., 2013), Markov chain models (Ghazal et al., 2013), hidden Markov models (Maqsud, 2015), VAE (Hu et al., 2017), GANS (Aghakhani et al., 2018), sequence-to-sequence (seq2seq) models (Yang et al., 2019; Hou et al., 2018; Santu et al., 2019), Grover method (Zellers et al., 2019), FactGen (Shu et al., 2020), GPT (Anaby-Tavor et al., 2020) etc for automatic text generation. Our proposed generation technique is similar in the sense that we also use a generative, particularly, summarization model. However, we employ simple yet effective techniques to ensure that the synthetic reference summary is a good representation of *Overlapping Information*, i.e., the "commonality" constraint is preserved.

3 Background

Here we first provide a brief description of the SOS task and the benchmark dataset that was introduced by Bansal et al. (2022).

3.1 Problem Formulation

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 thus, the *SOS* task is framed 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. Also, overlap summary should also have minimal repetition i.e. brevity is a desired property of *Semantic Overlap Summarization*. For example, if a particular piece of information or quote is repeated twice in both documents, we don't necessarily want it to be present in the output overlap summary two times. The output can either be an extractive summary or abstractive summary or a mixture of both, as per the use case. Additionally, *SOS* should follow the *commutative* property, i.e $D_A \cap_O D_B = D_B \cap_O D_A$.

3.2 The Benchmark Dataset

One of the key challenges with *SOS* task is that there is no existing dataset that we could readily use to evaluate it[1]. To this end, Bansal et al. (2022) recently presented the first benchmark dataset in the news domain by scraping the dataset from AllSides.com[2]. AllSides is a third-party online news platform that exposes individuals to news and information from all sides of the political spectrum (i.e. left, right, central), in order to provide an unbiased picture of the world to the general public. AllSides also includes a *factual description* (written by a human) of the reading content, titled "Theme", so that readers may see the so-called "neutral" point-of-view. Given two narratives ("Left" and "Right"), this theme-description was used as a proxy for ground truth for the *semantic overlap summary*. In total, a total of $2,925$ narrative pairs along with theme-descriptions (having a minimum length of 15 words) were collected. This data set was further separated into testing data (150 narrative pairs as explained below) and training data, AS_T (remaining samples).

Human Annotations[3]: Seq-to-seq tasks are often judged against multiple human-written references for robust evaluation. To create a testing benchmark with multiple human-written (ground-truth) references for semantic overlap summary, Bansal et al. (2022) randomly selected 150 narrative pairs and recruited three human volunteers to annotate our testing samples. Given a narrative pair, each an-

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

[2]AllSides is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

[3]The dataset and manual annotations can be found here.

notator was asked to read them carefully and then write a paragraph in their own words that capture the semantic overlap of the input narratives. This aided in the creation of a comprehensive testing benchmark for a more thorough assessment.

After the first round of annotations, they observed a discrepancy among the three annotators in terms of the *real* definition of "common/overlapping information". To mitigate the discrepancy, only the narrative pairs where at least two of the annotators wrote a minimum of 15 words as their reference summaries were retained. The idea was 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 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.

4 Challenges for Supervised Training

One of the major challenges with deep seq-to-seq models is that they require a huge amount of data samples to train the model. Even the largest pre-trained language models require at least a small amount of data to fine-tune them for a new domain (Howard and Ruder, 2018). In our case, we could fine-tune a large pre-trained language model on the AllSides data we collected with readily available reference summaries, which particularly covers narratives from the news domain; however, one may not find such annotated data in other domains like health, legal, education etc. In other words, while one may find multiple alternative narratives across many different domains, a dataset with the ground-truth overlap summaries of the narratives will not be available in most cases.

To address this challenge, we need to come up with an automatic technique for generating synthetic training data at a large scale. One naive idea is to divide a document into two segments such that some sentences are repeated in both segments and the repeated sentences can be regarded as the desired overlap. However, this approach will not work since the model will simply learn to copy the repeated sentences and ignore the actual semantics entirely. This is where our new synthetic data generation approach can help us circumvent this hurdle, the details of which are discussed in the following section.

5 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 simply 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 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 num-

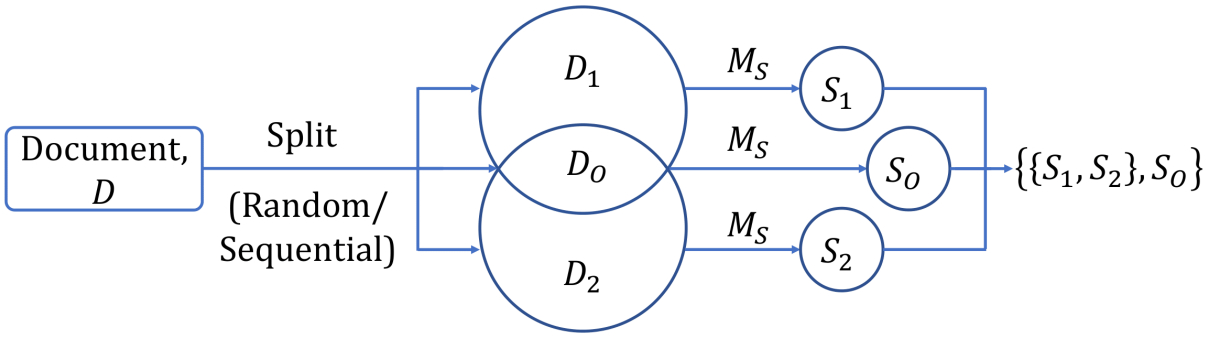

Figure 2: Synthetic Training Data Generation for Semantic Overlap

ber 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.

Algorithm 1 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

This process is described in algorithm 1 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.

5.1 Initial Qualitative Inspection

We started with a simple text dataset, i.e., the WikiHow dataset (Koupaee and Wang, 2018), 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 (Zhang et al., 2019), a state-of-the-art abstractive summarization model. Row 1.1 from Table 1 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 (See et al., 2017) 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 1.2). This mainly happened because D_O, the input document to the PEGASUS model, was too small and we were sim-

Table 1: 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 5.1. (. . .) denotes the sentences which for not shown for brevity.

	S_1	S_2	S_O
	WikiHow Sample		
1.1	. . . 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		
1.2	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**		
1.3	. . . 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.

ply expecting larger summaries from short input 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 1.3 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.

5.2 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

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

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*) (Reimers and Gurevych, 2019), stsb-roberta-large (*STSB*) (Reimers and Gurevych, 2019) and universal-sentence-encoder (*USE*) (Cer et al., 2018) 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 2, there is indeed enough overlap between synthetic summary pairs with 50% of variants showing higher overlap on the expected lines.

5.3 Further Validation by Humans

Following (Fabbri et al., 2021; West et al., 2021), 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 (Gehrmann et al., 2018; Kryściński et al., 2019; Fabbri et al., 2021). 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 (Dang, 2005) 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 information[4] 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.

We asked 3 humans[5] 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 assign 2 labels (*Coherence* and *Fluency*) for 3 documents (S_1, S_2, S_O) and another two labels (*Consistency*

[4]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.

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

	S_1	S_2	S_O
Coherence	4.13	4.23	4.28
Fluency	4.48	4.47	4.65
Consistency	-	-	3.93
Relevance	-	-	3.87

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

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 3, 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 (Fabbri et al., 2021).

6 Experiments and Results

As we mentioned in section 1, we aim is 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.

6.1 Baseline Models

We experimented with multiple SoTA pre-trained abstractive summarization models. These models are 1) **DistilBart** (Shleifer and Rush, 2020), distilled version of BART (Lewis et al., 2019), 2) **Distill-PEGASUS** (Shleifer and Rush, 2020), distilled version of PEGASUS (Zhang et al., 2019), and 3) **T5** (Raffel et al., 2019) 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** (Fabbri et al., 2019) (with default settings), since this is the only model trained in a supervised fashion compared to other available models (Zhao et al., 2020; Lebanoff et al., 2018).

6.2 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 (Lin, 2004) (from 137 samples) by comparing the machine-generated overlap summaries against the four human-written reference summaries.

6.3 Results

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 5.2. 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 6).

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".

The model performance in the test set is reported in table 4 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*.

		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 4: 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).

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 5, 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 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 5. 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 ($> 50\%$).

		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 5: 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**).

7 Conclusion

In this paper, we introduced a new and challenging task called Semantic Overlap Summarization (*SOS*) and proposed a novel data augmentation technique which helps us in creating a large number of synthetic training examples for this task. Although synthetic examples are not always 100% accurate, they can save a lot of time for humans and their efforts could be better directed towards evaluation (in line with West et al., 2021). In fact, qualitative inspection and human validation confirmed that the generated synthetic data was indeed meaningful. We further conduct quantitative experiments to confirm the efficacy of our approach. Additionally, one could create even larger datasets by using multiple seed values, different summarizers, varying overlap percentages etc. One particular limitation of this work is the need for an abstractive summarizer to generate the synthetic dataset and the quality of the generated samples is entirely dependent on it. Also, we have shown the efficacy of our approach in the news domain and leave its generalization capabilities to other domains for the future.

8 Acknowledgements

This work has been partially supported by Army Research Office (ARO) Grant Award #W911NF-22-1-0280 (ARO Proposal No. 79475-MI-II). We would also like to thank Auburn University College of Engineering and the Department of CSSE for their continuous support through Student Fellowships and Faculty Startup Grants.

References

Hojjat Aghakhani, Aravind Machiry, Shirin Nilizadeh, Christopher Kruegel, and Giovanni Vigna. 2018. Detecting deceptive reviews using generative adversarial networks. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 89–95. IEEE.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

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

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, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 6195–6207. International Committee on Computational Linguistics.

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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

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.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12.

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.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*.

Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. 2013. Bigbench: Towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD international conference on Management of data*, pages 1197–1208.

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*.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. *arXiv preprint arXiv:1807.01554*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

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*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International Conference on Machine Learning*, pages 1587–1596. PMLR.

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

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Quoc V Le, Navdeep Jaitly, and Geoffrey E Hinton. 2015. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. *arXiv preprint arXiv:1808.06218*.

Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. 2017. Data-driven news generation for automated journalism. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 188–197.

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*.

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*.

Umar Maqsud. 2015. Synthetic text generation for sentiment analysis. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 156–161.

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.

Zijian Ming, Chunjie Luo, Wanling Gao, Rui Han, Qiang Yang, Lei Wang, and Jianfeng Zhan. 2013. Bdgs: A scalable big data generator suite in big data benchmarking. In *Advancing big data benchmarks*, pages 138–154. Springer.

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*.

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*.

Nils Reimers and Iryna Gurevych. 2019. 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.

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

Shubhra Kanti Karmaker Santu, Kalyan Veeramachaneni, and Chengxiang Zhai. 2019. Tilm: Neural language models with evolving topical influence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 778–788.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Sam Shleifer and Alexander M Rush. 2020. Pre-trained summarization distillation. *arXiv preprint arXiv:2010.13002*.

Kai Shu, Yichuan Li, Kaize Ding, and Huan Liu. 2020. Fact-enhanced synthetic news generation. *arXiv preprint arXiv:2012.04778*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*.

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*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. 2017. Data noising as smoothing in neural network language models. *arXiv preprint arXiv:1703.02573*.

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*.

Ze Yang, Can Xu, Wei Wu, and Zhoujun Li. 2019. Read, attend and comment: a deep architecture for automatic news comment generation. *arXiv preprint arXiv:1909.11974*.

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*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *arXiv preprint arXiv:1905.12616*.

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*.

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

A Appendix

A.1 Fine-Tuning with CNN Synthetic Data

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