Multi-Narrative Semantic Intersection Task: Evaluation and Benchmark

Anonymous ACL submission

Abstract

In this paper, we introduce an important yet relatively unexplored NLP task called *Multi-Narrative Semantic Intersection* (MNSI), which entails generating a *Semantic Intersection* of multiple alternate narratives. As no benchmark dataset is readily available for this task, we created one by crawling $2,925$ alternative narrative pairs from the web and then, went through the tedious process of manually creating 411 different ground-truth semantic intersections by engaging human annotators. As a way to evaluate this novel task, we first conducted a systematic study by borrowing the popular *ROUGE* metric from text-summarization literature and discovered that *ROUGE* is not suitable for our task. Subsequently, we conducted further human annotations/validations to create 200 document-level and $1,518$ sentence-level ground-truth labels which helped us formulate a new precision-recall style evaluation metric, called **SEM-F1** (semantic F1), based on presence, partial-presence and absence of information. Experimental results show that the proposed SEM-F1 metric yields higher correlation with human judgement as well as higher inter-rater-agreement compared to ROUGE metric and thus, we recommend the community to use this metric for evaluating future research on this topic.

1 Introduction

Human beings can be viewed as subjective sensors who observe real-word events and report relevant information through their own narratives. 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 still remains as an open problem.

In this paper, we look deeper into this challenging yet relatively under-explored area of automated understanding of multiple alternative narratives. To be more specific, we formally introduce a new NLP task called *Multi-Narrative Semantic Intersection* (MNSI) and conduct the first systematic study of this task by creating a benchmark dataset as well as proposing a suitable evaluation metric for the task. MNSI essentially means the task of extracting / paraphrasing / summarizing the *overlapping information* from multiple alternative narratives coming from disparate sources. In terms of computational goal, we study the following research question:

Given two distinct narratives N_1 and N_2 of some event e expressed in ***unstructured natural language*** *format, how can we extract the overlapping information present in both N_1 and N_2?*

Figure 1 shows a toy example of *MNSI* task, where the *TextIntersect*[1] ($\cap_T$) operation is being applied on two news articles. Both articles cover the same story related to the topic "abortion", however, they report from different political perspectives, i.e., one from *left* wing and the other from *right* wing. For greater visibility, "Left" and "Right" wing reporting biases are represented by *blue* and *red* text respectively. *Green* text denotes the common information in both news articles. The goal of *TextIntersect* ($\cap_T$) operation is to extract the overlapping information conveyed by the *green* text.

At first glance, the MNSI task may appear similar to 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 *MNSI*, the goal is to provide summarized content with an additional constraint, i.e., the commonality criteria. There is no current baseline method as well as existing dataset that exactly match our task; more importantly, it is unclear which one is the right evaluation metric to

[1] We'll be using the terms *TextIntersect* operator and *Semantic Intersection* interchangeably throughout the paper.

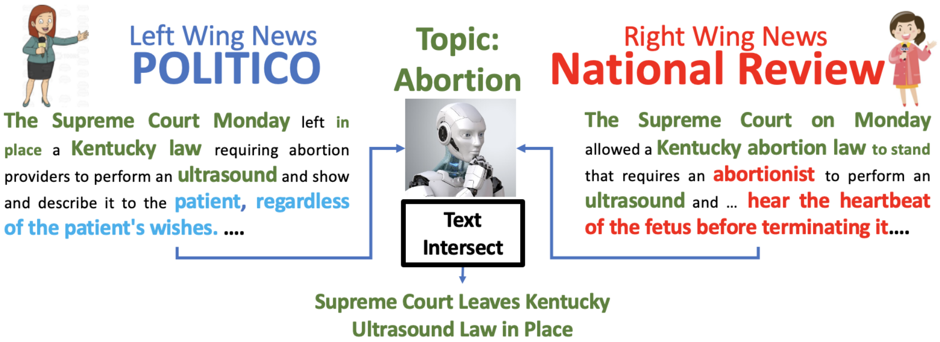

Figure 1: A toy use-case for Semantic Intersection Task (*TextIntersect*). A news on topic abortion has been presented by two news media (left-wing and right-wing). "Green" Text denotes the overlapping information from both news media, while "Blue" and "Red" text denotes the respective biases of *left*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

properly evaluate this task. As a starting point, we frame *MNSI* as a constrained summarization task where the goal is to generate a natural language output which conveys the overlapping information present in multiple input text documents. However, the bigger challenge we need to address first is the following: 1) How can we evaluate this task? and 2) How would one create a benchmark dataset for this task? To address these challenges, we make the following contributions in this paper.

1. We formally introduce *Multi-Narrative Semantic Intersection* (MNSI) as a new NLP task and conduct the first systematic study by formulating it as a constrained summarization problem.
2. We create and release the first benchmark dataset consisting of $2,925$ alternative narrative pairs for facilitating research on the *MNSI* task. Also, we went through the tedious process of manually creating 411 different ground-truth semantic intersections and conducted further human annotations/validations to create 200 document-level and $1,518$ sentence-level ground-truth labels to construct the dataset.
3. As a starting point, we experiment with *ROUGE*, a widely popular metric for evaluating text summarization tasks and demonstrate that *ROUGE* is NOT suitable for evaluation of *MNSI* task.
4. We propose a new precision-recall style evaluation metric, **SEM-F1** (semantic F1), for evaluating the *MNSI* task. Extensive experiments show that new SEM-F1 improves the inter-rater agreement compared to the traditional *ROUGE* metric, and also, shows higher correlation with human judgments.

2 Related Works

The idea of semantic text intersection is not entirely new, (Karmaker Santu et al., 2018) imagined a hypothetical framework for performing comparative text analysis, where, *TextIntersect* was one of the "hypothetical" operators proposed as part of the framework. However, the technical details and exact implementation were left as a future work.

As *Semantic Intersection* 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). Extractive approaches, as the name implies, generate summaries by extracting parts of the original document (usually sentences), while abstractive methods may generate new words or phrases which are not in the original document. 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.

Recently, encoder-decoder based neural models have become really popular for abstractive summarization (Rush et al., 2015; Chopra et al., 2016; Zhou et al., 2017; Paulus et al., 2017). It has become even prevalent to train a general language model on 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). However, *MNSI* task is different from traditional multi-document summarization tasks in that the goal here is to summarize content with an additional constraint: the **overlap** criteria, i.e., the output should only contain the common information from both input narratives.

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 intersection. However, these approaches are more targeted towards individual sentences and do not directly translate to arbitrarily long documents. Thus, MNSI task is still an open problem and there is no existing dataset, method or evaluation metric that have been systematically studied.

An idea conceptually similar to our work was applied on visual data (Alfassy et al., 2019), where the authors developed basic set-operators using neural network based approaches. However, we apply the idea on textual data which comes with entirely different set of challenges.

Along the evaluation dimension, *ROUGE* (Lin, 2004) is perhaps the most commonly used metric today for evaluating automated summarization techniques; due to its simplicity and automation. However, *ROUGE* has been criticized a lot for primarily relying on lexical overlap (Nenkova, 2006) of n-grams. Later, (Zhou et al., 2006) proposed for the use of a large broad domain-independent para table derived from a bilingual parallel corpus to allow para matching for summary evaluation. (Cohan and Goharian, 2016) demonstrated that *ROUGE* performs poorly in cases of terminology variation and paraphrasing. As of today, around 192 variants of *ROUGE* are available (Graham, 2015) including *ROUGE* with word embedding (Ng and Abrecht, 2015) and synonym (Ganesan, 2018), graph-based lexical measurement (ShafieiBavani et al., 2018), Vanilla *ROUGE* (Yang et al., 2018) and highlight-based *ROUGE* (Hardy et al., 2019). However, there has been no study yet whether *ROUGE* metric is appropriate for evaluating the *Semantic Intersection* task, which is one of central goals of our work.

3 Motivation

Multiple alternative narratives are very common across many domains like education, medicine, privacy etc., and thus, MNSI/TextIntersect operation can be very useful to digest such multi-narratives at scale and speed. Below are some use-cases.

Military Intelligence: If A and B are two intelligence reports related to a mission from two human agents, the *TextIntersect* operation can help verify the claims in each report w.r.t. the other.

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

Medical: *TextIntersect* can be applied on clinical notes of patients provided with the same treatment to understand the success/effects of the treatment.

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

4 Problem Formulation

What is *Semantic Intersection*? This is indeed a philosophical question and there is no single correct answer (various possible definitions are mentioned in appendix section A). 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 *TextIntersect* repeatedly. Also, let us define the output as $D_{int} \leftarrow D_A \cap_T D_B$. A human would mostly express the output in the form of natural language and this is why, we frame the *MNSI* task as a constraint summarization problem such that the output summary only contains information that is present in both the input documents. It can either be extractive summary or abstractive summary or a mixture of both, as per the use case. This task is inspired by the set-theoretic intersection operator. However, unlike set-intersection, our *Text Intersection* does not have to be the *maximal* set. The aim is summarize the overlapping information in an abstractive fashion. For example, if a particular piece of information or quote is repeated twice in both the documents, we don't necessarily want it to be present in target intersection summary two times. On the other hand, *Semantic Intersection* should follow the *commutative* property i.e $D_A \cap_T D_B = D_B \cap_T D_A$.

5 The Benchmark Dataset

As mentioned in section 1, there is no existing dataset which we could readily use to evaluate the *MNSI* task[2]. To address this challenge, we crawled data 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 their own *factual* description of the reading material, labeled as "Theme" so that readers can see the so-called "neutral" point-of-view. Table 1 gives an overview of the dataset created by crawling from AllSides.com, which consists of news articles (from at least one "Left" and one "Right" wing media) covering $2,925$ events in total and also having a minimum length of "theme-description" to be 15 words. Given two narratives ("Left" and "Right"), we used the theme-description as a proxy ground-truth *Text-Intersection* for this work. We divided this dataset into testing data (described next) and training data (remaining samples) and their statistics in provided in appendix (table 13).

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

Table 1: Overview of dataset scraped from AllSides

Human Annotations[3]: We decided to involve human volunteers to annotate our testing samples in order to create multiple human-written ground-truth semantic intersections for each event narrative pairs. This helped in creating a comprehensive testing benchmark for more rigorous evaluation. Specifically, we randomly sampled 150 narrative pairs (one from "Left" wing and one from "Right" wing) and then asked 3 (three) humans to write a a natural language description which conveys the semantic intersection of the information present in both narratives describing each event.

After the first round of annotation, we immediately observed that there was a discrepancy among the three annotators in terms of the *real* definition of "semantic intersection". For example, one annotator argued that *Semantic Intersection* of two narratives is non-empty as long as there is an overlap along one of the *5W1H* facets (Who, What, When, Where, Why and How), while another annotator argued that overlap in only one facet is not enough to decide whether there is indeed a semantic intersection. As an example, one of the annotators wrote only "Donald Trump" as the *Semantic Intersection* for a couple of cases where the narratives were substantially different, while others had those cases marked as "empty set".

To mitigate this issue, we only retained the narrative-pairs where at least two of the annotators wrote minimum 15 words as their ground-truth semantic intersection, with the hope that a human written description will contain 15 words or more only in cases where there is indeed a "significant" overlap between the two original narratives. This filtering step gave us 137 testing-samples at the end where each sample had 4 ground-truth semantic intersections, *one* from AllSides and *three* from human annotators.

6 Evaluating MNSI Task using ROUGE

As *ROUGE* (Lin, 2004) is the most popular metric used today for evaluating automated summarization techniques; we first conducted a case-study with *ROUGE* as the evaluation metric for the MNSI task.

6.1 Methods Used in the Case-Study

We experimented with multiple SoTA pre-trained abstractive summarization models as a proxy for *Semantic-Intersection* generator. These models are: 1. **BART** (Lewis et al., 2019), fine tuned on CNN and multi english Wiki news datasets [4], 2. **Pegasus** (Zhang et al., 2019), fine tuned on CNN and Daily mail dataset [5], and 3. **T5** (Raffel et al., 2019) fine tuned on multi english Wiki news dataset [6]. As our primary goal is to construct a benchmark dataset for the *MNSI* task and establish an appropriate metric for evaluating this task, experimenting with only 3 abstractive summarization models is not a barrier to our work. Proposing a custom method fine-tuned for the *Semantic-Intersection* task is an

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

[3]The dataset and manual annotations can be found in supplementary folder.

[4]WikinewsSum/bart-large-cnn-multi-en-wiki-news

[5]google/pegasus-cnn_dailymail

[6]WikinewsSum/t5-base-multi-en-wiki-news

orthogonal goal to this work and we leave it as a future work. Also, we'll use the phrases "summary" and "intersection-summary" interchangeably from here. To generate the summary, we concatenate a narrative pair and feed it directly to the model.

For evaluation, we first evaluated the machine generated intersection summaries for the 137 manually annotated testing samples using the rouge metric (Lin, 2004) and follow the procedure mentioned in the paper to compute the ROUGE-F_1 scores 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 4 F_1 scores and averaged them out across the test dataset. The raw rouge scores can be seen in the table 11 in appendix.

6.2 Results and Findings

We computed Pearson's correlation coefficients between each pair of Rouge-F_1 scores obtained using all of the 4 reference intersection-summaries (3 human written summary and 1 AllSides theme description) to test the robustness of *ROUGE* metric for evaluating the *MNSI* task. The corresponding correlations are shown in table 2. For each annotator pair, we report the 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 ROUGE metric is not stable across multiple human-written intersection-summaries and thus, *unreliable*. Indeed, only one out the 6 different annotator pairs has a value greater than 0.50 for all the 3 Rouge metrics (R1, R2, RL), which is *problematic*.

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 2: 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 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.

7 Can We Do Better than ROUGE?

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

7.1 Different trials of Human Judgement

Assigning a Single Numeric Score: As an initial trial, we decided to first label 25 testing samples using two human annotators (we call them label annotators L_1 and L_2). Both label-annotators read each of the 25 narrative pairs as well as the corresponding system generated intersection-summary (generated by fine-tuned BART) and assigned a numeric score between 1-10 (inclusive). This number reflects their judgement/confidence about how accurately the system-generated summary captures the *actual* intersection of the two input narratives. *Note that, the reference intersection summaries were not included in this label annotation process and the label-annotators judged the system-generated summary exclusively with respect to the input narratives*. To quantify the agreement between human scores, we computed the Kendall rank correlation coefficient (or Kendall's Tau) between two annotator labels since these are ordinal values. However, to our disappointment, the correlation value was 0.20 with p-value being 0.22[7]. This shows that even human annotators are disagreeing among themselves and we need to come up with a better labelling guideline to reach a reasonable agreement among the human annotators.

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

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

[7]The higher p-value means that the correlation value is insignificant because of the small number of samples, but the aim is to first find a labelling criterion where human can agree among themselves.

to compute F_1 score), we compute the Kendall's rank correlation coefficient among the precision scores and recall scores of all the annotator pairs separately. The corresponding correlation values can be seen in the table 3. As we notice, there is definitely some improvement in agreement among annotators compared to the one number annotation in 7.1, however, the average correlation is still 0.33 and 0.41 for precision and recall respectively, much lower than the 0.5.

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

Table 3: Kendall's rank correlation coefficients among the the precision and recall scores for pairs of human annotators (25 test samples). Here L_i refers to the i^{th} label annotator.

7.2 Sentence-wise Scoring

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

Overlap Labels: Label-annotators (L_1, L_2 and L_3) were asked to look at a machine-generated sentence 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, if the human feels there is more than 75% overlap (between each system-generated sentence and 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 total $3 \times 506 = 1,518$ sentence-level ground-truth labels.

To create the overlap labels from precision perspective as described above, we concatenated all the 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 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 sentences in each of the 4 reference summaries separately (provided by (I_1, I_2, I_3 and I_4)), with respect to the machine generated summary.

Human agreement in terms of Kendall's Rank Correlation				
	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 4: Average precision and recall Kendall rank correlation coefficients between sentence-wise annotation for different annotators. L_i refers to the i^{th} label annotator. All values are statistically significant (p<0.05).

Inter-Rater-Agreement: We use the Kendall rank correlation coefficient to compute the agreement among the ordinal labels assigned by human label annotators. Since there can be multiple sentences in the system generated or the reference summary, we flatten out the sentence labels and concatenate them for the entire dataset. To compute the Kendall Tau, we map the ordinal labels to numerical values using the mapping: $\{P: 1, PP: 0.5, A: 0\}$. As we can notice in table 4, inter-annotator correlation for both precision and recall are ≥ 0.50 and thus, signifying higher agreement among label annotators.

Reward-based Inter-Rater-Agreement: Alternatively, we first define a reward matrix (Table 5) 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 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 6. Both precision and recall reward scores are high (≥ 0.70) for all the different annotator pairs, thus signifying, high inter label-annotator agreement.

We believe, one of the reasons for higher reward scores could be that sentence-wise labelling puts less cognitive load on human mind in contrast to the single or double score(s) for the entire intersection summary and accordingly, shows high agreement in terms of human interpretation. Similar observation is also noted in Harman and Over (2004).

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 5: Reward function used to evaluate the labels assigned by two label annotators (or labels inferred using SEM-F1 metric and human annotated labels) for a given sentence. It acts as a form of correlation between annotator pairs.

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

Table 6: 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.

8 Semantic-F1: The New 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-F1**. The details of our **SEM-F1** metric are described in algorithm 1 and the respective notations are mentioned in table 7. 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 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-F1** is to compute the sentence-wise similarity (e.g., cosine similarity using a sentence embedding model) to infer the semantic overlap/intersection between two sentences from both precision and recall perspective and then, combine them into F_1 score.

Notations	Description
S_G	Machines generated summary
S_R	Reference summary
$T := (t_l, t_u)$	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 7: Table of notations for algorithm 1

8.1 Is SEM-F1 Reliable?

The SEM-F1 metric computes cosine similarity scores between sentence-pairs from both precision and recall perspectives. To see whether SEM-F1 metric correlates with human-judgement, we further converted the sentence-wise raw cosine scores into *Presence* (P), *Partial Presence* (PP) and *Absence* (A) labels using some user-defined thresholds as described in algorithm 2. This helped us to directly compare the SEM-F1 inferred labels against the human annotated labels.

As mentioned in section 8, we utilized state-of-the-art sentence embedding models to encode sentences from both the model generated summaries and the human written narrative intersections. To be more specific, we experimented with 3 sentence embedding 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 predict the sentence-wise *presence (P)*, *partial presence (PP)* and *absence (A)* labels to 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, threshold range $(45, 75)$ means that if similarity score $< 45\%$, infer label "absent", else if similarity score $\geq 75\%$, infer label "present" and else, infer label "partial-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-F1 metric. For this, we repeat the proce-

Algorithm 1 Semantic-F1 Metric

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

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

Reward/Kendall		Machine-Human Agreement in terms of Reward Function						
		$\mathbf{T = (25, 75)}$	$\mathbf{T = (35, 65)}$	$\mathbf{T = (45, 75)}$	$\mathbf{T = (55, 65)}$	$\mathbf{T = (55, 75)}$	$\mathbf{T = (55, 80)}$	$\mathbf{T = (60, 80)}$
Embedding: P-v1	**Precision**	0.75/0.57	0.8/0.63	0.76/0.59	0.8/0.63	0.78/0.6	0.74/0.6	0.73/0.58
	Recall	0.66/0.54	0.76/0.64	0.73/0.66	0.72/0.64	0.69/0.63	0.65/0.64	0.61/0.6
Embedding: STSB	**Precision**	0.73/0.6	0.73/0.62	0.73/0.6	0.73/0.62	0.73/0.63	0.73/0.59	0.73/0.58
	Recall	0.63/0.55	0.64/0.63	0.63/0.6	0.65/0.61	0.65/0.61	0.63/0.61	0.64/0.59
Embedding: USE	**Precision**	0.76/0.64	0.76/0.66	0.78/0.64	0.78/0.64	0.79/0.63	0.78/0.62	0.79/0.65
	Recall	0.63/0.53	0.66/0.6	0.67/0.58	0.68/0.61	0.67/0.62	0.64/0.62	0.65/0.61

Table 8: Average Precision and Recall correlation (Reward score/Kendall correlation) between label-annotators (L_i) and automatically inferred labels using SEM-F1 (average of 3 label annotators). The raw numbers for each annotators can be found in appendix (table 12). The results are shown for different embedding models (8.1) and multiple threshold levels $T = (t_l, t_u)$. Moreover, the both the Reward and Kendall values are consistent/stable across all the 5 embedding models and threshold values.

dure of Table 6, but this time comparing human labels against 'SEM-F1 labels'. The corresponding results are shown in Table 8. As we can notice, the average reward values are consistently high (≥ 0.50) for all the 3 label-annotators (L_i). Moreover, the reward values are consistent/stable across all the 3 embedding models and threshold values, signifying that SEM-F1 is indeed robust across various sentence embeddings and threshold used.

Following the procedure in table 4, we also compute the Kendall's Tau between human label annotators and automatically inferred labels using SEM-F1. Our results in table 8 are consistent with reward-based inter-rater-agreement and the correlation values are ≥ 0.50 with little variation along various thresholds for both precision and recall.

8.2 SEM-F1 Scores for Random Baselines

Here, we present the actual SEM-F1 scores for the three models described in section 6.1 along with scores for two intuitive baselines, namely, 1) Random Intersection 2) Random Annotation.

Random Intersection: For a given sample and model, we select a random intersection summary generated by the model out of the other 136 test samples. There random intersections are then evaluated using SEM-F1 against 4 reference summaries.

Random Annotation: For a given sample, we select a random reference summary out of the other 4 references among the other 136 test samples. The model generated summaries are then compared against these Random Annotations/References to compute SEM-F1 scores as reported in table 9.

As we notice, there is approximately 40-45 percent improvement over the baseline scores suggesting SEM-F1 can indeed distinguish *good* from *bad*.

8.3 Pearson Correlation for SEM-F1

Following the case-study based on Rouge in section 6, we again compute the Pearson's correlation coefficients between each pair of raw SEM-F1 scores

	Random Annotation SEM-F1 Scores			Random Intersection SEM-F1 Scores			SEM-F1 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 9: SEM-F1 Scores

obtained using all of the 4 reference intersection-summaries. The corresponding correlations are shown in table 10. 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. This shows a clear improvement over the ROUGE metric suggesting that SEM-F1 is more accurate than ROUGE metric.

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

Table 10: Max (across 3 models) Pearson's correlation between the SEM-F1 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.

9 Conclusions

In this work, we proposed a new NLP task, called Multi-Narrative Semantic Intersection (*MNSI*) and created a benchmark dataset through meticulous human effort to initiate a new research direction. As a starting point, we framed the problem as a constrained summarization task and showed that *ROUGE* is not a reliable evaluation metric for this task. We further proposed a more accurate metric, called *SEM-F1*, for evaluating *MNSI* task. Experiments show that SEM-F1 is more robust and yield higher agreement with human judgement.

References

Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogerio Feris, Raja Giryes, and Alex M Bronstein. 2019. Laso: Label-set operations networks for multi-label few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6548–6557.

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

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

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 autoencoders. *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 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. *ACM SIGKDD Explorations Newsletter*, 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*.

Dragomir Radev. 2000. A common theory of information fusion from multiple text sources step one: cross-document structure. In *1st SIGdial workshop on Discourse and dialogue*, pages 74–83.

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 Other definitions of Text Intersection

Below, we present a set of possible definitions of *Semantic Intersection* to encourage the readers to think more about other alternative definitions.

1. On a very simplistic level, one can think of *Semantic Intersection* to be just the common words between the two input documents. One can also include their frequencies of occurrences in such representation. More specifically, we can define D_{int} as a set of unordered pairs of words w_i and their frequencies of common occurrences f_i, i.e., $D_{int} = \{(w_i, f_i)\}$. We can further extend this approach such that *Semantic Intersection* is a set of common n-grams among the input documents. More specifically, $D_{int} = \{\big((w_1, w_2, ..., w_n)_i, f_i\big)\}$ such that the n-grams, $(w_1, w_2, ..., w_n)_i$, is present in both D_A (with frequency f_{iA}) and D_B (with frequency f_{iB}) and $f_i = min(f_{iA}, f_{iB})$.
2. Another way to think of *Semantic Intersection* is to find the common topics among two documents just like finding common object labels among two images (Alfassy et al., 2019), by computing the joint probability of their topic distributions. More specifically, *Semantic Intersection* can be defined by the following joint probability distribution: $P(T_i|D_{int}) = P(T_i|D_A) \times P(T_i|D_B)$. This representation is more semantic in nature as it can capture overlap in topics.
3. Alternatively, one can take the *5W1H* approach (Xie et al., 2008), where a given narrative D can be represented in terms of unordered sets of six facets: *5Ws* (Who, What, When, Where and Why) and *1H* (How). In this case, we can define *Semantic Intersection* as the common elements between the corresponding sets related to these 6 facets present in both narratives, i.e. $D_{int} = \{S_i\}$ where S_i is a set belonging to one of the six *5W1H* facets. It is entirely possible that one of these S_i's is an empty set (ϕ). The most challenging aspect with this approach is accurately inferring the 5W1H facets.
4. Another way could be to define a given document as a graph. Specifically, we can consider a document D as a directed graph $G = (V, E)$ where V represents the vertices and E represents the edges. Thus, *TextIntersect* can be defined as the set of common vertices or edges or both. Specifically, D_{int} can be defined as a maximum common subgraph of both G_A and G_B, where G_A and G_B are the corresponding graphs for the documents D_A and D_B respectively. However, coming up with a graph structure G which can align with both documents D_A and D_B, would itself be a challenge.
5. One can also define *TextIntersect* operator ($\cap$) between two documents based on historical context and prior knowledge. Given a knowledge base K, $D_{int} = \cap(D_A, D_B|K)$ (Radev, 2000).

All the approaches defined above have their specific use-cases and challenges, however, from a human-centered point of view, they may not reflect how humans generate semantic intersections. A human would mostly express it in the form of natural language and this is why, we frame the *TextIntersect* operator as a constraint summarization problem such that the information of the output summary is present in both the input documents.

B Threshold Algorithm

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
```

C Rouge Scores

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 11: Average Rouge-F1 Scores for all the test models across test dataset. For a particular sample, we take the maximum value out of the 4 F1 scores corresponding to the 4 reference summaries.

		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

(a) Average Precision and Recall reward/correlation (mean $\pm$ std) between label-annotators (L_i) and automatically inferred labels using SEM-F1. The results are shown for different embedding models (8.1) 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 the 5 embedding models and threshold values.

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

(b) Average Precision and Recall Kendall Tau between label-annotators (L_i) and automatically inferred labels using SEM-F1. The results are shown for different embedding models (8.1) 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 the 5 embedding models and threshold values. All values are statistically significant at p-value<0.05.

Table 12: Machine-Human Agreement

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

Table 13: Two input documents are concatenated to compute the statistics. Four numbers for reference (#words/#sents) in Test split corresponds to the 4 reference intersections.