Revisiting Automatic Evaluation of Extractive Summarization Task: Can We Do Better than ROUGE?

Mousumi Akter, Naman Bansal, Shubhra Kanti Karmaker Santu
BDI Lab, Auburn University, Alabama, USA
{mza0170, nzb0040, sks0086}@auburn.edu

Abstract

It has been the norm for a long time to evaluate automated summarization tasks using the popular *ROUGE* metric. Although several studies in the past have highlighted the limitations of *ROUGE*, researchers have struggled to reach a consensus on a better alternative until today. One major limitation of the traditional *ROUGE* metric is the lack of semantic understanding (relies on direct overlap of n-grams). In this paper, we exclusively focus on the extractive summarization task and propose a semantic-aware *nCG* (normalized cumulative gain)-based evaluation metric (called *Sem-nCG*) for evaluating this task. One fundamental contribution of the paper is that it demonstrates how we can generate more reliable semantic-aware ground truths for evaluating extractive summarization tasks without any additional human intervention. To the best of our knowledge, this work is the first of its kind. We have conducted extensive experiments with this new metric using the widely used CNN/DailyMail dataset. Experimental results show that the new *Sem-nCG* metric is indeed semantic-aware, shows higher correlation with human judgement (more *reliable*) and yields a large number of disagreements with the original *ROUGE* metric (suggesting that *ROUGE* often leads to inaccurate conclusions also verified by humans).

1 Introduction

Text summarization is a difficult NLP task and an automatic evaluation of this task is even more challenging. However, automatic evaluation is vital for large-scale experiments as it acts as a replacement for time consuming and pricey human evaluation. As such, the reliability and robustness of automatic evaluation is very crucial.

The most commonly used metric for evaluating text summarization is *ROUGE* (Lin, 2004). Although *ROUGE* has been criticized for considering direct lexical overlap and thus not being semantic-aware, the majority of summarization models' assessments today are still based on *ROUGE* scores. In this paper, we revisit the popular *ROUGE* metric exclusively in the context of evaluating *extractive* summarization task, a task where phrases and sentences from the original text are extracted to create a summary. As such, if the human-written summary includes more novel words than the original document, ROUGE will provide a poor score to extractive summaries due to a lack of semantic awareness. Another limitation of the *ROUGE* metric in the context of *extractive* summarization is the following: while the *extractive* summarization task is generally framed as a sentence ranking problem, the *ROUGE* metric was not originally proposed for evaluating the quality of a ranker. Indeed, the heavily used technique behind *extractive* summarization is to rank sentences from the original document according to how well they reflect the overall description and then create a summary by concatenating the top-ranked sentences. Thus, the "right" evaluation metric for the *extractive* summarization task should also consider the quality of the sentence ranker. Again think about a human-written summary which is highly abstractive in nature. A good ranker that ranks the most informative sentences at the top may still suffer from low ROUGE scores due to fewer direct lexical overlaps between the system summary and human-written summary.

To address these limitations, we propose an alternative *gain*-based evaluation metric in this paper (called *Sem-nCG*) for evaluating extractive summarization tasks, which is both 1) semantic-aware and 2) rewards a system-generated summary based on some groundtruth ranking of sentences from original document. *nCG* (Normalized cumulative gain) is a widely used metric for evaluating the performance of ranking systems, especially when conducting multi-level relevance judgements (Järvelin and Kekäläinen, 2002). Although *nCG* evaluation

Findings of the Association for Computational Linguistics: ACL 2022, pages 1547 - 1560

is not entirely new (Karmaker Santu et al., 2017; Kuzi et al., 2019; Karmaker et al., 2020), one fundamental contribution of this paper is that it demonstrates how we can automatically generate a reliable semantic-aware groundtruth ranking of sentences within a source document, which essentially enables automatic *Sem-nCG* based evaluation without any additional human intervention. To the best of our knowledge, this work is the first of its kind. To be more specific, given an original document and a human-written summary for evaluation purposes, we used several state-of-the-art sentence embedding techniques (including InferSent, Sentence Transformer, Elmo, Google Universal Sentence Encoding and their ensemble) to prepare groundtruth ranking of sentences from original document by computing semantic similarity between each individual sentence of original document and entire human written summary. Finally, this groundtruth ranking is compared against model-inferred ranking to compute *Sem-nCG* score, where a higher number means a better extractive summary.

We have conducted extensive experiments with this new metric using the CNN/DailyMail dataset and 6 state-of-the-art extractive summarization models ($\text{BERT}_{\text{base}}$, MobileBERT, DistilBERT, RoBERTa, XLNET, GPT-2). Experimental results show that the new *Sem-nCG* metric is: 1) semantic-aware, 2) shows higher correlation with human judgement (more reliable), and 3) yields a large number of disagreements with the original *ROUGE* metric (suggesting *ROUGE* often leads to inaccurate conclusions). When cross-examined by humans, we found *Sem-nCG* to be more accurate (62% of the time) than *ROUGE* on average where the two metrics disagreed on the relative performance of a pair of *extractive* summarization models. Thus, in response to the question of whether we can do better than *ROUGE* for evaluating extractive summarization tasks, the answer appears to be "YES".

2 Related Work

Evaluation of the text summarization task is challenging and has been studied vastly in the past. (Radev and Tam, 2003) proposed the *Relative Utility* (RU) metric, which evaluates extractive summarization as a ranking task (similar to our formulation), but has not gained much popularity, because their approach requires manual labor to rank each sentence of a document, and it is not practical to manually annotate such large data-sets.

ROUGE (Lin, 2004) is perhaps the most popular metric used today for the evaluation of the automated summarization techniques, mainly because it is a simple and automatic process. However, *ROUGE* has been criticized a lot for primarily relying on lexical overlap (Nenkova, 2006) of n-grams. Later, (Zhou et al., 2006) suggested using a broad domain-independent paraphrase table derived from a bilingual parallel corpus to enable paraphrase matching for summary evaluation. (Cohan and Goharian, 2016) showed that *ROUGE* suffers from poor performance in cases of terminology variation and paraphrasing. As of today, around 192 variants of *ROUGE* have been proposed (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, none of the variants of *ROUGE* considers the ranking quality (core technique of extractive summarization); let alone providing an automatic way to do it, which is the primary goal of our work.

Researchers have also proposed metrics alternative to *ROUGE*: factoids-based (atomic information units for sentence meaning) (Teufel and van Halteren, 2004) and pyramid-based (Nenkova and Passonneau, 2004) approaches are two of them. Multiple different reference summaries are a must for both approaches, where the pyramid-based approach requires additional manual labor to construct the pyramid. Since the pyramid must be built by hand and gives imprecise scores, this technique failed to gain much attraction. Many enhancements have been made to the pyramid-based approach: precise automated system for calculating pyramid ratings (Passonneau et al., 2013), pyramid evaluation via automated information extraction (Yang et al., 2016), lightweight sampling-based version that is crowdsourcable (Shapira et al., 2019) and facet-aware evaluation (Mao et al., 2020) for better assessment of knowledge coverage in extractive summarization. Still, the pyramid-based approach necessitates significant additional manual labour making it less appealing for large-scale evaluation.

Researchers also attempted to develop methods for evaluating reference-free model summaries (Louis and Nenkova, 2013; Xenouleas et al., 2019). Distance measures between the system summary and reference summary based on word

embeddings have also been proposed (Zhao et al., 2019; Sun and Nenkova, 2019). Moreover, model based evaluation for text generation (also adopted for text summarization) has also been a recent trend (Sellam et al., 2020; Zhang et al., 2020; Yuan et al., 2021). Yet, none of these metrics explicitly assess the quality of ranking performed by an *extractive* summarization method.

3 Background

nCG (Normalized Cumulative Gain) is a popular measure for evaluating information retrieval (IR) systems (Järvelin and Kekäläinen, 2002). Given a query and a ranked list of search results, computation of *nCG* involves summing the gains of the top k documents, and normalizing by the maximum possible gain that can be obtained for the query. Mathematically:

$$CG@k = \begin{cases} G@1, & \text{if } k = 1 \\ CG@[k-1] + G@k, & \text{otherwise} \end{cases} \quad (1)$$

Here, k is the cutoff position (e.g., $k = 5$ is a common choice), $G@k$ and $CG@k$ are the gain and cumulative gain, respectively, at the k-th position in the list. $nCG@k$ is $CG@k$ divided by the maximum achievable $CG@k$, also called Ideal CG ($ICG@k$), which is computed from the ideal ranking of the documents with respect to the query. The ideal ranking places the document(s) with the highest gain on the very top, followed by the documents with the next level of gain, etc. Mathematically:

$$nCG@k = \frac{CG@k}{ICG@k} \quad (2)$$

4 *Sem-nCG* for Extractive Summary

The main motivation for introducing the *Sem-nCG* metric is to ensure a fair evaluation of the *extractive* summarization task where the metric is both semantic-aware as well as captures the ranking quality of the extractive summarizer. Indeed, for *extractive* summarization, sentences in the original document are ranked based on how well they reflect the overall description, and thus, evaluating it with a rank-aware metric like *Sem-nCG* is more equitable. But, how can we develop a *Sem-nCG* metric for the extractive summarization task that was originally designed for *Information Retrieval* systems? What would be the query in this case? What would be the definition of a document? How do we define the gains? How can we compute the groundtruth ideal ranking? All of these are important questions we need to answer before one can use *Sem-nCG* evaluation for extractive summarization tasks.

Problem Formulation: We formulate *extractive* text summarization as a ranking problem, where the output is a ranked-list of sentences based on how well they convey the overall content of the original document. Let us assume that, input is a document $D = [S_D^1, S_D^2, S_D^3, S_D^4..., S_D^{|D|}]$, where S_D^i denotes i^{th} sentence of document D and output is the *Sem-nCG@k* score for the top-k sentences extracted from Document D by the *extractive* summarization model. Now, in order to compute *Sem-nCG@k*, we need to know what the gains of the top-k ranked sentences are, as well as the gains of the top-k ideal (desired) sentences. In other words, without knowing the groundtruth gains for each sentence in the original document, we cannot compute the *Sem-nCG@k* metric.

Groundtruth Gains: It is indeed a philosophical question to ask what should be the definition of gains in case of the extractive summarization. In this work, we define gain as the following:

Definition 4.1 *Given document D and a sentence s from D, gain of s with respect to D is proportional to the degree of how well s captures the overall semantic meaning of document D.*

One way to measure this capturing power is to ask human judges. However, human judgment in this case is problematic for multiple reasons as follows: 1) Human evaluation is time-consuming and expensive, 2) Some human raters have the tendency to give higher ratings than deserved, this is known as the *Leniency* problem, which results in higher variance (Harman and Over, 2004), 3) Natural language descriptions are noisy and ambiguous, which makes manual ordering of sentences by annotators even harder resulting in low inter-rater agreement. This is why we opted for an automated way to create the groundtruth gains without involving humans, as demonstrated by Algorithm 1.

Automatic Gain Computation: How can we automatically infer groundtruth gains in order to automate *Sem-nCG@k* computation? Fortunately, in most summarization benchmark datasets, one or more reference summaries written by humans are also provided along with the original documents. We leverage these human-written reference summaries to automatically infer groundtruth gains.

The exact process is presented in Algorithm 1, where we utilize the semantic similarity between

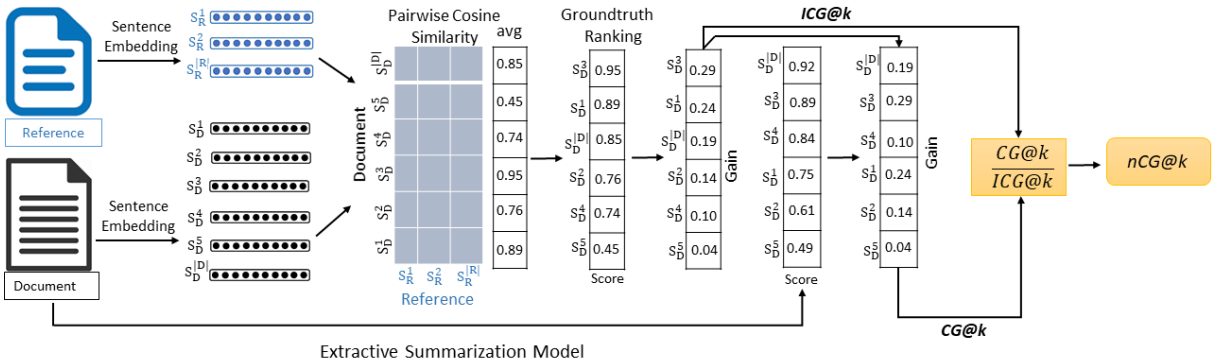

Figure 1: Pipeline for *Sem-nCG@k* evaluation of extractive summarization task, *CG@k* stands for Cumulative Gain at k[th] position and *ICG@k* for Ideal *CG@k*.

Algorithm 1 *Sem-nCG@k* Computation

INPUT: Document D, Reference R, Model's top-k extracted sentences, number of sentences in D as N
OUTPUT: *Sem-nCG@k* score

Phase 1: Groundtruth Gain Computation
$GT \leftarrow \{\}$
$GT_{gain} \leftarrow \{\}$
Represent sentences in D and R by embedding vectors
for each $S_D^i \in D$ **do**
  for each $S_R^j \in R$ **do**
    $Sim(S_D^i, S_R^j) \leftarrow Cosine\_Similarity(S_D^i, S_R^j)$
  end for
  $GT[S_D^i] \leftarrow mean(Sim)$
end for
$GT_{sorted} \leftarrow$ Sort GT based on $mean(Sim)$
$GT_{gain}[S_D^i] \leftarrow N - rank(S_D^i, GT_{sorted}) + 1$
Normalize GT_{gain} into a probabilistic gain
return GT_{gain}

Phase 2: *Sem-nCG@k* Computation
Compute *ICG@k* from GT_{gain}
$M \leftarrow$ Model's top-k extracted sentences
Retrieve M's gain from GT_{gain}
Compute *CG@k* for M
return *Sem-nCG@k* = $\frac{CG@k}{ICG@k}$

each sentence in the input document and the entire reference summary to generate groundtruth gains. For semantic similarity, we have experimented with different embeddings, including InferSent, Sentence Transformer, Elmo, Google Universal Sentence Encoding and their ensembles (details in section 5.3). Specifically, we measure the cosine similarity between each sentence in the original document and each reference sentence and then calculate an average cosine similarity for each source-sentence with respect to the whole reference. This average cosine similarity score is then used to rank all the sentences in the original document and a simple greedy approach is taken to assign the groundtruth gains as follows: sentences are assigned a groundtruth gain of $N, N-1, ..., 1$, sequentially from the top, where N denotes the number of sentences in the document. Later, the gain of each sentence is normalized to probabilistic scores ensuring the range of the *Sem-nCG* metric to be between 0 and 1. The intuition here is that a higher-ranked sentence gets more rewards than a lower-ranked one.

The gains computed by algorithm 1 are then used in equation 1 to compute the corresponding cumulative gain for ideal ranking (*ICG@k*) and for model's ranking (*CG@k*), respectively. The ratio of *CG@k* and *ICG@k*, which is *nCG@k* (equation 2), captures the quality of the system generated ranking with respect to the groundtruth ranking. Figure 1 visually demonstrates the pipeline for computing *Sem-nCG@k* metric.

5 Experimental Setup

5.1 Dataset

We conducted extensive experiments with our proposed *Sem-nCG* metric using the popular CNN/DailyMail (Hermann et al., 2015) benchmark dataset. The CNN/DailyMail dataset provides a collection of news articles and related highlights, and these highlights are used as a reference (gold summary). Also, the reference summaries are somewhat *extractive* in nature (a few bullet points providing a brief overview of the article) (Liu and Lapata, 2019). We collected the dataset from huggingface (Wolf et al., 2020)[1]. As we are not explicitly doing any training/fine-tuning of the summarizer models, we have only used the testing set for our experimental evaluation. We excluded any sample that has a sentence count less than 5 from our analysis as we report *Sem-nCG@5* scores. There were 64 such samples in the testing set, which brings

[1] https://huggingface.co/datasets/cnn_dailymail

our sample size to 11,426 (Details can be found in Table 1).

Feature	Description
Train/Validation/Test	287113/13368/11490
#Mean Tokens	781 per Article/56 per Highlights
Reference	Single
Strategy	Extractive

Table 1: Overview of CNN/DailyMail Dataset

5.2 *Extractive* Summarization Models

We collected six pre-trained models: $\text{BERT}_{\text{base}}$ (Liu and Lapata, 2019), MobileBERT (Sun et al., 2020), DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), GPT-2 (Radford et al., 2019), from huggingface (Wolf et al., 2020) that were fine-tuned on the CNN/DailyMail dataset for the *extractive* summarization task[2]. We then evaluated these six models using both our proposed *Sem-nCG@k* metric and traditional *ROUGE* metric.

5.3 Embedding Sensitivity

We recognize that the groundtruth gains we considered are not absolute since they are derived from a pre-trained sentence embedding. Therefore, we investigated the sensitivity of the gains by varying eight cutting-edge sentence embedding techniques. Specifically, we experimented with Infersent (v1&v2) (Conneau et al., 2017), Semantic Textual Similarity benchmark (STSb - bert/roberta/distilbert) (Reimers and Gurevych, 2019), Elmo (Peters et al., 2018) and Google Universal Sentence Encoder (USE) (Cer et al., 2018): i) enc-2 (Iyyer et al., 2015) based on the deep average network ii) enc-3 (Vaswani et al., 2017) based on transformers. We also created an ensemble method to aggregate the gains (in terms of raw similarity, rank and relevance) provided by different embeddings and combine them into a single gain with an expectation that the ensemble technique will provide a more reliable way for preparing the groundtruth gains. Furthermore, we have also experimented with 3 different variations of the ensemble technique: $\text{Ensemble}_{\text{sim}}$, $\text{Ensemble}_{\text{rank}}$ and $\text{Ensemble}_{\text{rel}}$, with the hope of obtaining more robust groundtruth gains. Specifically, $\text{Ensemble}_{\text{sim}}$ aggregates the cosine similarity first and then gives gains according to Algorithm 1, $\text{Ensemble}_{\text{rank}}$ generates a sentence ranking for each embedding variation and then aggregates the ranking to create a more robust ranking and then provide the gains according to Algorithm 1 and $\text{Ensemble}_{\text{rel}}$ calculates the gain first according to Algorithm 1 for all embedding variations and then takes an average over the gains. Please note that we compare sentences from original documents with highlights (written by humans) to prepare these groundtruth gains.

[2]Appendix contains model architecture details

6 Quantitative Evaluation

6.1 *ROUGE* is not Robust to Perturbation

One of the major criticisms of ROUGE is that it is not semantic-aware. Table 2 confirms that the ROUGE score highly varies if the original document is perturbed with synonyms [3]. Clearly, this is not desired from a "good" summary evaluation metric. Indeed, humans have various ways to express the same thing and often humans write summaries in their own words rather than picking the same key words from the original document (for example if the document uses "vacation", human references can have "trip", "tour", "break" etc.). For our experiments, we substituted around 20% of the words (excluding stop words) of the original document with their synonyms and computed ROUGE scores for these perturbed documents using the CNN/DailyMail dataset, assuming a 5-sentence summary. We utilized *wordnet*[4] from *nltk.corpus* to perform synonym replacement. As seen from Table 2, for ROUGE-1 and ROUGE-3, the score drop was around 5-7%, where for ROUGE-L it was around 3-5%. Interestingly, for ROUGE-2, the score drop was 5-16%.

As the groundtruth gain computation of *Sem-nCG* is dependent of embeddding techniques, we have also inspected whether the ROUGE variant with word embedding (ROUGE-we) (Ng and Abrecht, 2015) is also sensitive to perturbation. Interestingly, table 2 shows ROUGE-we scores are also sensitive to perturbation. For all ROUGE-we-{1,2,3}, the score drop was around 5-6%. One can reasonably expect that the score drop would be more significant if more words are replaced in original document ($> 20\%$).[5]

6.2 *Sem-nCG* is Robust to Perturbation

We have conducted the same experiment mentioned in Section 6.1 with *Sem-nCG* metric for

[3]The objective was to reduce the lexical overlap between extractive summary and reference. The reference can also be perturbed to do this experiment.

[4]www.nltk.org/howto/wordnet.html

[5]More evidence are included in Appendix

		$\text{BERT}_{\text{base}}$		MobileBERT		DistilBERT		RoBERTa		XLNet		GPT-2	
		Actual	Perturbed	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed
ROUGE-1	Precision	28.69	23.85	28.42	23.64	29.19	24.11	25.86	21.93	26.26	22.03	25.98	21.95
	Recall	65.04	57.79	62.8	56.02	65.37	58.07	57.93	52.08	57.53	51.5	57.68	51.4
	F1	38.62	32.74	37.86	32.18	39.18	33.08	34.69	29.93	34.96	29.9	34.71	29.77
ROUGE-2	Precision	13.47	8.82	12.93	8.49	13.95	9.07	10.56	7.01	10.77	7.03	10.64	6.96
	Recall	30.73	21.55	28.86	20.34	31.38	22.02	24.07	17.02	24.03	16.82	23.99	16.7
	F1	**18.15**	**2.13**	17.27	11.59	18.73	12.46	14.23	9.62	14.41	9.61	14.26	9.5
ROUGE-3	Precision	7.91	4.02	7.5	3.86	8.26	4.16	5.86	3.03	6	3.05	5.91	3.01
	Recall	17.85	9.72	16.56	9.14	18.42	10.02	13.31	7.35	13.35	7.29	13.28	7.22
	F1	10.61	5.5	9.97	5.24	11.04	5.69	7.87	4.15	8.01	4.16	7.9	4.1
ROUGE-L	Precision	17.62	14.63	17.37	14.28	18.25	14.9	16.05	13.37	16.34	13.39	16.12	13.39
	Recall	40.56	36.01	38.94	34.32	41.42	36.37	36.6	32.38	36.45	31.94	36.38	31.99
	F1	23.83	20.18	23.24	19.52	24.59	20.52	21.65	18.36	21.88	18.29	21.64	18.27
ROUGE-we-1	Precision	28.17	23.12	27.90	22.90	28.69	23.40	25.41	21.28	25.80	21.37	25.51	21.30
	Recall	63.73	55.96	61.51	54.20	64.12	56.30	56.77	50.46	56.37	49.91	56.48	49.82
	F1	37.90	31.73	37.13	31.16	38.48	32.10	34.06	29.02	34.33	29.00	34.06	28.89
ROUGE-we-2	Precision	18.18	13.39	17.65	13.03	18.70	13.68	15.25	11.48	15.50	11.51	15.32	11.46
	Recall	41.51	32.72	39.34	31.20	42.14	33.23	34.54	27.63	34.34	27.34	34.34	27.26
	F1	24.51	18.41	23.56	17.78	25.13	18.80	20.51	15.71	20.70	15.69	20.51	15.60
ROUGE-we-3	Precision	20.37	15.21	19.74	14.77	20.94	15.56	17.00	12.87	17.31	12.92	17.10	12.81
	Recall	47.22	37.70	44.67	35.86	47.90	38.31	39.09	31.42	38.94	31.09	38.92	30.91
	F1	27.56	20.98	26.45	20.22	28.24	21.45	22.95	17.67	23.20	17.66	22.98	17.50

Table 2: *ROUGE* and *ROUGE-we* scores (Precision, Recall and F1) for the extractive summarization models ($\text{BERT}_{\text{base}}$, MobileBERT, DistilBERT, RoBERTa, XLNet, GPT-2) on CNN/DailyMail test dataset. The results are for top-5 extracted sentences when the outputs are in actual and perturbed.

Embedding	$\text{BERT}_{\text{base}}$		MobileBERT		DistilBERT		RoBERTa		XLNet		GPT-2	
	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed
Infersent-v1	75.06	72.85	70.38	69.29	76.4	73.75	68.49	65.56	68.73	65.46	68.11	65.26
Infersent-v2	74.98	72.93	69.84	69.1	76.75	74.33	68.24	65.71	68.67	65.75	67.97	65.46
STSb-bert	75.46	74.68	70.8	70.47	76.99	76.76	68.99	66.88	69.81	67.37	68.98	66.79
STSb-roberta	75.23	74.53	70.72	70.45	76.69	76.33	69.02	66.97	69.77	67.4	69.04	66.8
STSb-distilbert	74.57	73.83	70.01	69.7	76.14	75.93	68.46	66.42	69.18	66.89	68.37	66.17
Elmo	74.64	70.3	69.72	67.64	75.91	70.76	68.03	64.91	68.83	64.55	67.89	64.77
USE-enc2	76.64	76.06	71.1	70.69	78.87	78.92	69.58	67.14	70.62	68.05	69.6	67.11
USE-enc3	76.03	74.96	70.17	69.37	78.14	77.73	68.16	65.72	69.14	66.49	68.08	65.66
$\text{Ensemble}_{\text{sim}}$	77.18	76.62	71.76	71.75	79.06	78.81	69.78	67.6	70.64	68.1	69.71	67.44
$\text{Ensemble}_{\text{rank}}$	77.15	76.41	71.81	71.8	78.94	78.37	69.74	67.55	70.53	67.91	69.63	67.34
$\text{Ensemble}_{\text{rel}}$	**78.74**	**78.93**	**73.54**	**74.48**	**80.47**	**80.85**	**71.74**	**70.81**	**72.5**	**71.17**	**71.62**	**70.64**
std	1.32	2.30	1.13	1.81	1.50	2.84	1.08	1.58	1.16	1.75	1.12	1.59

Table 3: *Sem-nCG@5* scores for the top-5 sentences of the extractive summarization models ($\text{BERT}_{\text{base}}$, MobileBERT, DistilBERT, RoBERTa, XLNet, GPT-2) on CNN/DailyMail test dataset for different embedding variations.

top-5 extracted sentences. As shown in Table 3, we can see that Ensemble techniques (especially for $\text{Ensemble}_{\text{rel}}$) show more robustness which is somewhat expected as it utilizes the benefits of multiple sentence embeddings. Among non-ensemble techniques, STSb-distilbert seems to be the most robust. If computational time is a bottleneck (Table 4), we would recommend utilizing the STSb-distilbert embedding for our proposed *Sem-nCG* metric.

6.3 *Sem-nCG* is Robust across Multiple Sentence Embedding Techniques

In this experiment, we tested the sensitivity of the proposed *Sem-nCG* metric with respect to the sentence embedding used to create the groundtruth gains. Specifically, we experimented with eight different sentence embedding techniques (Table 3). The findings reveal that the *Sem-nCG@k* score is stable across different sentence embeddings as evident from the low standard deviation of both *Sem-nCG@5* scores for the top-5 extracted sentences. Also, the relative performance of the models always remain same (DistilBERT $>$ $\text{BERT}_{\text{base}}$ $>$ MobileBERT $>$ XLNet $>$ GPT-2 $>$ RoBERTa) for all embedding variations.

6.4 *Sem-nCG* often disagrees with *ROUGE*

Although *ROUGE* and *Sem-nCG@k* agree on relative performances of multiple summarization models in the average case, as we explored further, we discovered that the agreement does not hold for individual document samples. As shown in Table 5, there is a considerable amount of disagreements between *ROUGE* and *Sem-nCG@k* for each pair of models. Here, disagreement means when comparing Model_{A} and Model_{B}, *Sem-nCG@k* in-

dicates Model_A's output is better, while *ROUGE* implies Model_B's output is better and vice-versa. To resolve these conflicts, we further involved humans to perform meta-evaluation of *ROUGE* and *Sem-nCG@k*, where human judgement agreed with *Sem-nCG@k* most of the time (see Section 7).

7 Human Evaluation

7.1 Human judgment favors *Sem-nCG* over *ROUGE* in case of disagreements

We next took a deeper look into the cases where *Sem-nCG* disagreed with *ROUGE* (Table 5) while comparing two extractive summarization models. We asked humans to blindly evaluate the quality of the summaries generated by two models and make a judgement on which summary was better as suggested by (Peyrard, 2019) as well. Specifically, we considered 5 pairs of models (BERT_base vs. MobileBERT, MobileBERT vs. DistilBERT, DistilBERT vs. RoBERTa, RoBERTa vs. XLNet, and XLNet vs. GPT-2) and provided humans with outputs for each pair of models, hiding the model's name. We took 10 conflicting examples between *Sem-nCG* and *ROUGE-L* for each pair of models. This means that humans evaluated $10 \times 2 = 20$ summaries, each 5 sentences long, for each model pair. In total, annotators labeled $5 \times 20 \times 5 = 500$ sentences for model output, after reading around $5 \times 10 \times 50 = 2500$ sentences for articles and $5 \times 10 \times 3 = 150$ sentences for highlights. We asked the annotators to say which *extractive* summary is better and matched their decision against both *ROUGE* and *Sem-nCG@k*'s conclusions. Our annotators were three doctoral students all working in NLP. We took the majority voting judgement from annotators and the results are reported in Table 6. As summarized in Table 6, blind evaluation by humans indicated *Sem-nCG@k* was more accurate than *ROUGE* in the case of disagreements between the two, thus confirming that *Sem-nCG@k* captures semantics better than *ROUGE*.

7.2 Meta-Evaluation of *Sem-nCG*

We further performed meta-evaluation of the *Sem-nCG* metric using data provided by (Fabbri et al., 2021)[6]. The dataset includes summaries generated by 16 models (both extractive and abstractive) from 100 source news articles (1600 summaries in total). For our experiments, we only considered the extractive summaries and omitted samples containing less than 3 sentences (as we report *Sem-nCG@3*), and that resulted in 252 samples. Each of these summaries was annotated by 5 indepedent crowdsource workers and 3 independent experts (8 annotations in total). Summaries were evaluated across 4 dimensions: consistency, fluency, coherence, relevance after looking into the CNN/DailyMail reference and 10 additional crowd-sourced reference summaries. As mentioned in (Gillick and Liu, 2010), non-expert annotation can be risky, so we only considered expert annotations as followed by (Fabbri et al., 2021) as well. Next, we computed kendall's tau correlation between the *Sem-nCG* score and each of consistency, fluency, coherence, relevance scores rated by humans in the case of single reference setting for the following 3 different scenarios (example in Table 8):

- *Less Overlapping Reference (LOR)*: Highly abstractive references with fewer lexical overlap with the original document.
- *Medium Overlapping Reference (MOR)*: Somewhat extractive references (CNN/DailyMail) with moderate lexical overlap.
- *Highly Overlapping Reference (HOR)*: Highly extractive references with high lexical overlap.

Table 7 shows that our proposed metric outperforms ROUGE in terms of consistency (the most crucial dimension perhaps) for all 3 types of references (even for HOR) with a considerable margin. Interestingly, we found that there is not a clear winner among the embedding choices. However, the STSb-distilbert embedding shows good performance in the consistency dimension both for less overlapping and high overlapping references. Note that STSb-distilbert also takes less computation time (Table 4) and can be a better choice for low-resource evaluation scenarios.

Along the fluency dimension, our proposed *sem-nCG@k* correlates better with humans for all types of references (except for less overlapping references with a comparable performance). Of particular interest from Table 7 are the more abstractive (LOR) references with little overlaps, where *sem-nCG@k* correlation is higher than *ROUGE* for 3 dimensions, including consistency, coherence, and relevance. For medium and highly overlapping references, *ROUGE* correlation along the coherence and relevance dimension was higher, which is somewhat expected, since *ROUGE* mainly computes lexical overlaps. *These results suggest that,*

[6] https://github.com/Yale-LILY/SummEval

Embedding	Infersent-v1	Infersent-v2	STSb-bert	STSb-roberta	STSb-distilbert	Elmo	USE-enc2	USE-enc3
Time (Second)	0.36	0.41	0.33	0.34	**0.13**	79.1	20.1	27.5

Table 4: The average computational time (in CPU) required to run the evaluation of a single test instance for different pre-trained embeddings. Apparently, STSb-distilbert is very fast when compared to the other embeddings.

Model	Paired with	R1	R2	R3	RL
BERTbase	MobileBERT	6478	6258	6461	6397
	DistilBERT	6443	6326	6486	6336
	RoBERTa	4408	3994	4345	4511
	XLNet	3853	4121	4447	4643
	GPT-2	4380	3989	4376	4478
MobileBERT	DistilBERT	6152	5699	6040	5850
	RoBERTa	5397	5027	5261	5269
	XLNet	5533	5024	5222	5287
	GPT-2	5488	5050	5250	5286
DistilBERT	RoBERTa	7786	3800	4173	4285
	XLNet	4296	3917	4251	4458
	GPT-2	4040	3759	4147	4282
RoBERTa	XLNet	5772	4489	4923	4725
	GPT-2	4911	4583	5008	4787
XLNet	GPT-2	4820	4471	4850	4693

Table 5: Disagreement between *Sem-nCG@5* (with $\text{Ensemble}_{\text{rel}}$) and *ROUGE* (F1) out of 11426 samples for different extractive summarization model pairs.

	Win	Lose	Tie
Sem-nCG@5 vs. *ROUGE*	**62%**	36%	2%

Table 6: Statistics for *Sem-nCG@5* wins, loses, and ties against *ROUGE-L (F1)*. Results report average of 5 pairs ($\text{BERT}_{\text{base}}$ vs. MobileBERT, MobileBERT vs. DistilBERT, DistilBERT vs. RoBERTa, RoBERTa vs. XLNet and XLNet vs. GPT-2) evaluated by humans.

while there may not be a clear winner between sem-nCG and ROUGE when the testing corpus mostly contains medium and highly overlapping references, however, ***sem-nCG@k is clearly a superior metric when evaluating summaries against a more abstract (low overlap) reference****.*

8 Discussions and Conclusion

In this paper, we revisited the problem of automatic evaluation for the *extractive* summarization task, exclusively focusing on the popular $ROUGE$ metric. We first argued that any summary evaluation should be more semantic-aware and demonstrated that $ROUGE$ fails to capture semantics through comprehensive experiments. Indeed, $ROUGE$ score drops (5-7%) even only for small percentages (20%) of synonym perturbation, and thus is not optimal for evaluating any summarization task.

Next, we argued that a "good" metric for evaluating extractive summarization task should assess its core ranking quality, which $ROUGE$ does not. To address this issue, we proposed a new metric called *Sem-nCG* which is both semantic-aware and considers ranking quality. More importantly, *Sem-nCG* provides an automated way to compare a set of top-ranked model-extracted sentences (the system-extracted summary) against an ideal ranking of sentences, where the ideal ranking is automatically inferred by computing gains based on some human-written summary. This saves us from tedious process of manual annotation of each sentence within the original document, thus making it practically suitable for large scale automated evaluation.

The correctness of the *Sem-nCG* metric depends largely on the reliability of the groundtruth gains computed by algorithm 1. Therefore, to verify the quality of the groundtruth gains, we conducted extensive quantitative evaluations which confirm that the *Sem-nCG* metric is stable across multiple sentence embedding techniques (very robust) [section 6.2]. Through additional experiments, we have demonstrated the following as well: 1) *Sem-nCG* correlates better with humans [section 7.2]; 2) *Sem-nCG* often disagrees with *ROUGE* for pairwise comparison of summarization methods [section 6.4]; 3) In the cases of such disagreements, further verification from human judges confirmed that *Sem-nCG* is more reliable than the *ROUGE* metric [section 7.1]; and 4) *Sem-nCG* is a superior metric when evaluating summaries against a more abstract (low overlap) reference [section 7.2]. To conclude, we recommend the following practice:

- For extractive summarization evaluation, please refrain from overemphasizing a substantial improvement over $ROUGE$ solely.
- While evaluating extractive summaries, mitigate the limitations of the $ROUGE$ metric by reporting additional metrics which are semantic-aware and can generate reliable gains from human references (e.g., *Sem-nCG*), especially when the human-references are more abstractive in nature.
- Human judgment must still be the gold standard, and while making a conclusion of making substantial improvement over previous work, make sure it is backed by human evaluation.

We recognize that our proposed *Sem-nCG* metric overlooks redundancy when computing groundtruth gains; thus, our immediate future goal is to design a redundancy-aware *Sem-nCG*, as well as expand *Sem-nCG* for multi-references and multi-document summarization settings.

Sem-nCG@3												
	Consistency			**Fluency**			**Coherence**			**Relevance**		
Embedding	LOR	MOR	HOR	LOR	MOR	HOR	LOR	MOR	HOR	LOR	MOR	HOR
Infersent-v1	0.06	0.07	0.07	0.03	0.02	0.10	0.04	0.01	-0.01	0.07	0.08	0.05
Infersent-v2	0.08	0.05	0.08	0.05	0.02	**0.12**	0.06	0.07	0.04	0.07	0.13	0.09
STSb-bert	0.11	0.07	0.09	0.01	0.03	0.10	-0.01	0.07	0.01	0.03	0.14	0.12
STSb-distilbert	**0.17**	0.09	**0.12**	0.00	0.02	0.04	0.06	0.04	-0.01	0.06	0.11	0.07
STSb-roberta	0.12	**0.13**	0.05	-0.01	0.01	0.04	0.02	0.01	0.00	0.07	0.08	0.09
Elmo	0.06	0.08	0.09	0.00	0.00	0.06	0.02	0.02	0.01	0.02	0.08	0.06
USE-enc2	0.05	0.03	0.04	0.03	**0.06**	0.08	**0.07**	0.09	0.02	0.11	0.13	0.08
USE-enc3	0.01	0.01	0.09	-0.08	0.00	0.04	0.02	0.03	0.04	0.01	0.12	0.05
$\text{Ensemble}_{\text{sim}}$	0.08	0.07	0.08	0.00	0.03	0.07	0.05	0.05	-0.03	0.08	0.13	0.07
$\text{Ensemble}_{\text{rank}}$	0.10	0.09	0.09	-0.02	0.01	0.08	0.04	0.04	-0.01	0.08	0.12	0.07
$\text{Ensemble}_{\text{rel}}$	0.09	0.08	0.09	0.01	0.03	0.08	**0.07**	0.06	0.00	**0.11**	0.14	0.07
ROUGE												
ROUGE-1	0.08	0.04	-0.01	**0.06**	0.05	0.07	0.02	0.13	**0.13**	0.07	0.21	**0.22**
ROUGE-2	0.05	0.02	-0.05	0.04	0.04	0.01	-0.05	0.14	0.13	0.01	0.23	0.21
ROUGE-3	0.08	0.03	-0.05	0.06	0.05	0.00	-0.08	**0.15**	0.12	0.02	**0.24**	0.19
ROUGE-L	0.02	0.06	-0.02	0.02	0.04	-0.04	-0.01	0.13	0.07	0.04	0.18	0.14

Table 7: Kendall's tau correlation coefficients of expert annotations computed at single reference setting for ROUGE and *Sem-nCG* along four quality dimensions (for top-3 sentences). The correlation was demonstrated for low overlapping references (LOR), Medium Overlapping CNN/DailyMail Reference (MOR), and high overlapping references (HOR) chosen from 11 reference summaries per example. The outperformed correlated values in each column have been bolded both for *Sem-nCG* and *ROUGE*.

Article		
Paul Merson has restarted his row with Andros Townsend after the Tottenham midfielder was brought on with only seven minutes remaining in his team's 0-0 draw with Burnley on Sunday. 'Just been watching the game, did you miss the coach? #RubberDub #7minutes,' Merson put on Twitter. Merson initially angered Townsend for writing in his Sky Sports column that 'if Andros Townsend can get in (the England team) then it opens it up to anybody.' Paul Merson had another dig at Andros Townsend after his appearance for Tottenham against Burnley Townsend was brought on in the 83rd minute for Tottenham as they drew 0-0 against Burnley Andros Townsend scores England's equaliser in their 1-1 friendly draw with Italy in Turin on Tuesday night The former Arsenal man was proven wrong when Townsend hit a stunning equaliser for England against Italy and he duly admitted his mistake. 'It's not as though I was watching hoping he wouldn't score for England, I'm genuinely pleased for him and fair play to him – it was a great goal,' Merson said. 'It's just a matter of opinion, and my opinion was that he got pulled off after half an hour at Manchester United in front of Roy Hodgson, so he shouldn't have been in the squad. 'When I'm wrong, I hold my hands up. I don't have a problem with doing that - I'll always be the first to admit when I'm wrong.' Townsend hit back at Merson on Twitter after scoring for England against Italy Sky Sports pundit Merson (centre) criticised Townsend's call-up to the England squad last week Townsend hit back at Merson after netting for England in Turin on Wednesday, saying 'Not bad for a player that should be 'nowhere near the squad' ay @PaulMerse?' Any bad feeling between the pair seemed to have passed but Merson was unable to resist having another dig at Townsend after Tottenham drew at Turf Moor.		
LOR	**MOR**	**HOR**
An athlete was brought in to save the game during an event against a rival team. Although many disagreed with this decision as players have been known to get in trouble from time to time.	Andros Townsend an 83rd minute sub in Tottenham's draw with Burnley. He was unable to find a winner as the game ended without a goal. Townsend had clashed with Paul Merson last week over England call-up.	Paul Merson and Andros Townsend have been in about for a while now, Merson felt that Townsend did not deserve a place in the English team. Townsend scored for England with a crucial goal to which Merson apologized and acknowledge the performance of Townsend in that game and wished him well on his performance. The back and forth between the two men has continued regardless but it appears that now their bad feelings have subsided despite some light jest between the two.

Table 8: An example of the three scenarios highlighted in the human evaluation

Acknowledgments

This research has been supported by the Department of Computer Science and Software Engineering, Auburn University. We thank the members of BDI lab for helping with human evaluation and the anonymous reviewers for their valuable feedback.

References

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant,

Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

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).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 670–680. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

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

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

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.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1693–1701.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Shubhra (Santu) K Karmaker, Parikshit Sondhi, and ChengXiang Zhai. 2020. Empirical analysis of impact of query-specific customization of ndcg: a case-study with learning-to-rank methods. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3281–3284.

Shubhra Kanti Karmaker Santu, Parikshit Sondhi, and ChengXiang Zhai. 2017. On application of learning to rank for e-commerce search. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 475–484.

Saar Kuzi, Sahiti Labhishetty, Shubhra Kanti Karmaker Santu, Prasad Pradip Joshi, and ChengXiang Zhai. 2019. Analysis of adaptive training for learning to rank in information retrieval. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2325–2328.

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP-IJCNLP*, pages 3728–3738.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Comput. Linguistics*, 39(2):267–300.

Yuning Mao, Liyuan Liu, Qi Zhu, Xiang Ren, and Jiawei Han. 2020. Facet-aware evaluation for extractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 4941–4957. Association for Computational Linguistics.

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

Ani Nenkova and Rebecca J. Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL*, pages 145–152. The Association for Computational Linguistics.

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.

Rebecca J. Passonneau, Emily Chen, Weiwei Guo, and Dolores Perin. 2013. Automated pyramid scoring of summaries using distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 2: Short Papers*, pages 143–147. The Association for Computer Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the ACL 2019, Volume 1: Long Papers*, pages 5093–5100. Association for Computational Linguistics.

Dragomir R. Radev and Daniel Tam. 2003. Summarization evaluation using relative utility. In *CIKM International Conference on Information and Knowledge Management*, pages 508–511. ACM.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

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.

Ori Shapira, David Gabay, Yang Gao, Hadar Ronen, Ramakanth Pasunuru, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2019. Crowdsourcing lightweight pyramids for manual summary evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 682–687. Association for Computational Linguistics.

Simeng Sun and Ani Nenkova. 2019. The feasibility of embedding based automatic evaluation for single document summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, pages 1216–1221. Association for Computational Linguistics.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic BERT for resource-limited devices. In *ACL*, pages 2158–2170.

Simone Teufel and Hans van Halteren. 2004. Evaluating information content by factoid analysis: Human annotation and stability. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 419–426. ACL.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural*

Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Stratos Xenouleas, Prodromos Malakasiotis, Marianna Apidianaki, and Ion Androutsopoulos. 2019. SUM-QE: a bert-based summary quality estimation model. In *Proceedings of Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, pages 6004–6010. Association for Computational Linguistics.

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.

Qian Yang, Rebecca J. Passonneau, and Gerard de Melo. 2016. PEAK: pyramid evaluation via automated knowledge extraction. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2673–2680. AAAI Press.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5754–5764.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *CoRR*, abs/2106.11520.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR*. OpenReview.net.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 563–578. Association for Computational Linguistics.

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.

A Appendix

A.1 Extractive Summarization Models

$\text{BERT}_{\text{base}}$: Transformer models achieve state of the art performance on different NLP tasks. A simple variant of BERT for extractive summarization has been shown in paper (Liu and Lapata, 2019) which consists of 2 parts: a BERT encoder and a summarization classifier. The BERT model here consists of the pretrained $\text{BERT}_{\text{base}}$ encoder from masked language model by (Devlin et al., 2019).

MobileBERT (Sun et al., 2020): In an effort to make BERT available for low resource devices, MobileBERT has been proposed which is a thin version of $\text{BERT}_{\text{large}}$, with carefully designed balance between self-attentions and feed-forward.

DistilBERT (Sanh et al., 2019): Model size reduction has been studied extensively in the literature due to huge computational expenses of large models. DistilBERT uses knowledge distillation during pre-training to reduce the size of BERT model. It has 40% less parameter than $\text{BERT}_{\text{base}}$ and runs 60% faster while achieving 97% of BERT's performance.

RoBERTa (Liu et al., 2019): RoBERTa is another variant of BERT that modified key hyperparameters and removed the next sentence prediction objective while training with larger mini batches and learning rates. The authors have shown the importance of design choices in BERT architecture while improving the performance.

XLNet (Yang et al., 2019) : While BERT has been pre-trained on mask language model, XLNet proposes a generalized autoregressive method for pre-training and an extension of the Transformer-XL that outperforms BERT on 20 NLP tasks.

GPT-2 (Radford et al., 2019): GPT-2 is similar to decoder only transformer but trained on a very large dataset which outperforms BERT on NLP tasks like question answering, reading comprehension, summarization.

A.2 Various Sentence Embeddings used for *Sem-nCG*

Infersent (Conneau et al., 2017): BiLSTM network with max-pooling generates 4096-

Model	Layers	Hidden Units	Parameters	Size
$BERT_{base}$	12	768	108M	100%
MobileBERT	24	128	25.3M	25%
DistilBERT	12	768	66M	60%
RoBERTa	12	768	125M	113%
XLNet	12	768	110M	100%
GPT-2	24	1024	355M	328%

Table 9: Summary of the Model Architecture

dimensional sentence embedding. Infersent-v1 (trained with GloVe) and Infersent-v2 (trained with fastText) are the two versions of Infersent sentence embedding.

Semantic Textual Similarity benchmark (STSb) (Reimers and Gurevych, 2019): Sentence Transformer allows to generate dense vector representations of sentences. We considered three of the best available models that were optimized for semantic textual similarity (STSb-bert, STSb-roberta and STSb-distilbert).

Elmo (Peters et al., 2018): A fixed mean-pooling of all contextualized word representations with shape 1024 has been considered, effectively transforming the contextualized word-embedding into a sentence embedding.

Google Universal Sentence Encoder (USE) (Cer et al., 2018): USE converts the input text to a 512-dimensional vector. There are two kinds available, i) enc-2 (Iyyer et al., 2015) based on the deep average network ii) enc-3 (Vaswani et al., 2017) based on transformer.

A.3 *Sem-nCG@k* and ROUGE Scores for Top-3 Sentences

To generalize our remarks, we repeated the same experiments (mentioned in Section 5) for ROUGE and *Sem-nCG@k* for the top-3 sentences. Table 10 demonstrates that ROUGE is sensitive to synonym perturbation for the top-3 sentences of extractive models. Table 11, on the other hand, confirms that *Sem-nCG@k* is merely sensitive to sentence perturbation (especially $Ensemble_{rel}$) and also robust across various sentence embedding variations (confirms from lower standard deviation).

A.4 Dimensions of Human Evaluation

We have considered four quality dimensions following (Fabbri et al., 2021) to measure the Kendall's tau rank correlation between *Sem-nCG@3* and ROUGE.

Consistency: facts between the summary and its source are consistent. Factually consistent summaries contain just assertions from the summarized source, and do not include any trippy facts.

Fluency: sentence structure and quality. Referring to the DUC quality criteria (Dang, 2005), summary sentences "should have no formatting problems, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read."

Coherence: the overall quality of summary sentences while retaining a coherent body of information about a topic rather than just a jumble of related information (Dang, 2005).

Relevance: extracting the most significant information from the source. Summaries with redundancy and extra information were to be penalized by the annotators. Only relevant information from the original should be provided in the summary.

		BERT-base		**MobileBERT**		**DistilBERT**		**RoBERTa**		**XLNet**		**GPT-2**	
		Actual	Perturbed	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed
ROUGE-1	Precision	36.64	30.24	35.2	29.15	37.77	30.97	32.99	27.78	33.31	27.83	33.04	27.76
	Recall	52.8	46.46	50.66	44.6	53.19	46.75	47.42	42.42	47.01	41.99	47.18	41.82
	F1	41.72	35.3	39.93	33.9	42.67	35.93	37.55	32.36	37.6	32.24	37.48	32.12
ROUGE-2	Precision	16.67	10.68	15.47	9.95	17.72	11.28	13.29	8.73	13.44	8.71	13.35	8.62
	Recall	24.01	16.48	22.34	15.28	24.89	17.02	19.36	13.55	19.19	13.38	19.28	13.24
	F1	18.95	12.48	17.55	11.58	19.97	13.06	15.19	10.22	15.23	10.15	15.19	10.04
ROUGE-3	Precision	9.77	4.84	8.96	4.5	10.52	5.19	7.38	3.78	7.46	3.77	7.41	3.72
	Recall	13.86	7.38	12.71	6.79	14.64	7.74	10.64	5.82	10.54	5.76	10.6	5.7
	F1	11.01	5.62	10.06	5.19	11.79	5.97	8.38	4.4	8.39	4.38	8.38	4.32
ROUGE-L	Precision	22.8	18.73	21.87	18.01	24.19	19.61	20.64	17.22	20.86	17.18	20.7	17.23
	Recall	33.14	29.05	31.67	27.7	34.33	29.83	29.98	26.61	29.73	26.26	29.85	26.28
	F1	26.03	21.93	24.85	20.96	27.41	22.8	23.58	20.14	23.62	19.99	23.56	20.01

Table 10: *ROUGE* scores for the extractive summarization models ($\text{BERT}_{\text{base}}$, MobileBERT, DistilBERT, RoBERTa, XLNet, GPT-2) on CNN/DailyMail test dataset. The results are for top-3 extracted sentences when the outputs are in actual and perturbed.

	BERT-base		MobileBERT		DistilBERT		RoBERTa		XLNet		GPT-2	
	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed	Actual	Perturbed
Infersent-v1	78.03	75.04	73.33	71.14	79.25	75.17	72.72	69.47	72.72	69.25	72.48	69.11
Infersent-v2	77.73	75.08	72.31	70.59	79.64	75.99	72.02	69.45	72.18	69.36	71.86	69.12
STSb-bert	78.08	77	72.93	71.35	79.46	78.91	72.93	70.67	73.43	71.01	73.08	70.64
STSb-roberta	77.66	76.84	72.88	71.44	79.06	78.42	72.79	70.79	73.27	71.07	73.02	70.6
STSb-distilbert	76.95	75.96	71.98	70.42	78.38	77.78	72.02	69.82	72.48	70.11	72.08	69.63
Elmo	77.28	71.08	72.33	68.79	78.34	70.52	71.7	67.57	71.98	67.06	71.68	67.3
USE-enc2	79.43	78.58	73.11	71.27	81.52	81.29	73.93	71.44	74.5	72	74.03	71.28
USE-enc3	78.37	76.98	72.37	70.14	80.28	79.53	71.97	69.36	72.36	69.8	71.94	69.16
$\text{Ensemble}_{\text{sim}}$	80.17	79.37	74.37	73.15	81.91	81.19	74.26	72.2	74.69	72.46	74.29	71.97
$\text{Ensemble}_{\text{rank}}$	80.17	79.15	74.5	73.37	81.8	80.66	74.26	72.19	74.62	72.3	74.26	71.92
$\text{Ensemble}_{\text{rel}}$	**81.2**	**80.98**	**75.7**	**75.48**	**82.81**	**82.46**	**75.56**	**74.6**	**75.93**	**74.76**	**75.57**	**74.38**
std	1.38	2.69	1.15	1.83	1.55	3.43	1.24	1.89	1.29	2.05	1.27	1.91

Table 11: *Sem-nCG@3* scores for the top-3 sentences of the extractive summarization models ($\text{BERT}_{\text{base}}$, MobileBERT, DistilBERT, RoBERTa, XLNet, GPT-2) on CNN/DailyMail test dataset for different embedding variations.