Redundancy Aware Multiple Reference Based Gainwise Evaluation of Extractive Summarization

Mousumi Akter
Research Center Trustworthy
Data Science and Security
Technical University Dortmund, Germany
mousumi.akter@tu-dortmund.de

Santu Karmaker
Big Data Intelligence (BDI) Lab
Auburn University
Alabama, USA
sks0086@auburn.edu

Abstract

The ROUGE metric is commonly used to evaluate extractive summarization task, but it has been criticized for its lack of semantic awareness and its ignorance about the ranking quality of the extractive summarizer. Previous research has introduced a gain-based automated metric called *Sem-nCG* that addresses these issues, as it is both rank and semantic aware. However, it does not consider the amount of redundancy present in a model summary and currently does not support evaluation with multiple reference summaries. It is essential to have a model summary that balances importance and diversity, but finding a metric that captures both of these aspects is challenging. In this paper, we propose a redundancy-aware *Sem-nCG* metric and demonstrate how the revised *Sem-nCG* metric can be used to evaluate model summaries against multiple references as well which was missing in previous research. Experimental results demonstrate that the revised *Sem-nCG* metric has a stronger correlation with human judgments compared to the previous *Sem-nCG* metric and traditional ROUGE and BERTScore metric for both single and multiple reference scenarios.

1 Introduction

For the past two decades, ROUGE (Lin, 2004b) has been the most used metric for evaluating extractive summarization tasks. Nonetheless, ROUGE has long been criticized for its lack of semantic awareness (Graham, 2015; Ng and Abrecht, 2015; Ganesan, 2018; Yang et al., 2018) and its ignorance about the ranking quality of the extractive summarizer (Akter et al., 2022).

To address these issues, previous work has proposed a gain-based metric called *Sem-nCG* (Akter et al., 2022) to evaluate extractive summaries by incorporating rank and semantic awareness. Redundancy, a crucial factor in evaluating extractive summaries, was not, however, included in the *Sem-nCG* metric. Additionally, their proposed *Sem-nCG* metric does not support the evaluation of model summaries against multiple references. However, it is well recognized that a set of documents can have multiple, very different, and equally valid summaries; as such, obtaining multiple reference summaries can improve the stability of the evaluation (Nenkova, 2005; Lin, 2004a). It's quite challenging to come up with a metric that takes into account the balance between importance and diversity in model summary. Therefore, it's necessary to carry out a systematic study on how to integrate redundancy and multiple references to the existing *Sem-nCG* metric.

In this paper, we first incorporate redundancy into the previously proposed *Sem-nCG* metric. In other words, we propose a redundancy-aware *Sem-nCG* metric by exploring different ways of incorporating redundancy into the original metric. Through extensive experiments, we demonstrate that the redundancy-aware *Sem-nCG* exhibits a notably stronger correlation with humans than the original *Sem-nCG* metric.

Next, we demonstrate how this redundancy-aware metric could be applied to evaluate model summaries against multiple references. This is a non-trivial task because *Sem-nCG* evaluates a model-generated summary by considering it as a ranked list of sentences and then comparing it against an automatically inferred *ground-truth* ranked list of sentences within a source document based on a single human written summary (Akter et al., 2022). However, in the case of multiple references, the *ground-truth* ranked list of source sentences must be inferred based on all available human-written reference summaries, not just one.

When there are multiple reference summaries available, incorporating them into evaluation poses significant challenge. This is because the quality of human-written summaries differs not only in writing style but also in focus. Moreover, in-

cluding multiple reference summaries with a lot of terminology variations and paraphrasing makes the automated evaluation metric less stable (Cohan and Goharian, 2016). In this work, we have also shown how to infer a single/unique ground-truth ranking based on multiple reference summaries with the proposed redundancy-aware *Sem-nCG* metric. Our findings suggest that, compared to the conventional ROUGE and BERTScore metric, the redundancy-aware *Sem-nCG* exhibits a stronger correlation with human judgments for evaluating model summaries when both single and multiple references are available. Therefore, we encourage the community to use redundancy-aware *Sem-nCG* to evaluate extractive summarization tasks. Our contributions are:

- Redundancy of extracted sentences is a common problem in extractive summarization systems. We have demonstrated how to consider redundancy awareness in the already-designed *Sem-nCG* metric.
- We present how to use the redundancy-aware *Sem-nCG* metric for summary evaluation with multiple references which poses unique challenges of variability.
- The revised *Sem-nCG* metric exhibits a stronger correlation with human judgments for evaluating model summaries when both single and multiple references are available, not only with the previous *Sem-nCG* metric but also with conventional ROUGE and BERTScore metric.

2 Redundancy-aware *Sem-nCG* Metric

***Sem-nCG* Score:** Normalized Cumulative Gain (*nCG*) is a popular evaluation metric in information retrieval to evaluate the quality of a ranker. nCG compares the model ranking with an *ideal* ranking and assigns a certain score to the model based on some pre-defined gain. (Akter et al., 2022) has utilized the idea of *nCG* in the evaluation of extractive summarization. The basic concept of Sem-nCG is to compute the gain (*CG@k*) obtained by a top k extracted sentences and divide that by the maximum/ideal possible gain (*ICG@k*), where the gains are inferred by comparing the input document against a human written summary. Mathematically:

$$\textit{Sem-nCG@k} = \frac{\textit{CG@k}}{\textit{ICG@k}} \tag{1}$$

Redundancy Score: We followed (Chen et al., 2021) to compute self-referenced redundancy score which is computationally efficient and less ambiguous than classical approaches. The summary, X, itself is used as the reference to determine the degree of semantic similarity between each summary token/sentence and the other tokens/sentences. The average of maximum semantic similarity is used to determine the redundancy score. For a given summary, $X = \{x_1, x_2, ..., x_n\}$, the calculation is as follows:

$$Score_{\text{red}} = \frac{\sum_i max_{j:i \neq j} Sim(x_j, x_i)}{|\mathbf{X}|} \tag{2}$$

where, $j : i \neq j$ denotes that the similarity between x_i and itself has not been considered. Note that $Score_{red} \in [0, 1]$ in our case and lower is better.

Final Score: We used the following formula to calculate the final score after obtaining the scores of *Sem-nCG* and $Score_{red}$:

$$Score = \lambda * \textit{Sem-nCG} + (1 - \lambda) * (1 - Score_{red}) \tag{3}$$

Here, $\lambda \in [0, 1]$ is a hyper-parameter to scale the weight between $Score_{red}$ and *Sem-nCG*. $Score \in [0, 1]$ where higher score means better summary.

3 Experimental Setup

Dataset: Human correlation is an essential attribute to consider while assessing the quality of a metric. To compute the human correlation of the revised redundancy-aware *Sem-nCG* metric, we utilized SummEval dataset from (Fabbri et al., 2021)[1]. The annotations include summaries generated by 16 models (abstractive and extractive) from 100 news articles (1600 examples in total) on the CNN/DailyMail Dataset. Each source news article includes the original CNN/DailyMail reference summary as well as 10 additional crowd-sourced reference summaries. Each summary was annotated by 5 independent crowd-sourced workers and 3 independent experts (8 annotations in total) along the four dimensions: *Consistency*, *Relevance*, *Coherence* and *Fluency* (Fabbri et al., 2021)[2]. As this work focuses on the evaluation of extractive summarization, we considered the output generated by extractive models and filtered out samples comprising less than 3 sentences (as we report *Sem-nCG@3*). Additionally, we considered the expert

[1]We used the dataset by (Fabbri et al., 2021), the only available benchmark "meta-evaluation dataset" for **extractive summarization**, to the best of our knowledge. *Sem-nCG*'s authors have demonstrated its correlation with human judgment on this dataset. To ensure a fair comparison, we maintained the same settings as the original *Sem-nCG* when assessing the redundancy-aware *Sem-nCG*.

[2]See Appendix A.2 for details

annotations for the meta-evaluation, as non-expert annotations can be risky (Gillick and Liu, 2010).

As was done in (Akter et al., 2022), for each sample, from the 11 available reference summaries, we considered 3 settings: Less Overlapping Reference/LOR (highly abstractive references with fewer lexical overlap with the original document), Medium Overlapping Reference/MOR (medium lexical overlap with the original document) and Highly Overlapping Reference/HOR (highly extractive references with high lexical overlap with the original document).

Embedding for Groundtruth Ranking: The core of the *Sem-nCG* metric is to automatically create the groundtruth/ideal ranking against which the model ranking is compared. To create the groundtruth ranking, (Akter et al., 2022) used various sentence embeddings. Similarly, we utilized various sentence embeddings as well since our goal is to compare the new redundancy-aware *Sem-nCG* metric to the original *Sem-nCG* metric. Specifically, we considered Infersent (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) with enc-2 (Iyyer et al., 2015) based on the deep average network, to infer the groundtruth/ideal ranking of the sentences within the input document with guidance from the human written summaries.

***Score$_{red}$* Computation:** To compute the self-referenced redundancy score, we used the top-3 sentences from the model generated summary (as we report *Sem-nCG@3*). We calculated each sentence's maximum similarity to other sentences and then averaged it to get the desired *Score$_{red}$*. We experimented with four distinct variations to compare the sentences: cosine similarity (by converting sentences to STSb-distilbert (Reimers and Gurevych, 2019) embeddings), ROUGE (Lin, 2004b), MoverScore (Zhao et al., 2019) and BERTScore (Zhang et al., 2020).

4 Results

4.1 Redundancy-aware *Sem-nCG*

We first considered how redundancy-aware SemnCG performs in extractive summarization with single reference. As shown in Table 1, we computed Kendall's tau (τ) correlation between the expert given score for model summary and the SemnCG score with/without redundancy along the four meta-evaluation criteria: *Consistency*, *Relevance*, *Coherence*, and *Fluency*, for different embedding variations (to create the groudtruth ranking) and different approaches to compute *Score$_{red}$*. We utilized Equation 3 to compute the redundancy-aware *SemnCG* score, where lambda (λ) is a hyper-parameter choice and is set to $\lambda = 0.5$ empirically. In Table 1 w/o redundancy refers to Equation 1.

Table 1 shows that the redundancy-aware *SemnCG* metric outperforms the original *Sem-nCG* metric in terms of *Consistency*, *Relevance*, and *Coherence*; with a 5% improvement in *Relevance* and a 14% improvement in *Coherence* for less overlapping references (LOR). We also observe improvements in the *Relevance* (9%) and *Coherence* (20%) dimensions for medium overlapping references (MOR). For High Overlapping References (HOR), the improvement is 8% and 22% for *Relevance* and *Coherence*, respectively.

We also observe that STSb-distilbert embedding is a better choice in the *Consistency* dimension, whereas USE with enc-2 is a better choice in the *Relevance* and *Coherence* dimensions to construct the groundtruth ranking. Therefore, we recommend STSb-distilbert to create groundtruth ranking if *Consistency* is a top priority, otherwise, we recommend using USE with enc-2. A groundtruth ranking was also created by combining STSb-distilbert and USE into an ensemble, which showed balanced performance across all four dimensions. It also appears that ROUGE and BERTScore provide comparable performances while computing *Score$_{red}$*. However, using ROUGE score as self-referenced redundancy will be a better choice as evident from Section 4.3.

In Table 2 Kendall's tau correlation of ROUGE and BERTScore has been demonstrated to get an idea of the advantage of redundancy-aware *SemnCG* and it is clearly evident that redundancy-aware Sem-nCG also exhibits stronger correlation than these metrics.

4.2 Hyperparameter Choice

In figure 1, we have varied $\lambda \in [0, 1]$ for the 3 scenarios (LOR, MOR and HOR) and computed human correlation along four dimensions (*Consistency*, *Relevance*, *Coherence* and *Fluency*) when different embeddings are used to create the groundtruth ranking and ROUGE score is used to compute *Score$_{red}$*. Human correlations with BERTScore-based redundancy are presented in Appendix. For both redundancy penalties, it shows

Embedding	Type	Consistency			Relevance			Coherence			Fluency		
		LOR	MOR	HOR	LOR	MOR	HOR	LOR	MOR	HOR	LOR	MOR	HOR
Inferesent	w/o redundancy	0.08	0.06	0.08	0.07	0.12	0.09	0.06	0.06	0.04	**0.05**	0.03	**0.12**
+ Redundancy penalty	Cosine Similarity	0.04	0.02	0.06	0.08	0.15	0.13	0.14	0.19	0.18	0.02	-0.02	0.08
	ROUGE-1	0.07	0.05	0.11	0.11	0.18	0.17	0.18	0.25	**0.26**	-0.01	-0.04	0.05
	MoverScore	0.05	0.06	0.11	0.09	0.15	0.12	0.11	0.13	0.11	0.03	0.01	0.11
	BERTScore	0.05	0.02	0.08	0.13	0.19	0.18	0.18	0.22	0.24	-0.01	-0.04	0.04
Elmo	w/o redundancy	0.06	0.07	0.09	0.02	0.08	0.06	0.02	0.02	0.01	0.00	0.01	0.06
+ Redundancy penalty	Cosine Similarity	0.03	0.03	0.05	0.04	0.13	0.10	0.12	0.14	0.14	-0.06	-0.05	0.02
	ROUGE-1	0.08	0.05	0.08	0.07	0.15	0.14	0.17	0.20	0.20	-0.06	-0.06	0.01
	MoverScore	0.08	0.07	0.10	0.04	0.10	0.09	0.07	0.06	0.06	-0.02	-0.01	0.05
	BERTScore	0.06	0.03	0.05	0.09	0.17	0.16	0.17	0.19	0.18	-0.06	-0.07	0.00
STSb-bert	w/o redundancy	0.11	0.08	0.09	0.03	0.13	0.12	-0.01	0.06	0.01	0.03	**0.10**	0.03
+ Redundancy penalty	Cosine Similarity	0.08	0.01	0.06	0.05	0.17	0.13	0.10	0.18	0.16	-0.05	0.02	0.05
	ROUGE-1	0.12	0.05	0.09	0.08	**0.22**	0.18	0.14	0.25	0.22	-0.04	-0.04	0.01
	MoverScore	0.12	0.06	0.10	0.05	0.16	0.15	0.04	0.11	0.09	-0.01	0.02	0.08
	BERTScore	0.10	0.01	0.06	0.11	**0.22**	**0.20**	0.14	0.24	0.20	-0.06	-0.04	0.01
STSb-roberta	w/o redundancy	0.12	**0.14**	0.07	0.07	0.07	0.05	0.04	0.00	-0.02	-0.01	0.01	0.06
+ Redundancy penalty	Cosine Similarity	0.09	0.07	0.05	0.08	0.11	0.06	0.13	0.13	0.10	-0.06	-0.05	-0.01
	ROUGE-1	0.12	0.11	0.09	0.11	0.16	0.10	0.18	0.20	0.17	-0.07	-0.07	-0.04
	MoverScore	0.13	0.13	0.10	0.09	0.10	0.07	0.08	0.07	0.04	-0.03	0.00	0.04
	BERTScore	0.10	0.08	0.05	0.13	0.18	0.12	0.17	0.18	0.15	-0.08	-0.06	-0.04
USE	w/o redundancy	0.05	0.04	0.04	0.11	0.14	0.08	0.07	0.08	0.02	0.03	0.05	0.08
+ Redundancy penalty	Cosine Similarity	0.02	-0.01	0.03	0.10	0.16	0.09	0.16	0.19	0.16	-0.05	0.01	0.03
	ROUGE-1	0.06	0.02	0.07	0.13	0.21	0.14	0.20	**0.26**	0.23	-0.06	0.00	0.00
	MoverScore	0.07	0.03	0.07	0.13	0.16	0.11	0.13	0.13	0.10	0.01	0.03	0.06
	BERTScore	0.03	-0.01	0.05	**0.15**	**0.22**	0.17	**0.21**	0.24	0.22	-0.06	0.00	0.00
STSb-distilbert	w/o redundancy	0.17	0.09	**0.12**	0.06	0.09	0.07	0.06	0.03	-0.01	0.01	0.03	0.04
+ Redundancy penalty	Cosine Similarity	0.16	0.04	0.06	0.07	0.12	0.07	0.14	0.16	0.11	-0.05	-0.03	-0.04
	ROUGE-1	0.16	0.06	0.08	0.10	0.16	0.12	0.17	0.21	0.17	-0.06	-0.04	-0.05
	MoverScore	**0.18**	0.08	0.10	0.08	0.12	0.09	0.09	0.09	0.04	-0.02	0.01	0.01
	BERTScore	0.14	0.03	0.05	0.12	0.18	0.14	0.17	0.20	0.16	-0.06	-0.05	-0.05
$\text{Ensemble}_{\text{sim}}$	w/o redundancy	0.12	0.08	0.07	0.10	0.12	0.07	0.08	0.06	0.00	0.01	0.04	0.05
+ Redundancy penalty	Cosine Similarity	0.11	0.02	0.04	0.10	0.16	0.09	0.16	0.20	0.15	-0.06	-0.01	-0.01
	ROUGE-1	0.13	0.05	0.08	0.13	0.21	0.14	0.20	**0.26**	0.21	-0.05	-0.03	-0.03
	MoverScore	0.14	0.06	0.08	0.12	0.15	0.10	0.14	0.13	0.08	-0.01	0.03	0.03
	BERTScore	0.10	0.03	0.05	**0.15**	**0.22**	0.16	**0.21**	0.25	0.20	-0.05	-0.02	-0.03

Table 1: Kendall's tau (τ) correlation coefficients of expert annotations for different embedding variations of *Sem-nCG* along with various redundancy penalties when $\lambda = 0.5$. Low overlapping reference (LOR), medium overlapping CNN/DailyMail reference (MOR), and high overlapping reference (HOR) were chosen from 11 reference summaries per example to demonstrate the correlation. The highest value in each column is in bold green.

	Consistency			Relevance			Coherence			Fluency		
	LOR	MOR	HOR	LOR	MOR	HOR	LOR	MOR	HOR	LOR	MOR	HOR
ROUGE-1	0.08	0.05	0.01	0.07	0.21	0.22	0.03	0.13	0.13	0.05	0.05	0.05
ROUGE-L	0.02	0.06	-0.01	0.03	0.19	0.15	-0.02	0.14	0.08	0.01	0.04	-0.07
BERTScore	0.06	0.10	0.07	0.10	0.18	0.20	0.06	0.15	0.11	0.08	0.05	0.04

Table 2: Kendall's tau correlation coefficients of ROUGE and BERTScore for Low overlapping reference (LOR), medium overlapping CNN/DailyMail reference (MOR), and high overlapping reference (HOR) chosen from 11 reference summaries per example to demonstrate the correlation.

that higher lambda ($\lambda \geq 0.6$) achieves better correlation for the *Consistency* dimensions, which makes sense because higher lambda means giving more weight to *Sem-nCG*. For *Relevance* and *Coherence* dimensions, a lower lambda (λ) value between $[0.3 - 0.5]$ is a better choice as lower λ means more penalty to redundancy. It appears that for *Fluency* all metric variations struggle. It is evident that $\lambda = 0.5$ gives comparable performance in all four quality dimensions (consistency, relevance, coherence and fluency) and thus we recommend using $\lambda = 0.5$ while adopting Equation 3 to compute redundancy-aware *Sem-nCG*. Table 3 shows a qualitative example for the evaluation of a model-extracted summary.

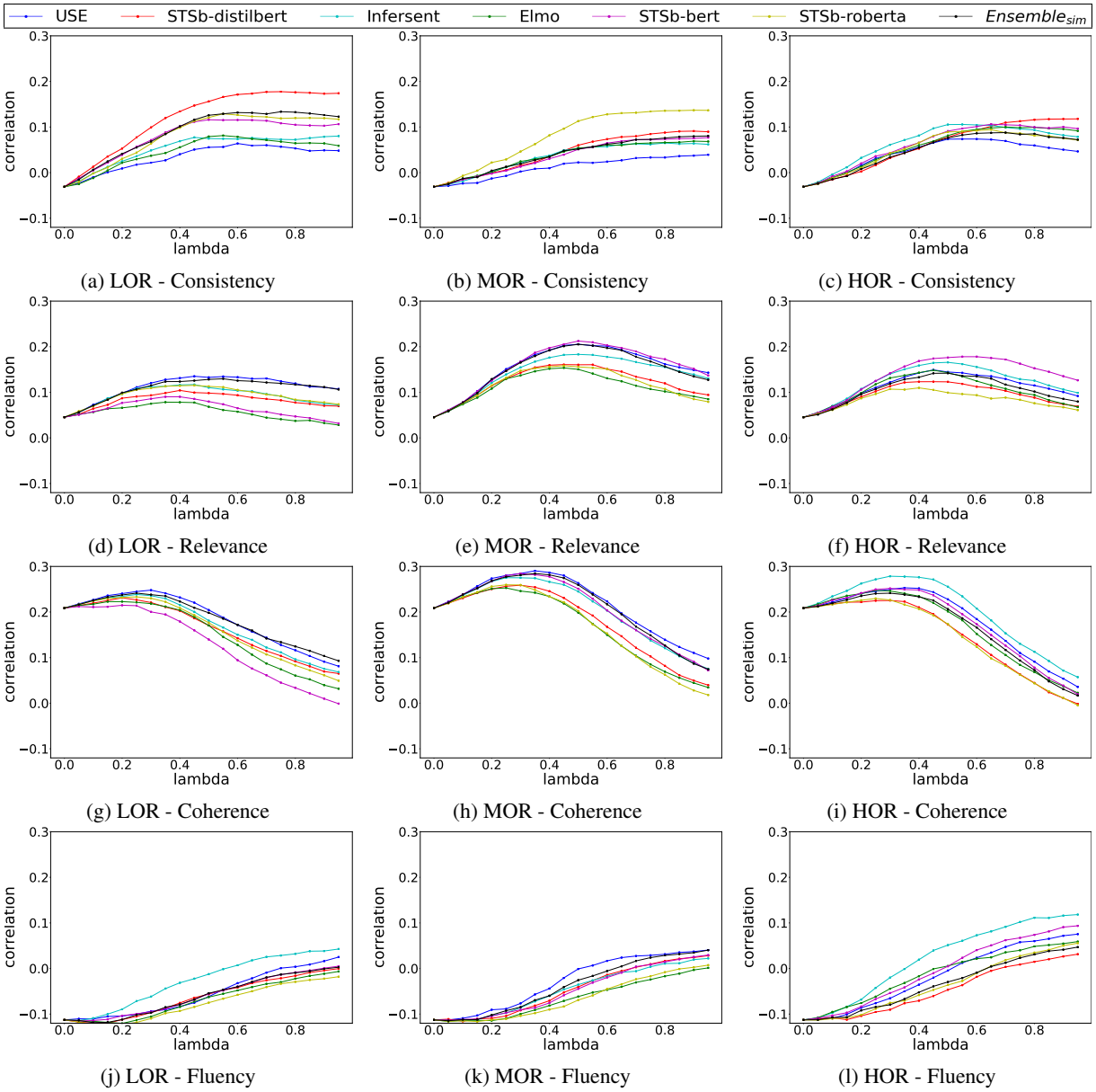


Figure 1: Kendall Tau (τ) Correlation coefficient when lambda (λ) $\in [0, 1]$ from (a)-(c) for Consistency, (d)-(f) for relevance, (g)-(i) for coherence and (j)-(l) for Fluency dimension when ROUGE score is used as redundancy penalty for less overlapping reference (LOR), medium overlapping reference (MOR) and high overlapping reference (HOR).

4.3 Redundancy-aware *Sem-nCG* for Evaluation with Multiple References

SummEval (Fabbri et al., 2021) dataset contains 11 reference summaries. For summary evaluation with multiple references, we considered the lexical overlap of the reference summaries with the original document to demonstrate the terminology variations. Then we considered 3 less overlapping references as Multi-Ref LORs, 3 medium overlapping references as Multi-Ref MORs and 3 high overlapping references as Multi-Ref HORs. We have also mixed up 1 LOR, 1 MOR and 1 HOR and considered this set as Muti-Ref LOR, MOR, HOR to see how the evaluation metric correlates in different terminology variations. Table 4 confirms that ROUGE shows very poor correlation in all the dimensions (consistency, relevance, coherence, and fluency) in all the scenarios and shows slightly better correlation in Multi-Ref HORs (which is somewhat expected as ROUGE considers direct lexical overlap). Interestingly, BERTScore also shows poor correlation in all the settings supporting that the traditional evaluation metric becomes less stable for multiple reference summaries with lots of terminology variations (Cohan and Goharian, 2016).

Article: Last week she was barely showing – but Demelza Poldark is now the proud mother to the show's latest addition. Within ten minutes of tomorrow night's episode, fans will see Aidan Turner's dashing Ross Poldark gaze lovingly at his new baby daughter. As Sunday night's latest heartthrob, women across the country have voiced their longing to settle down with the brooding Cornish gentleman – but unfortunately, it seems as if his heart is well and truly off the market. Scroll down for the video. Last week she was barely showing – but Demelza Poldark is now the proud mother to the show's latest addition He may have married his red-headed kitchen maid out of duty, but as he tells her that she makes him a better man, audiences can have little doubt about his feelings. What is rather less convincing, however, is the timeline of the pregnancy. With the climax of the previous episode being the announcement of the pregnancy, it is quite a jump to the start of tomorrow's installment where Demelza, played by Eleanor Tomlinson, talks about being eight months pregnant. Just minutes after – once again without any nod to the passing of time – she is giving birth, with the last month of her pregnancy passing in less than the blink of an eye. With the climax of the previous episode being the announcement of the pregnancy, it is quite a jump to the start of tomorrow's instalment where Demelza, played by Eleanor Tomlinson, talks about being eight months pregnant As Sunday night's latest heartthrob, women across the country have voiced their longing to settle down with Poldark – but unfortunately, it seems as if his heart is well and truly off the market Their fast relationship didn't go unnoticed by fans. One posted on Twitter: 'If you are pregnant in Poldark times expect to have it in the next 10 minutes' It is reminiscent of the show's previous pregnancy that saw Elizabeth, another contender for Ross's affection, go to full term in the gap between two episodes. This didn't go unnoticed by fans, who posted on Twitter: 'Poldark is rather good, would watch the next one now. Though if you are pregnant in Poldark times expect to have it in the next 10 minutes.
Model Summary: Within ten minutes of tomorrow night's episode, fans will see aidan turner's dashing ross poldark gaze lovingly at his new baby daughter. Last week she was barely showing – but demelza poldark is now the proud mother to the show's latest addition. Last week she was barely showing – but demelza poldark is now the proud mother to the show's latest addition. (clearly redundant extractive summary)
Score$_{red}$ for model summary: 0.40
Less Overlapping Reference (LOR): A celebrity recently welcomed a baby into the world and the wife discusses her experiences with her pregnancy. She has wanted to settle down for a while and is glad her pregnancy wasn't noticeable on television.
Medium Overlapping/CNN Reference (MOR): SPOILER ALERT: Maid gives birth to baby on Sunday's episode. Only announced she was pregnant with Poldark's baby last week.
High Overlapping Reference (HOR): In the latest episode, Demelza Poldark talks about being 8 months pregnant. Ross Poldark, who is off the market and in love with Demelza, will be shown gazing lovingly at his new baby daughter tomorrow night.
Sem-nCG Score only according to equation 1 for LOR: 0.67 MOR: 0.733 HOR: 0.8
Revised Sem-nCG Score along with Score$_{red}$ according to equation 3 for LOR: 0.532 MOR: 0.565 HOR: 0.599
Human Evaluation (annotated by experts and score ranged between 0-1) Coherence: 0.47 Consistency: 1 Fluency: 1 Relevance: 0.67

Table 3: An example of the model summary evaluation using the redundancy-aware Sem-nCG metric.

Metric	Multi-Ref LOR, MOR, HOR				Multi-Ref LORs				Multi-Ref MORs				Multi-Ref HORs			
	Con	Rel	Coh	Flu	Con	Rel	Coh	Flu	Con	Rel	Coh	Flu	Con	Rel	Coh	Flu
ROUGE-1	0.00	-0.01	-0.09	-0.01	-0.05	0.05	0.00	0.01	-0.05	0.09	0.04	-0.01	-0.02	0.21	0.13	0.10
ROUGE-L	0.00	-0.01	-0.09	-0.01	0.00	0.04	-0.01	0.01	-0.06	0.07	0.04	0.00	-0.01	0.15	0.09	-0.04
BERTScore	0.09	0.19	0.14	0.03	0.01	0.07	-0.01	0.04	-0.04	0.05	0.03	0.05	0.04	0.20	0.12	0.06

Table 4: Kendall Tau (τ) correlation coefficient for ROUGE and BERTScore for consistency (con), relevance (rel), coherence (coh) and fluency (flu) dimension for evaluating extractive model summaries with multiple references.

In the original *Sem-nCG* metric, a groundtruth ranking is prepared by considering the cosine similarity between each sentence of the document and reference summary but the evaluation with multiple-reference was left as future work. As a starting point, how to incorporate multiple-reference summaries in the original *Sem-nCG* metric, we designed how to create the groundtruth ranking by considering multiple references. Here, we took the naive approach, first computing cosine similarity of each sentence of the document with each reference among multiple references. Then average it, which we called Ensemble$_{sim}$. For Ensemble$_{rel}$, for each groundtruth ranking prepared for each reference among multiple reference summaries, we took the average of relevance (as it was computed in previously proposed *Sem-nCG* metric (Akter et al., 2022)) and based on that we merged the groundtruth rankings into one groundtruth ranking. Then we use this groundtruth ranking to compute *Sem-nCG* for model extracted summary. With the original Sem-nCG metric, we have also incorporated redundancy into the *Sem-nCG* metric utilizing equation 3. We have only considered ROUGE and BERTScore as redundancy penalty both in Table 5 and 6 when $\lambda = 0.5$ (as evident from Section 4.2 that this setting gives better performance). We have also considered different embedding variations to create the groundtruth ranking.

From Table 5, we can see that redundancy-aware *Sem-nCG* shows better correlations for all the scenarios (multi-ref LORs, multi-ref MORs, multi-

Multi-Ref LOR, MOR, HOR

Embedding	w/o Redundancy				+ Redundancy Penalty (ROUGE)				+ Redundancy Penalty (BERTScore)			
	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency
Infersent	0.07	0.11	0.08	**0.06**	0.11	0.18	**0.27**	0.01	0.09	0.18	0.20	0.03
Elmo	0.09	0.06	0.01	0.00	0.09	0.12	0.18	-0.05	0.09	0.12	0.11	-0.03
STSb-bert	0.10	0.12	0.04	0.06	0.09	0.19	0.24	-0.02	0.10	**0.20**	0.18	0.01
STSb-roberta	0.14	0.10	0.01	0.02	0.12	0.17	0.21	-0.06	**0.13**	0.17	0.13	-0.02
USE	0.04	0.12	0.08	0.05	0.06	0.19	0.26	-0.03	0.05	0.19	0.20	0.01
STSb-distilbert	0.14	0.13	0.05	0.02	0.10	0.19	0.23	-0.04	0.12	**0.20**	0.17	-0.01

Multi-Ref LORs

Embedding	w/o Redundancy				+ Redundancy Penalty (ROUGE)				+ Redundancy Penalty (BERTScore)			
	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency
Infersent	0.03	0.10	0.09	**0.07**	0.05	**0.16**	**0.25**	0.02	0.02	0.15	0.18	0.04
Elmo	0.04	0.05	-0.04	0.03	0.05	0.12	0.15	-0.04	0.03	0.11	0.06	-0.01
STSb-bert	0.08	0.10	0.02	0.01	0.09	0.15	0.20	-0.06	0.06	0.15	0.13	-0.04
STSb-roberta	0.10	0.07	-0.04	0.00	**0.11**	0.15	0.17	-0.07	0.09	0.15	0.09	-0.04
USE	0.02	0.05	0.01	0.03	0.04	0.12	0.19	-0.04	0.02	0.10	0.12	0.00
STSb-distilbert	0.10	0.04	-0.02	-0.02	**0.11**	0.09	0.15	-0.09	0.09	0.09	0.09	-0.07

Multi-Ref MORs

Embedding	w/o Redundancy				+ Redundancy Penalty (ROUGE)				+ Redundancy Penalty (BERTScore)			
	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency
Infersent	0.08	0.08	0.03	**0.06**	0.10	**0.15**	**0.23**	-0.02	0.08	**0.15**	0.16	0.02
Elmo	0.06	0.05	-0.02	0.00	0.04	0.13	0.16	-0.07	0.05	0.11	0.08	-0.05
STSb-bert	0.07	0.05	0.02	0.01	0.09	0.13	0.22	-0.08	0.07	0.12	0.15	-0.04
STSb-roberta	0.05	0.07	-0.01	0.02	0.07	0.14	0.21	-0.07	0.04	0.14	0.14	-0.03
USE	0.02	0.08	0.05	0.01	0.04	0.15	**0.25**	-0.06	0.02	0.14	0.17	-0.03
STSb-distilbert	**0.11**	0.01	0.00	-0.01	0.09	0.07	0.17	-0.10	0.10	0.06	0.10	-0.05

Multi-Ref HORs

Embedding	w/o Redundancy				+ Redundancy Penalty (ROUGE)				+ Redundancy Penalty (BERTScore)			
	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency
Infersent	0.07	0.08	0.05	0.03	**0.11**	0.16	0.23	-0.02	0.07	0.15	0.15	0.01
Elmo	0.04	0.09	0.02	0.06	0.06	0.16	0.19	0.00	0.04	0.14	0.11	0.03
STSb-bert	0.08	0.11	0.04	0.05	0.12	0.18	**0.24**	-0.03	0.08	0.18	0.16	0.01
STSb-roberta	0.10	0.09	0.01	0.03	0.14	0.17	0.22	-0.04	0.10	0.16	0.13	0.00
USE	0.04	0.14	0.07	0.05	0.07	0.20	**0.24**	-0.03	0.04	**0.21**	0.18	0.01
STSb-distilbert	0.08	0.09	0.02	0.05	**0.11**	0.15	0.22	-0.03	0.09	0.15	0.14	0.02

Table 5: Kendall Tau (τ) correlation coefficient for **Ensemble$_{\text{sim}}$** when lambda (λ) $= 0.5$ for consistency, relevance, coherence and fluency dimension without redundancy and when ROUGE and BERTScore is used as redundancy penalty for different terminology variations of multiple references (highly abstractive (LORs), medium overlapping (MORs) and highly extractive (HORs) references). The best value in each dimension has been bold green.

ref HORs and mixture of LOR, MOR & HOR). Both ROUGE and BERTScore provide comparable results for self-referenced redundancy penalties, with ROUGE score-based redundancy providing a marginally superior result. Interestingly, redundancy-aware *Sem-nCG* shows robust performance in all the scenarios while showing 25% improvement in coherence and 10% improvement in relevance dimension. Same patterns are observed when Ensemble$_{\text{rel}}$ is also used for the evaluation of multiple reference (See Table 6).

From our empirical evaluation, we would recommend USE embedding to create Ensemble$_{\text{sim}}$ (merging sentence-wise similarities across different references) with ROUGE redundancy penalty to evaluate extractive summary with multiple references.

5 Related Work

The most common method for evaluating model summaries has been to compare them against human-written reference summaries. ROUGE (Lin, 2004b) considers direct lexical overlap and afterwards different version of ROUGE (Graham, 2015) has also been proposed 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) to mitigate the limitations of original ROUGE. Metrics based on semantic similarity between reference and model summaries have also been proposed to capture the semantics, including S+WMS (Clark et al., 2019), MoverScore (Zhao et al., 2019), and BERTScore (Zhang et al., 2020). Reference-free evaluation has also been a recent trend to avoid dependency on human reference (Böhm et al., 2019; Peyrard, 2019; Sun and Nenkova, 2019; Gao et al., 2020; Wu et al., 2020).

Although the *extractive* summarizing task is typically framed as a sentence ranking problem, none of the mentioned metrics evaluate the quality of the ranker. To address this, recently (Akter et al., 2022) has proposed a rank-aware and gain-based evaluation metric for extractive summarization called *Sem-nCG*, but it does not incorporate redundancy and also lacks evaluation with multiple references. These are two significant limitations that need to be addressed, and hence, the focus of this work.

Multi-Ref LOR, MOR, HOR

Embedding	w/o Redundancy				+ Redundancy Penalty (ROUGE)				+ Redundancy Penalty (BERTScore)			
	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency
Infersent	0.09	0.10	0.04	0.08	0.11	0.17	0.24	0.01	0.09	0.18	0.18	0.04
Elmo	0.09	0.06	0.02	0.01	0.09	0.13	0.20	-0.05	0.09	0.12	0.13	-0.03
STSb-bert	0.12	0.15	0.04	0.05	0.12	0.22	0.24	-0.03	0.12	0.24	0.18	0.01
STSb-roberta	**0.14**	0.08	0.01	0.01	0.13	0.15	0.21	-0.05	0.13	0.15	0.12	-0.02
USE	0.04	0.16	0.11	**0.08**	0.05	0.21	**0.29**	0.00	0.04	0.22	0.24	0.05
STSb-distilbert	0.14	0.10	0.03	0.02	0.10	0.16	0.22	-0.04	0.11	0.18	0.16	-0.01

Multi-Ref LORs

Embedding	w/o Redundancy				+ Redundancy Penalty (ROUGE)				+ Redundancy Penalty (BERTScore)			
	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency
Infersent	0.03	0.09	0.07	**0.08**	0.05	0.15	**0.23**	0.04	0.02	0.14	0.16	0.05
Elmo	0.03	0.04	-0.04	0.03	0.04	0.10	0.14	-0.03	0.03	0.09	0.06	-0.01
STSb-bert	0.09	0.10	0.00	0.01	0.10	**0.16**	0.19	-0.06	0.09	0.17	0.13	-0.03
STSb-roberta	0.10	0.05	-0.06	0.00	0.11	0.13	0.15	-0.08	0.09	0.12	0.07	-0.04
USE	0.04	0.08	0.03	0.04	0.05	0.14	0.22	-0.04	0.03	0.13	0.15	0.01
STSb-distilbert	**0.13**	0.06	0.01	-0.01	0.12	0.11	0.17	-0.09	0.12	0.12	0.12	-0.06

Multi-Ref MORs

Embedding	w/o Redundancy				+ Redundancy Penalty (ROUGE)				+ Redundancy Penalty (BERTScore)			
	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency
Infersent	0.06	0.10	0.05	0.06	0.07	0.19	**0.26**	-0.01	**0.06**	0.18	**0.19**	0.02
Elmo	0.06	0.06	0.00	0.02	0.04	0.13	0.17	-0.06	0.04	0.12	0.11	-0.02
STSb-bert	0.08	0.01	-0.02	0.01	**0.09**	0.09	0.18	-0.08	0.08	0.08	0.11	-0.04
STSb-roberta	0.05	0.07	0.00	0.02	0.06	0.14	0.20	-0.07	0.05	0.14	0.13	-0.02
USE	0.01	0.09	0.05	0.01	0.04	0.16	0.24	-0.05	0.01	0.16	0.19	-0.02
STSb-distilbert	0.08	0.02	0.00	-0.01	0.07	0.09	0.18	-0.09	0.07	0.08	0.12	-0.06

Multi-Ref HORs

Embedding	w/o Redundancy				+ Redundancy Penalty (ROUGE)				+ Redundancy Penalty (BERTScore)			
	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency	Consistency	Relevance	Coherence	Fluency
Infersent	0.09	0.11	0.06	0.05	0.13	0.18	0.25	-0.01	0.09	0.18	0.18	0.02
Elmo	0.05	0.08	0.02	0.05	0.07	0.16	0.19	-0.01	0.05	0.14	0.12	0.02
STSb-bert	0.07	0.11	0.04	0.05	0.11	0.18	0.25	-0.02	0.06	0.19	0.17	0.02
STSb-roberta	0.10	0.08	0.01	0.04	**0.13**	0.16	0.21	-0.04	0.11	0.15	0.13	0.00
USE	0.06	0.13	0.07	**0.05**	0.09	**0.20**	**0.26**	-0.02	0.06	0.20	0.19	0.02
STSb-distilbert	0.09	0.09	0.03	0.03	0.12	0.15	0.22	-0.05	0.10	0.15	0.15	0.00

Table 6: Kendall Tau (τ) correlation coefficient for **Ensemble$_{\text{rel}}$** when lambda (λ) $= 0.5$ for consistency, relevance, coherence and fluency dimension without redundancy and when ROUGE and BERTScore is used as redundancy penalty for different terminology variations of multiple references (highly abstractive (LORs), medium overlapping (MORs) and highly extractive (HORs) references). The best value in each dimension has been bold green.

Redundancy in extracted sentences is a prominent issue in extractive summarization systems. Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) is a classic algorithm to penalize redundancy in model summary. There are several approaches that explicitly model redundancy and use algorithms to avoid selecting sentences that are too similar to those that have already been extracted (Ren et al., 2016). Trigram blocking (Paulus et al., 2018) is another popular approach to reduce redundancy in model summary. Chen et al. (2021) has shown how to compute self-referenced redundancy score while evaluating the model summary.

When multiple reference summaries are available, Researchers have also suggested Pyramid-based (Nenkova and Passonneau, 2004) approaches for summary evaluation. However, this method requires more manual labor and has undergone numerous improvements (Passonneau et al., 2013; Yang et al., 2016; Shapira et al., 2019; Mao et al., 2020), it still needs a substantial amount of manual effort, making it unsuitable for large-scale evaluation. Recently, for NLG evaluation different unified frameworks and models (Deng et al., 2021; Zhong et al., 2022; Liu et al., 2023a; Wu et al., 2024) to predict different aspects of the generated text has been proposed. Even though these metrics can be applied to text summarization, it is still a data-driven approach and it is unclear why the model produces such scores.

Uniqueness to our work: We improved the sem-nCG metric for extractive summarization, in order to make it more aware of redundancy. This was tricky, as it requires a balance of importance and diversity during evaluation. We also showed how to use the updated metric for multiple references, which was challenging due to variations in human references and terminology.

6 Conclusion

Previous work has proposed the *Sem-nCG* metric exclusively for evaluating extractive summarization task considering both rank awareness and semantics. However, the *Sem-nCG* metric ignores redundancy in a model summary and does not support evaluation with multiple reference summaries, which are two significant limitations. In this paper, we proposed a redundancy-aware multi-reference based *Sem-nCG* metric which is superior compared to the previous *Sem-nCG* metric along *Consistency*,

Relevance and *Coherence* dimensions. Additionally, for summary evaluation using multiple references, we created a unique ground-truth ranking by incorporating multiple references rather than trivial max/average score computation with multiple references. Our empirical evaluation shows that the traditional metric becomes unstable when multiple references are available and the revised redundancy-aware *Sem-nCG* shows a notably higher correlation with human judgments than ROUGE and BERTScore metric both for single and multiple references. Thus we encourage the community to evaluate extractive summaries using the revised redundancy-aware *Sem-nCG* metric.

7 Limitations

One limitation of the work is that the dataset for human evaluation is not big (252 samples). We used the dataset from (Fabbri et al., 2021), the only available benchmark "meta-evaluation dataset" for extractive summarization, to the best of our knowledge. (Akter et al., 2022) have demonstrated the correlation of *Sem-nCG* with human judgment on this dataset. To ensure a fair comparison, we maintained the same settings as the original *Sem-nCG* when assessing the redundancy-aware *Sem-nCG*. To evaluate the redundancy-aware *Sem-nCG* we will need a similar kind of evaluation benchmark and we can not do anything here. Even though (Liu et al., 2023b) has published a new dataset, that work focuses mainly on how to increase human annotation reliability for summary evaluation with respect to Atomic Content Unit (ACU) and doesn't provide human judgment for model's summary along four summary quality dimensions: coherence, consistency, fluency and relevance.

Another limitation of the work may seem like that the ablation study does not show any consistent pattern. We understand that it's difficult to come up with a single evaluation metric that can account for different qualities such as coherence, consistency, fluency, and relevance. It requires careful consideration to balance these different qualities. However, we noticed that extractive sentences are inherently grammatically correct, so we can exclude fluency from the hyperparameter choice. After analyzing the data, we found that a balanced λ value of 0.5 worked well across all four quality dimensions. This suggests that this configuration strikes a reasonable tradeoff between importance and diversity. It addresses the complexities inherent in assessing summarization quality comprehensively with a single score from the metric.

8 Ethics Statement

For the experiments, we used a publicly accessible dataset and anonymous human annotations. As a result, to the best of our knowledge, there are no ethical violations. Additionally, the evaluation of extractive summarization is a major aspect of this work. Hence, we consider it a low-risk research study.

9 Acknowledgements

This work has been partially supported by the Research Center Trustworthy Data Science and Security https://rc-trust.ai, one of the Research Alliance centers within the https://uaruhr.de. This work has also received partial support from the National Science Foundation (NSF) Standard Grant Award #2302974 and the Air Force Office of Scientific Research Grant/Cooperative Agreement Award #FA9550-23-1-0426.

References

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

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. 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 3108–3118. Association for Computational Linguistics.

Jaime G. Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 335–336. ACM.

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. Association for Computational Linguistics.

Wang Chen, Piji Li, and Irwin King. 2021. A training-free and reference-free summarization evaluation metric via centrality-weighted relevance and self-referenced redundancy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 404–414. Association for Computational Linguistics.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. 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 2748–2760. 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.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7580–7605. Association for Computational Linguistics.

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.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1347–1354. Association for Computational Linguistics.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, USA, June 6, 2010*, pages 148–151. Association for Computational Linguistics.

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.

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.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Chin-Yew Lin. 2004a. Looking for a few good metrics: Automatic summarization evaluation - how many samples are enough? In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization, NTCIR-4, National Center of Sciences, Tokyo, Japan, June 2-4, 2004*. National Institute of Informatics (NII).

Chin-Yew Lin. 2004b. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*,

pages 2511–2522. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

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. 2005. Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *Proceedings, The Twentieth National Conference on Artificial Intelligence and the Seventeenth Innovative Applications of Artificial Intelligence Conference, July 9-13, 2005, Pittsburgh, Pennsylvania, USA*, pages 1436–1441. AAAI Press / The MIT Press.

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.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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. A simple theoretical model of importance for 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 1059–1073. Association for Computational Linguistics.

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 and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Pengjie Ren, Furu Wei, Zhumin Chen, Jun Ma, and Ming Zhou. 2016. A redundancy-aware sentence regression framework for extractive summarization. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 33–43. ACL.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 1216–1221. Association for Computational Linguistics.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free

summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3612–3621. Association for Computational Linguistics.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is more for long document summary evaluation by llms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 2: Short Papers, St. Julian's, Malta, March 17-22, 2024*, pages 330–343. 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.

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 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. 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.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *CoRR*, abs/2210.07197.

A Appendix

A.1 Explanation of Metrics for $Score_{red}$

ROUGE (Lin, 2004b): Between the generated summary and reference summary, ROUGE counts the overlap of textual units (n-grams, word sequences).

MoverScore (Zhao et al., 2019): uses the Word Mover's Distance (Kusner et al., 2015) to calculate the semantic distance between a summary and a reference text, pooling n-gram embedding from BERT representations.

BERTScore (Zhang et al., 2020): calculates similarity scores by matching generated and reference summaries on a token level. The cosine similarity between contextualized token embeddings from BERT is maximized by computing token matching greedily.

Cosine Similarity: Sentences are converted to sentence embedding using STSb-distilbert (Reimers and Gurevych, 2019). Then the semantic similarity of sentences is measured using cosine similarity between sentence vectors.

The code for the metrics used can be found here.[3]

A.2 Human Evaluation Components

To calculate the Kendall's Tau (τ) rank correlation for the redundancy-aware *Sem-nCG* metric, we used four quality dimensions following (Akter et al., 2022; Fabbri et al., 2021).

Consistency: refers to the fact that the contents in the summary and the source are the same. Only assertions from the source are included in factually consistent summaries, which do not include any trippy facts.

Relevance: getting the most important information from a source. The annotators were to penalize summaries with redundancy and excessive information. In the summary, only important information from the source should be included.

Coherence: overall summary sentence quality while keeping a coherent body of information on a topic rather than a tangle of related information (Dang, 2005).

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

A.3 Computational Infrastructure & Runtime

A.4 Sentence Embedding Used in Setion 4

Infersent (Conneau et al., 2017): Infersent-v2 is trained with fastText word embedding and generates 4096-dimensional sentence embedding using a BiLSTM network with max-pooling.

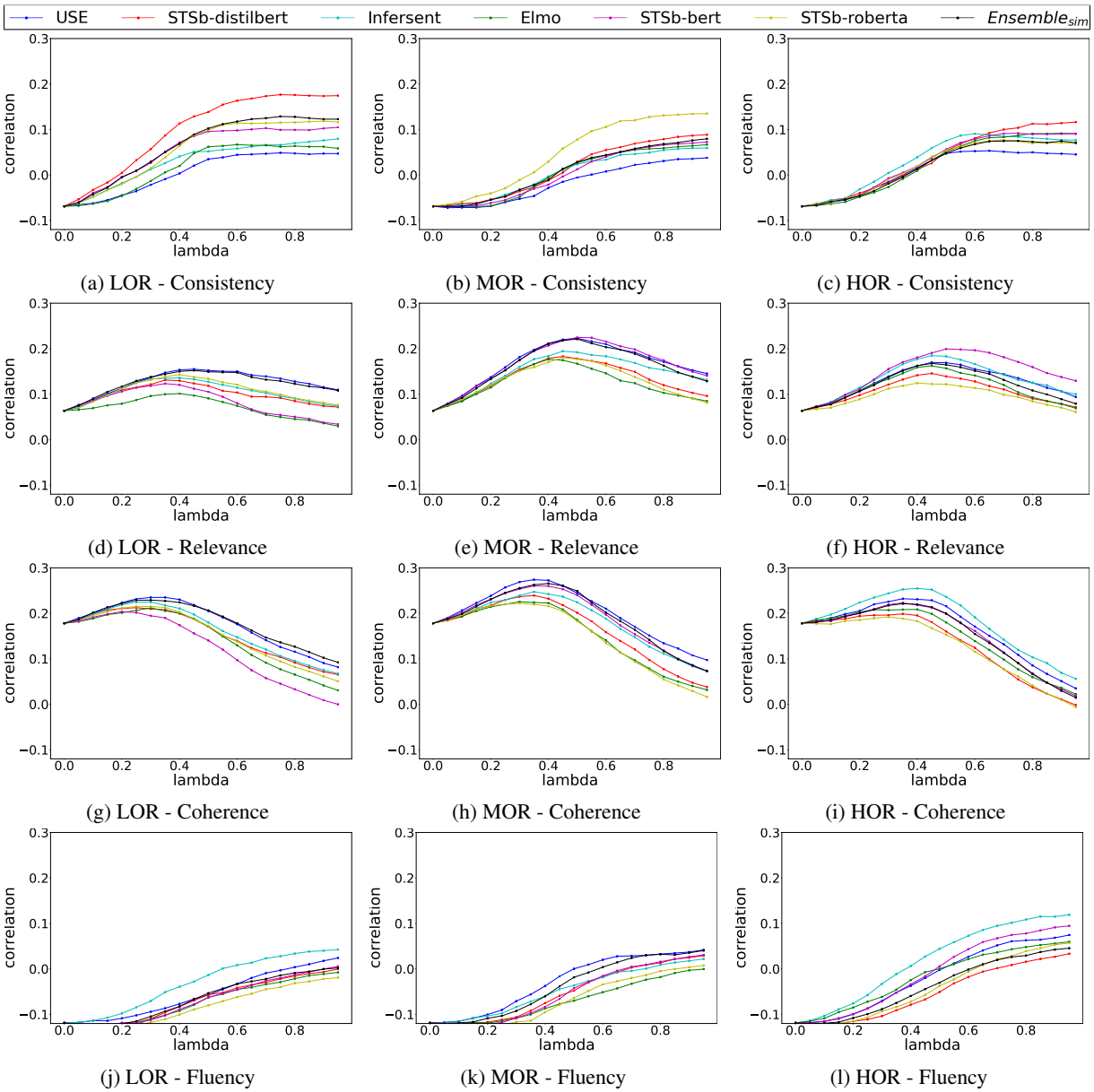
[3] https://github.com/Yale-LILY/SummEval/tree/master/evaluation/summ_eval

Computational Infrastructure		
NVIDIA Quadro RTX 5000 GPUs		
Hyperparameter Search		
$\lambda \in [0, 1]$ uniform-integer distribution		
Type	**Variation**	**Runtime (s)**
$Score_{red}$	Cosine Similarity	0.06
	ROUGE	0.44
	MoverScore	0.23
	BERTScore	14.7
Sem-nCG	Infersent	0.4
	Elmo	79.1
	STSb-bert	0.33
	STSb-roberta	0.34
	USE	20.2
	STSb-distilbert	0.13
	$\text{Ensemble}_{\text{sim}}$	20.33

Elmo (Peters et al., 2018): The contextualized word embedding was transformed into a sentence embedding using a fixed mean-pooling of all contextualized word representations with embedding shape 1024.

Google Universal Sentence Encoder (USE) (Cer et al., 2018): We utilized USE with enc-2 (Iyyer et al., 2015) which is based on the deep average network to transform input text to a 512-dimensional sentence embedding.

Semantic Textual Similarity benchmark (STSb) (Reimers and Gurevych, 2019): Sentence Transformer allows to generate dense vector representations of sentences. Three of the best available models that were optimized for semantic textual similarity were considered: STSb-bert (embedding size 1024), STSb-roberta (embedding size 1024) and STSb-distilbert (embedding size 768).



(a) LOR - Consistency (b) MOR - Consistency (c) HOR - Consistency

(d) LOR - Relevance (e) MOR - Relevance (f) HOR - Relevance

(g) LOR - Coherence (h) MOR - Coherence (i) HOR - Coherence

(j) LOR - Fluency (k) MOR - Fluency (l) HOR - Fluency

Figure 2: Kendall Tau (τ) correlation coefficient when lambda (λ) $\in [0, 1]$ from (a)-(c) for consistency, (d)-(f) for relevance, (g)-(i) for coherence and (j)-(l) for fluency dimension when BERTScore is used as redundancy penalty for less overlapping reference (LOR), medium overlapping reference (MOR) and high overlapping reference (HOR).