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.
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, the 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 parallel 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 $D_{\text{int}} \leftarrow D_A \cap_T D_B$. 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, \textit{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 \textit{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 \textit{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 \textit{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).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>theme</td>
<td>headlines by AllSides</td>
</tr>
<tr>
<td>theme-description</td>
<td>news description by AllSides</td>
</tr>
<tr>
<td>right/left head</td>
<td>right/left news headline</td>
</tr>
<tr>
<td>right/left context</td>
<td>right/left news description</td>
</tr>
</tbody>
</table>

Table 1: Overview of dataset scraped from AllSides

5 \textbf{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 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 \textit{real} definition of “semantic intersection”. For example, one annotator argued that \textit{Semantic Intersection} of two narratives is non-empty as long as there is an overlap along one of the 5W1H facets (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 \textit{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 \textit{ROUGE} (Lin, 2004) is the most popular metric used today for evaluating automated summarization techniques; we first conducted a case-study with \textit{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 \textit{Semantic-Intersection} generator. These models are: 1. \textbf{BART} (Lewis et al., 2019), fine tuned on CNN and multi english Wiki news datasets \(^4\), 2. \textbf{Pegasus} (Zhang et al., 2019), fine tuned on CNN and Daily mail dataset \(^5\), and 3. \textbf{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 \textit{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 \textit{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.

\(^4\)\textit{WikinewsSum/bart-large-cnn-multi-en-wiki-news}

\(^5\)\textit{google/pegasus-cnn_dailymail}

\(^6\)\textit{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 correlation pairs 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 \(R_1\), \(R_2\) 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 (\(R_1\), \(R_2\), RL), which is problematic.

<table>
<thead>
<tr>
<th>Pearson’s Correlation Coefficients</th>
<th>(R_1)</th>
<th>(R_2)</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I_1)</td>
<td>0.62</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>(I_2)</td>
<td>0.3</td>
<td>0.38</td>
<td>—</td>
</tr>
<tr>
<td>(I_3)</td>
<td>0.07</td>
<td>0.27</td>
<td>0.69</td>
</tr>
<tr>
<td>(I_4)</td>
<td>0.17</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>Average</td>
<td>0.36</td>
<td>0.33</td>
<td>0.38</td>
</tr>
</tbody>
</table>

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 (\(\leq 0.50\)), 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 0.227. 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
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.

<table>
<thead>
<tr>
<th>Human agreement in terms of Kendall Tau</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_1$</td>
<td>0.52</td>
<td>0.37</td>
</tr>
<tr>
<td>$L_2$</td>
<td>0.18</td>
<td>0.29</td>
</tr>
<tr>
<td>Average</td>
<td>0.33</td>
<td>0.41</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Human agreement in terms of Kendall’s Rank Correlation</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_1$</td>
<td>0.68</td>
<td>0.75</td>
</tr>
<tr>
<td>$L_2$</td>
<td>0.59</td>
<td>0.64</td>
</tr>
<tr>
<td>$L_3$</td>
<td>0.49</td>
<td>0.71</td>
</tr>
<tr>
<td>Average</td>
<td>0.64</td>
<td>0.72</td>
</tr>
</tbody>
</table>

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 $\geq 0.50$ 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 ($\geq 0.70$) 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).
The SEM-F1 metric computes cosine similarity 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 (STSBI) (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 ≥ 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 (Li) and the labels inferred by SEM-F1 metric. For this, we repeat the proce-
Table 8: Average Precision and Recall correlation (Reward score/Kendall correlation) between label-annotators (L) and automatically inferred labels using SEM-F1 (average of 3 label annotators). The raw numbers for each annotator can be found in appendix (table 12). The results are shown for different embedding models (8.1) and multiple threshold levels \(T = (t_1, t_2)\). Moreover, the both the Reward and Kendall values are consistent/stable across all the 5 embedding models and threshold values.

<table>
<thead>
<tr>
<th>Embedding/Model</th>
<th>Precision</th>
<th>Recall</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-V1 STSB USE</td>
<td>0.75/0.57</td>
<td>0.66/0.54</td>
<td>0.72/0.53</td>
<td>0.63/0.53</td>
</tr>
<tr>
<td>T5 STSB USE</td>
<td>0.64/0.55</td>
<td>0.73/0.64</td>
<td>0.65/0.61</td>
<td>0.67/0.61</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>SEM-F1 Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-V1</td>
<td>0.65 0.52</td>
</tr>
<tr>
<td>STSB</td>
<td>0.58 0.60</td>
</tr>
<tr>
<td>USE</td>
<td>0.61 0.52</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pearson’s Correlation Coefficients</th>
<th>P-V1</th>
<th>STSB</th>
<th>USE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i_1)</td>
<td>0.69</td>
<td>0.65</td>
<td>0.71</td>
</tr>
<tr>
<td>(i_2)</td>
<td>0.40</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>(i_3)</td>
<td>0.33</td>
<td>0.44</td>
<td>0.37</td>
</tr>
<tr>
<td>Average</td>
<td>0.49</td>
<td>0.49</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Table 10: Max (across 3 models) Pearson’s correlation between the SEM-F1 scores corresponding to different annotators. Here \(i\) 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.

8
References


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, $D_{int} = \{((w_1, w_2, ..., w_n), f_i)\}$ such that the n-grams, $(w_1, w_2, ..., w_n)$, 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 5WH 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 5WH facets. It is entirely possible that one of these $S_i$’s is an empty set ($\phi$). The most challenging aspect with this approach is accurately inferring the 5WH 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

1: procedure Threshold(rawSs, T)
2: initialize Labels ← []
3: for each element $e$ in rawSs do
4: if $e \geq t_o$% then
5: Labels.append($P$)
6: else if $t_i$% $\leq e \leq t_o$% then
7: Labels.append($PP$)
8: else
9: Labels.append($A$)
10: end if
11: end for
12: return Labels
13: end procedure

C Rouge Scores

<table>
<thead>
<tr>
<th>Model</th>
<th>R1</th>
<th>R2</th>
<th>RL</th>
</tr>
</thead>
<tbody>
<tr>
<td>BART</td>
<td>40.73</td>
<td>25.97</td>
<td>29.95</td>
</tr>
<tr>
<td>T5</td>
<td>38.50</td>
<td>24.63</td>
<td>27.73</td>
</tr>
<tr>
<td>Pegasus</td>
<td>46.36</td>
<td>29.12</td>
<td>37.41</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>T</th>
<th>(25, 75)</th>
<th>(35, 65)</th>
<th>(45, 75)</th>
<th>(55, 65)</th>
<th>(55, 75)</th>
<th>(55, 80)</th>
<th>(60, 80)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentence Embedding: P-v1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Precision Reward</strong></td>
<td>L₁</td>
<td>0.73 ± 0.27</td>
<td>0.81 ± 0.25</td>
<td>0.77 ± 0.26</td>
<td>0.85 ± 0.23</td>
<td>0.80 ± 0.24</td>
<td>0.77 ± 0.24</td>
</tr>
<tr>
<td></td>
<td>L₂</td>
<td>0.72 ± 0.30</td>
<td>0.73 ± 0.29</td>
<td>0.73 ± 0.30</td>
<td>0.78 ± 0.27</td>
<td>0.79 ± 0.27</td>
<td>0.75 ± 0.26</td>
</tr>
<tr>
<td></td>
<td>L₃</td>
<td>0.81 ± 0.23</td>
<td>0.86 ± 0.21</td>
<td>0.79 ± 0.24</td>
<td>0.78 ± 0.28</td>
<td>0.74 ± 0.28</td>
<td>0.69 ± 0.28</td>
</tr>
<tr>
<td><strong>Recall Reward</strong></td>
<td>L₁</td>
<td>0.66 ± 0.19</td>
<td>0.79 ± 0.16</td>
<td>0.75 ± 0.16</td>
<td>0.76 ± 0.18</td>
<td>0.71 ± 0.17</td>
<td>0.66 ± 0.17</td>
</tr>
<tr>
<td></td>
<td>L₂</td>
<td>0.67 ± 0.19</td>
<td>0.78 ± 0.16</td>
<td>0.76 ± 0.15</td>
<td>0.73 ± 0.19</td>
<td>0.72 ± 0.18</td>
<td>0.70 ± 0.18</td>
</tr>
<tr>
<td></td>
<td>L₃</td>
<td>0.66 ± 0.15</td>
<td>0.72 ± 0.17</td>
<td>0.68 ± 0.17</td>
<td>0.68 ± 0.22</td>
<td>0.64 ± 0.20</td>
<td>0.59 ± 0.19</td>
</tr>
</tbody>
</table>

(a) Average Precision and Recall reward/correlation (mean ± std) between label-annotators (Lᵢ) and automatically inferred labels using SEM-F1. The results are shown for different embedding models (8.1) and multiple threshold levels T = (Tᵢ, Tᵣ). For all the annotators Lᵢ (i ∈ {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

<table>
<thead>
<tr>
<th>T</th>
<th>(25, 75)</th>
<th>(35, 65)</th>
<th>(45, 75)</th>
<th>(55, 65)</th>
<th>(55, 75)</th>
<th>(55, 80)</th>
<th>(60, 80)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentence Embedding: P-v1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Precision Reward</strong></td>
<td>L₁</td>
<td>0.55</td>
<td>0.6</td>
<td>0.58</td>
<td>0.59</td>
<td>0.57</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>L₂</td>
<td>0.61</td>
<td>0.67</td>
<td>0.63</td>
<td>0.67</td>
<td>0.64</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>L₃</td>
<td>0.54</td>
<td>0.62</td>
<td>0.56</td>
<td>0.64</td>
<td>0.6</td>
<td>0.56</td>
</tr>
<tr>
<td><strong>Recall Reward</strong></td>
<td>L₁</td>
<td>0.53</td>
<td>0.64</td>
<td>0.66</td>
<td>0.62</td>
<td>0.61</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>L₂</td>
<td>0.55</td>
<td>0.64</td>
<td>0.67</td>
<td>0.63</td>
<td>0.63</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>L₃</td>
<td>0.54</td>
<td>0.65</td>
<td>0.64</td>
<td>0.66</td>
<td>0.65</td>
<td>0.65</td>
</tr>
</tbody>
</table>

(b) Average Precision and Recall Kendall Tau between label-annotators (Lᵢ) and automatically inferred labels using SEM-F1. The results are shown for different embedding models (8.1) and multiple threshold levels T = (Tᵢ, Tᵣ). For all the annotators Lᵢ (i ∈ {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
<table>
<thead>
<tr>
<th>Split</th>
<th>#words (docs)</th>
<th>#sents (docs)</th>
<th>#words (reference/s)</th>
<th>#sents (reference/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>1613.69</td>
<td>66.70</td>
<td>67.30</td>
<td>2.82</td>
</tr>
<tr>
<td>Test</td>
<td>959.80</td>
<td>44.73</td>
<td>65.46/38.06/21.72/32.82</td>
<td>3.65/2.15/1.39/1.52</td>
</tr>
</tbody>
</table>

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.