SEM-F_1: an Automatic Way for Semantic Evaluation of Multi-Narrative Overlap Summaries at Scale

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

Recent work has introduced an important yet relatively under-explored NLP task called **S**emantic **O**verlap **S**ummarization (**SOS**) that entails generating a summary from multiple alternative narratives which conveys the *common information* provided by those narratives. Previous work also published a benchmark dataset for this task by collecting $2,925$ alternative narrative pairs from the web and manually annotating 411 different reference summaries by engaging human annotators. In this paper, we exclusively focus on the automated evaluation of the *SOS* task using the benchmark dataset. More specifically, we first use the popular *ROUGE* metric from text-summarization literature and conduct a systematic study to evaluate the SOS task. Our experiments discover that *ROUGE* is not suitable for this novel task and therefore, we propose a new sentence-level precision-recall style automated evaluation metric, called **SEM-F_1** (Semantic F_1). It is inspired by the benefits of the sentence-wise annotation technique using overlap labels reported by the previous work. Our experiments show that the proposed **SEM-F_1** metric yields a higher correlation with human judgment and higher inter-rater agreement compared to the ROUGE metric.

1 Introduction

Human beings can be viewed as subjective sensors who observe real word events and report relevant information through their narratives (Karmaker Santu, 2019). Thus, multiple alternative narratives provide a robust way to comprehend the complete picture of an event being reported and verify corresponding facts and opinions from different perspectives. Despite great progress in NLP research in recent years, computers are still far from being able to accurately interpret multiple alternative narratives, which remains an open problem (Karmaker et al., 2021). In this paper, we study this challenging area of automatic summarization of multiple alternative narratives from different perspectives. More precisely, we exclusively focus on the *automated* evaluation of a new NLP task called **S**emantic **O**verlap **S**ummarization (**SOS**) from multiple alternative narratives. The *SOS* task has been introduced very recently by Bansal et al. (2022), where they conducted a systematic study of this task by creating a benchmark dataset as well as exploring how to manually evaluate this task. *SOS* essentially means the task of *summarizing the overlapping information* present in multiple alternate narratives by cross-verifying their information contents against each other. Computationally, the *SOS* task is defined as follows:

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

Multiple-perspective alternative narratives are frequent in a variety of domains, including education, the health sector, military intelligence, content analysis and privacy. 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 presents an example of the *SOS* task, where two human agents are reporting about the potential hiding location of a terrorist and the military general in charge of the mission wants to get a concise summary of the common information (reported by both parties) from both narratives. As shown in figure 1, both agents report that terrorist leader Y has been located (**Semantic Overlap**). However, Agent 342 reports the hiding location to be San Francisco (represented by *blue* text), whereas Agent 463 reports the location to be Portland (Oregon) (represented by *red* text). Agent 342 suspects that the target is wearing a suicide vest (represented by *blue* text), while Agent 463 mentions that the target is hiding in a tunnel (represented by *red* text).

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 780–792
December 7-11, 2022 ©2022 Association for Computational Linguistics

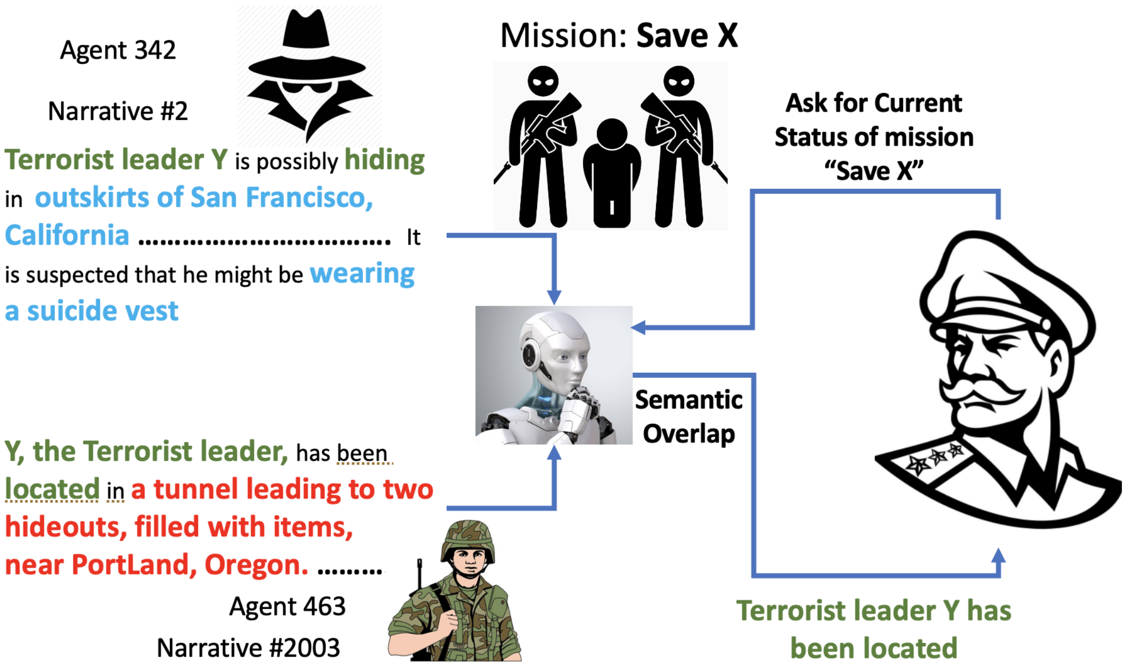

Figure 1: A toy example of *Semantic Overlap Summarization* (**SOS**) Task (from multiple alternative narratives). Here two human agents are reporting about the potential hiding location of a terrorist and the military general in charge of the mission wants to get a concise summary of the common information (reported by both parties) from both narratives. "Green" Text denotes the common information from both reports (Semantic Overlap), while "Blue" and "Red" text denotes the unique perspectives of each report.

The goal of *SOS* task is to generate a summary that conveys the common/overlapping information provided by the *green* text, i.e., the terrorist leader has been located.

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 that exactly matches our task; more importantly, it is unclear how to properly evaluate this task in an automated fashion. Therefore, as a starting point, we frame the *SOS* task as a constrained seq-to-seq problem 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 first address is the evaluation of the task. To address these challenges, we make the following contributions in this paper.

1. We frame *Semantic Overlap Summarization* (**SOS**) (from multiple alternative narratives) as a constrained multi-seq-to-seq problem and exclusively study how automatic evaluation of this task can be performed at a large scale.

2. As a starting point, we experiment with *ROUGE*, a widely popular metric for evaluating text summarization tasks, and demonstrate that *ROUGE* is NOT suitable for the automatic evaluation of *SOS* task.

3. Based on the findings of our previous work, we propose a new precision-recall style evaluation metric, **SEM-F$_1$** (Semantic F$_1$), for evaluating the *SOS* task. Extensive experiments show that new SEM-F$_1$ improves the inter-rater agreement compared to the traditional *ROUGE* metric, and also, shows a higher correlation with human judgments.

2 Related Works

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

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

Alternatively, one could aim to recover verb predicate-alignment structure (Roth and Frank, 2012; Xie et al., 2008; Wolfe et al., 2013) from a sentence and further, use this structure to compute the overlapping information (Wang and Zhang, 2009; Shibata and Kurohashi, 2012). Sentence Fusion is another related area which aims to combine the information from two given sentences with some additional constraints (Barzilay et al., 1999; Marsi and Krahmer, 2005; Krahmer et al., 2008; Thadani and McKeown, 2011). A related but simpler task is to retrieve parallel sentences (Cardon and Grabar, 2019; Nie et al., 1999; Murdock and Croft, 2005) without performing an actual overlap summary generation. However, these approaches are more targeted towards individual sentences and do not directly translate to arbitrarily long documents. Thus, the *SOS* task is still an open problem and there is no existing dataset, method or evaluation metric that has been systematically studied (Karmaker Santu et al., 2018). Recently, Bansal et al. (2022) conducted an initial exploration of the Semantic Overlap Summarization problem and created a benchmark dataset for further research in this area.

Along the evaluation dimension, *ROUGE* (Lin, 2004) is perhaps the most commonly used metric today for evaluating automated summarization techniques; due to its simplicity and automation. However, *ROUGE* has been criticized a lot for primarily relying on lexical overlap (Akter et al., 2022; Nenkova, 2006; Zhou et al., 2006; Cohan and Goharian, 2016) of n-grams. As of today, around 192 variants of *ROUGE* are available (Graham, 2015) including *ROUGE* with word embedding (Ng and Abrecht, 2015) and synonym (Ganesan, 2018), graph-based lexical measurement (ShafieiBavani et al., 2018), Vanilla *ROUGE* (Yang et al., 2018) and highlight-based *ROUGE* (Hardy et al., 2019). A recent study by Bansal et al. (2022) showed that the *ROUGE* metric is not appropriate for evaluating the *SOS* task. However, there has been no study yet on what can be an alternative to the ROUGE metric which is automatic and scalable, which is one of the central goals of our work.

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 $D_A \cap_O D_B = D_B \cap_O D_A$.

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 1: 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 "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.

3.2 The Benchmark Dataset

One of the key challenges with *SOS* task[1] is that there is no existing dataset for it. To this end, Bansal et al. (2022) presented the first benchmark dataset in the news domain by scraping the dataset from AllSides.com. AllSides is a third-party online news forum which 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 *factual* description of the reading material, labelled as "Theme" so that readers can see the so-called "neutral" point-of-view. Given two narratives ("Left" and "Right"), this theme-description is used as a proxy for ground truth reference summaries. They also engage human volunteers to thoroughly annotate the testing samples (narrative pairs) in order to create multiple reference overlap summaries for each pair. This helped in creating a comprehensive testing benchmark of 137 samples for more rigorous evaluation. Each narrative pair has 4 reference summaries, *one* from AllSides and *three* from human annotators, resulting in a total of 548 reference summaries.

4 Evaluating SOS Task using ROUGE

As *ROUGE* (Lin, 2004) is the most popular metric used today for evaluating summarization tasks; we first conducted a case study with *ROUGE* as the evaluation metric for the *SOS* task. For methods, we experimented with multiple SoTA pretrained abstractive summarization models as *naive baselines* for *Semantic-Overlap Summarizer (SOS)*. These models are 1) **BART** (Lewis et al., 2019), fine-tuned on CNN and multi english Wiki news datasets, 2) **Pegasus** (Zhang et al., 2019), fine-tuned on CNN and Daily mail dataset, and 3) **T5** (Raffel et al., 2019), fine-tuned on multi english Wiki news dataset. As our primary goal is to establish an appropriate metric for evaluating the *SOS* 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'll use the phrases "summary" and "overlap-summary" interchangeably from here.

Generating the summary: In order to handle two input documents, we concatenate them and feed the concatenated input directly to the model. The maximum summary length model hyper-parameter was set to 300 based on the max words across samples in the training data. The default values were used for all other hyper-parameters for each respective model.

Post-Processing: After the generation of model summaries, we did very basic post-processing. For example, for the Pegasus model, the new line character '<n>' was simply replaced by a blank space following the code from Huggingface.

For evaluation, we first evaluated the machine-generated overlap summaries for the 137 manually annotated testing samples using the ROUGE metric and followed the procedure mentioned in the paper (Lin, 2004) to compute the ROUGE-F_1 scores

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

with 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 appendix A).

Results and Findings: We computed Pearson's correlation coefficients 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 the *ROUGE* metric for evaluating the *SOS* task. The corresponding correlations are shown in table 1. 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*.

5 Sentence-wise Manual Scoring

Bansal et al. (2022) proposed 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 either absent, partially present or present 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 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*, and else, assign *PP* otherwise. 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 which is present in any of the references, it should be considered a hit. For recall, label-annotators were asked to assign labels to each sentence in each of the 4 reference summaries separately (provided by (I_1, I_2, I_3 and I_4)), with respect to the machine summary.

Inter-Rater-Agreement: After annotating each system-generated sentence (for precision) and reference sentence (for recall) with the labels (*A*, *PP* or *P*), we used the Kendall rank correlation coefficient to compute the pairwise annotator agreements among these ordinal labels. Table 2 shows that the correlations for both precision and recall are ≥ 0.50, signifying higher inter-annotator agreement.

Human agreement in terms of Kendall's Tau for 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 2: 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).

Reward-based Inter-Rater-Agreement: Alternatively, we defined a reward matrix (Table 3) 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 4. Both precision and recall reward scores are high (≥ 0.70) for all the different annotator pairs, thus signifying, high inter-label-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 3: Reward matrix used to compare the labels assigned by two label annotators for a given sentence to compute the agreement between the annotator pairs.

	Human agreement in terms of Reward function for Sentence-wise Scoring			
	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 4: 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.

We believe, one of the reasons for higher reward/Kendall scores could be that sentence-wise labelling puts a less cognitive load on the human mind and therefore, shows high agreement in terms of human interpretation. Similar observation is also noted in Harman and Over (2004).

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 5: Notations for algorithm 1

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) $\triangleright$ Sentence-wise precision and recall values
$pV \leftarrow$ Mean(raw_{pV})
$rV \leftarrow$ Mean(raw_{rV})
$f_1 \leftarrow \frac{2 * pV * rV}{pV + rV}$
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

6 Semantic-F_1: 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 5. 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.

6.1 Is SEM-F_1 Reliable?

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

		Machine-Human Agreement in terms of Kendall Rank Correlation						
		$\mathbf{T} = (\mathbf{25}, \mathbf{75})$	$\mathbf{T} = (\mathbf{35}, \mathbf{65})$	$\mathbf{T} = (\mathbf{45}, \mathbf{75})$	$\mathbf{T} = (\mathbf{55}, \mathbf{65})$	$\mathbf{T} = (\mathbf{55}, \mathbf{75})$	$\mathbf{T} = (\mathbf{55}, \mathbf{80})$	$\mathbf{T} = (\mathbf{60}, \mathbf{80})$
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 6: 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 (6.1) 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.

Algorithm 2 Threshold Function

```
procedure THRESHOLD(rawSs, T)
    initialize Labels ← []
    for each element e in rawSs do
        if e ≥ t_u% then
            Labels.append(P)
        else if t_l% ≤ e ≤ t_u% then
            Labels.append(PP)
        else
            Labels.append(A)
        end if
    end for
    return Labels
end procedure
```

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, we experimented with 3 sentence encoder models: 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). Along with the various embedding models, we also experimented with multiple threshold values used to infer the 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 4, but this time compared human labels against "SEM-F_1" inferred labels. The corresponding results are shown in 7. As

		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 7: Average Precision and Recall reward/correlation (mean ± std) between label-annotators (L_i) and automatically inferred labels using SEM-F_1. The results are shown for different embedding models (6.1) 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 8: *Actual SEM-F_1* and SEM-F_1 Scores for Random Baselines. The model-generated summaries are compared against a random reference summary in the case of *Random References* whereas, in the case of *Random Output*, randomly selected model output is compared against the true reference summary. As expected, *Actual SEM-F_1* scores are much higher than the random baselines.

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 2, we also compute Kendall's Tau between human label annotators and automatically inferred labels using SEM-F_1. Our results in table Table 6 are consistent with both reward-based inter-rater-agreement (Table 4) and Kendall rank correlation -based inter-rater-agreement (Table 2); the correlation values are ≥ 0.50 with little variation along various thresholds for both precision and recall.

6.2 SEM-F_1 Scores and Distinguishability

Here, we present the actual SEM-F_1 scores for the three models (BART, T5 and Pegasus) described in

Pearson's Correlation Coefficients for SEM-F_1									
	P-V1			**STSB**			**USE**		
	I_1	I_2	I_3	I_1	I_2	I_3	I_1	I_2	I_3
I_2	$\mathbf{0.69}$	—		$\mathbf{0.65}$	—		$\mathbf{0.71}$	—	
I_3	$\mathbf{0.40}$	$\mathbf{0.50}$	—	$\mathbf{0.50}$	$\mathbf{0.52}$	—	$\mathbf{0.51}$	$\mathbf{0.54}$	—
I_4	$\mathbf{0.33}$	$\mathbf{0.44}$	$\mathbf{0.60}$	$\mathbf{0.33}$	$\mathbf{0.36}$	$\mathbf{0.56}$	$\mathbf{0.37}$	$\mathbf{0.42}$	$\mathbf{0.66}$
Average		**0.49**			**0.49**			**0.54**	

Table 9: 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.

section 4 along with scores for two random baselines: 1) Random Reference, 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 8, 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"*.

6.3 Pearson Correlation for SEM-F_1

Following the case-study based on ROUGE in section 4, 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 9. 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.

7 Conclusions

In this work, we proposed a more accurate metric, called **SEM-$\mathbf{F_1}$**, for evaluating the *SOS* task. This metric compares the model-generated overlap summaries with the reference summary on a per-sentence basis using overlap labels and combines them to generate F_1 scores. Our experiments show that **SEM-$\mathbf{F_1}$** is more robust and yields higher agreement with human judgement and most importantly, can be computed automatically making it suitable for large-scale evaluation.

8 Limitations

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

We use the benchmark dataset proposed by Bansal et al., 2022 as our test set which has (~ 150 examples) and thus, makes it difficult to do a rigorous evaluation. We agree that having more samples in the test dataset would definitely help. But this is both time and money-consuming. We are working towards it and would like to increase the number of test samples in future.

9 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

Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker Santu. 2022. Revisiting automatic evaluation of extractive summarization task: Can we do better than rouge? In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1547–1560. Association for Computational Linguistics.

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

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.

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

Shubhra Kanti Karmaker, Md Mahadi Hassan, Micah J Smith, Lei Xu, Chengxiang Zhai, and Kalyan Veeramachaneni. 2021. Automl to date and beyond: Challenges and opportunities. *ACM Computing Surveys (CSUR)*, 54(8):1–36.

Shubhra Kanti Karmaker Santu. 2019. *Influence mining from unstructured big data*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A Appendix

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 10: 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.