Facet-Aware Evaluation for Extractive Summarization

Yuning Mao[1], **Liyuan Liu**[1], **Qi Zhu**[1], **Xiang Ren**[2], **Jiawei Han**[1]
[1]Department of Computer Science, University of Illinois at Urbana-Champaign, IL, USA
[2]Department of Computer Science, University of Southern California, CA, USA
[1]{yuningm2, ll2, qiz3, hanj}@illinois.edu [2]xiangren@usc.edu

Abstract

Commonly adopted metrics for extractive summarization focus on lexical overlap at the token level. In this paper, we present a facet-aware evaluation setup for better assessment of the information coverage in extracted summaries. Specifically, we treat each sentence in the reference summary as a *facet*, identify the sentences in the document that express the semantics of each facet as *support sentences* of the facet, and automatically evaluate extractive summarization methods by comparing the indices of extracted sentences and support sentences of all the facets in the reference summary. To facilitate this new evaluation setup, we construct an extractive version of the CNN/Daily Mail dataset and perform a thorough quantitative investigation, through which we demonstrate that facet-aware evaluation manifests better correlation with human judgment than ROUGE, enables fine-grained evaluation as well as comparative analysis, and reveals valuable insights of state-of-the-art summarization methods.[1]

1 Introduction

Text summarization has enjoyed increasing popularity due to its wide applications, whereas the evaluation of text summarization remains challenging and controversial. The most commonly used evaluation metric of summarization is lexical overlap, *i.e.*, ROUGE (Lin, 2004), which regards the system and reference summaries as sequences of tokens and measures their n-gram overlap.

However, recent studies (Paulus et al., 2017; Schluter, 2017; Kryscinski et al., 2019) reveal the limitations of ROUGE and find that in many cases, it fails to reach consensus with human judgment. Since lexical overlap only captures information coverage at the surface (token) level, ROUGE favors system summaries that share more tokens with the reference summaries. Nevertheless, such summaries may not always convey the desired semantics. For example, in Table 1, the document sentence with the highest ROUGE score has more lexical overlap but expresses rather different semantic meaning. In contrast, the sentence manually extracted from the document by our annotators, which conveys similar semantics, is over-penalized as it involves other details or uses alternative words.

Reference: Three people in **Kansas** have died from a **listeria outbreak**. **Lexical Overlap**: But they did not appear identical to **listeria** samples taken from patients infected in the **Kansas outbreak**. (***ROUGE-1 F1=37.0, multiple token matches but totally different semantics***) **Manual Extract**: Five people were infected and **three died** in the past year **in Kansas from listeria** that might be linked to blue bell creameries products, according to the CDC. (***ROUGE-1 F1=36.9, semantics covered but lower ROUGE due to the presence of other details)***
Reference: Chelsea boss **Jose Mourinho** and United manager **Louis van Gaal** are pals. **Lexical Overlap**: Gary Neville believes **Louis van Gaal**'s greatest achievement as a football manager is the making of **Jose Mourinho**. **Manual Extract**: The duo have been friends since they first worked together at Barcelona in 1997 where they enjoyed a successful relationship at the Camp Nou. (***ROUGE Recall/F1=0, no lexical overlap at all***)

Table 1: Lexical overlap — finding the document sentence with the highest ROUGE against one reference sentence — could be misleading. Examples are from the CNN/Daily Mail dataset (Nallapati et al., 2016).

In this paper, we argue that the information coverage in summarization can be better evaluated by *facet overlap*, *i.e.*, whether the system summary covers the facets in the reference summary. Specifically, we treat each *reference sentence* as a facet, identify *document sentences* that express the semantics of each facet as *support sentences* of the facet, and measure information coverage by Facet-Aware Recall (**FAR**), *i.e.*, how many facets are covered. We focus on extractive summarization for the following two reasons. Theoretically, since extractive methods cannot paraphrase or compress the document sentences as abstractive methods, it is somewhat unfair to penalize them for extracting long sentences that cover the facets. Pragmatically,

[1]Data can be found at https://github.com/morningmoni/FAR.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4941–4957
July 5 - 10, 2020.

we can evaluate extractive methods automatically by comparing the indices of extracted sentences and support sentences. We denote the mappings from each facet (sentence) in the reference summary to its support sentences in the document as Facet-Aware Mappings (**FAMs**). FAMs can be used as labels indicating which sentences should be extracted but they are grouped with respect to each facet, while conventional extractive labels correspond to the entire reference summary rather than individual facets (detailed explanations in Sec. 2.1). Compared to treating one summary as a sequence of n-grams, *facet-aware evaluation* considers information coverage at a semantically richer granularity, and thus can contribute to a more accurate assessment on the summary quality.

To verify the effectiveness of facet-aware evaluation, we construct an *extractive* version of the CNN/Daily Mail dataset (Nallapati et al., 2016) by annotating its FAMs (Sec. 2). We revisit state-of-the-art extractive methods using this new extractive dataset (Sec. 3.2), the results of which show that FAR correlates better with human evaluation than ROUGE. We also demonstrate that FAMs are beneficial for fine-grained evaluation of both abstractive and extractive methods (Sec. 3.3). We then illustrate how facet-aware evaluation can be useful for comparing different extractive methods in terms of their capability of extracting salient and non-redundant sentences (Sec. 3.4). Finally, we explore the feasibility of automatic FAM creation by evaluating sentence regression approaches against the ground-truth annotations (*i.e.*, FAMs), and generalize facet-aware evaluation to the entire CNN/Daily Mail dataset *without any human annotation* (Sec. 4). We believe that the summarization community will benefit from the proposed setup for better assessment of information coverage and gain deeper understandings of the current benchmark dataset and state-of-the-art methods through our analysis.

Contributions. (1) We propose a facet-aware evaluation setup that better assesses information coverage for extractive summarization. (2) We build the first dataset designed specifically for extractive summarization by creating facet-aware mappings from reference summaries to documents. (3) We revisit state-of-the-art summarization methods in the proposed setup and discover valuable insights. (4) To our knowledge, our work is also the first thorough quantitative analysis regarding the characteristics of the CNN/Daily Mail dataset.

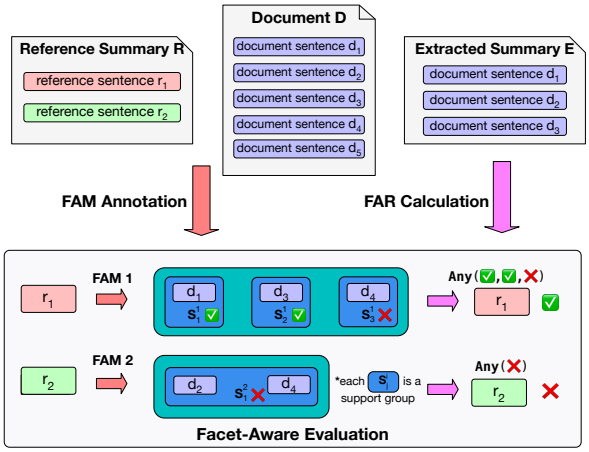

Figure 1: **An illustration of facet-aware evaluation**. Two of three support groups of facet 1 (r_1) are covered. Facet 2 (r_2) cannot be covered as document sentence 4 (d_4) is missing in the extracted summary. The illustration corresponds to the example in Sec. 3.1.

2 Dataset Creation

In this section, we describe the process of creating an *extractive* summarization dataset to facilitate facet-aware evaluation, which involves annotating FAMs between the documents and *abstractive* reference summaries. We first formalize the FAMs and then describe the FAM annotation on the CNN/Daily Mail dataset (Nallapati et al., 2016).

2.1 FAMs: Facet-Aware Mappings

We denote one document-summary pair as $\{\mathbf{D}, \mathbf{R}\}$, where $\mathbf{D} = [d_1, d_2, ..., d_D]$, $\mathbf{R} = [r_1, r_2, ..., r_R]$, and D, R denote the numbers of document sentences and reference sentences, respectively. We conceptualize *facet* as one unique semantic aspect presented in the summary. In practice, we hypothesize that each reference sentence r_i corresponds to one facet.[2] We define *support sentences* as the sentences in the document that express the semantics of one facet r_i. We define *support group* $\mathcal{S}$ of facet r_i as a set of support sentences that can fully cover the information of r_i. For each facet r_i in the reference summary, we try to find all its support sentences in the document and put them into support groups. Since we focus on single-document

[2]It is possible to define facet at sub-sentence or multi-sentence level as in Pyramid (Nenkova and Passonneau, 2004). However, such definitions inevitably incur more annotation effort and lower inter-annotator agreement, while the current definition balances cost and effectiveness.

Category	#Samples	#Facets	Example (full documents, reference summaries, and the FAMs can be found in Appendix C)
Noise (**N**)	41 (27.3%)	137 (27.1%)	• **Reference**: "Furious 7" opens Friday. (**unimportant detail**) • **Reference**: Click here for all the latest Floyd Mayweather vs Manny Pacquiao news. (**not found in the document**) • **Reference**: Vin Diesel: "This movie is more than a movie". (**random quotation**) • **Reference**: "I had a small moment of awe," she said. (**random quotation**)
Low Abstraction (**L**)	89 (59.3%)	310 (61.2%) $\overline{M}$=1: 275 (88.7%) $\overline{M}$=2: 35 (11.3%)	• **Reference**: Willis never trademarked her most-famous work, calling it "my gift to the city". • **Support**: Willis never trademarked her most-famous work, calling it "my gift to the city." (**identical**) • **Reference**: Thomas K. Jenkins, 49, was arrested last month by deputies with the Prince George's County sheriff's office, authorities said. • **Support**: Authorities said in a news release Thursday that 49-year-old Thomas K. Jenkins of capitol heights, Maryland, was arrested last month by deputies with the Prince George's County sheriff's office. (**compression**)
High Abstraction (**H**)	20 (13.3%)	59 (11.7%)	• **Reference**: College-bound basketball star asks girl with down syndrome to high school prom. Pictures of the two during the "prom-posal" have gone viral. (**highly abstractive**) • **Reference**: While Republican Gov. Asa Hutchinson was weighing an Arkansas religious freedom bill, Walmart voiced its opposition. Walmart and other high-profile businesses are showing their support for gay and lesbian rights. (**unable to find support sentences**)

Table 2: **Category breakdown of Facet-Aware Mappings (FAMs)**. Nearly 60% samples are of low abstraction while more than a quarter of samples contain noisy facets. $\overline{M}$ denotes the average number of support sentences.

summarization in this work, most facets only have one support group. But some may contain multiple and extracting any of them would suffice (see example in Appendix C Table 10). Allowing multiple support groups also makes FAMs easily extendable to multi-document summarization where redundant sentences prevail.

Formally, for each r_i, we annotate a Facet-Aware Mapping (FAM) $r_i \rightarrow \{\mathcal{S}_1^i, \mathcal{S}_2^i, ..., \mathcal{S}_N^i\}$, where N is the number of support groups. Each $\mathcal{S}_j^i = \{d_{I_1}, d_{I_2}, ..., d_{I_{M_j}}\}$ is a support group, where $I_1, I_2, ..., I_{M_j}$ are the indices of support sentences and M_j is the number of support sentences in $\mathcal{S}_j^i$. One illustrative example is presented in Fig. 1. The support sentences are likely to be verbose, but we consider whether the support sentences express the semantics of the facet regardless of their length.[3] The reason is that we believe extractive summarization should focus on information coverage since it cannot alter the original sentences and once salient sentences are extracted, one can then compress them in an abstractive manner (Chen and Bansal, 2018; Hsu et al., 2018).

Relation w. Extractive Labels. Extractive methods (Nallapati et al., 2017; Chen and Bansal, 2018; Narayan et al., 2018c) typically require binary labels of every document sentence indicating whether it should be extracted during model training. Such labels are called *extractive labels* and usually created heuristically based on reference summaries since existing datasets do not provide extractive labels but only *abstractive* references. Our assumption that each reference sentence corresponds to one facet is similar to that during the creation of extractive labels. The major differences are that (1) We allow an arbitrary number of support sentences while extractive labels usually limit to one support sentence for each reference sentence, *i.e.*, we do not specify M_j. For example, we would put *two* support sentences to *one* support group if they are complementary and only combining them can cover the facet. (2) We try to find multiple support groups ($N > 1$), as there could be more than one set of support sentences that cover the same facet. In contrast, there is no notion of support group in extractive labels as they inherently form one such group ($N = 1$). Also, we allow $N = 0$ if such a mapping cannot be found even by humans. (3) The FAMs are more accurate as they are created by human annotators while extractive methods use sentence regression approaches (which we evaluate in Sec. 4.1) to obtain extractive labels approximately.

Comparison w. SCUs. Some may mistake FAMs for Summarization Content Units (SCUs) in Pyramid (Nenkova and Passonneau, 2004), but they are different in that (1) FAMs utilize both the documents and reference summaries while SCUs ignore the documents; (2) FAMs are at the sentence level and can thus be used to *automatically* evaluate extractive methods once created — simply by matching sentence indices we can know how many facets are covered, while SCUs have to be *manually* annotated for each system (refer to Appendix B Fig. 4).

[3] We ignore coreference (*e.g.*, "he" vs. "the writer") and short fragments when considering the semantics of one facet, as we found that the wording of the reference summaries regarding such choices is also capricious.

2.2 Creation of Extractive CNN/Daily Mail

To verify the effectiveness of facet-aware evaluation, we annotate the FAMs of 150 document-summary pairs from the test set of CNN/Daily Mail. Specifically, we take the first 50 samples in the test set, the 20 samples used in the human evaluation of Narayan et al. (2018c), and randomly draw another 80 samples. The annotators are graduate students who are required to read through the document and mark support groups for each facet. The most similar document sentences to each facet found by ROUGE and cosine similarity of average word embeddings are provided as the baselines for annotation. 310 non-empty FAMs are created by three annotators with high agreement (pairwise Jaccard index 0.714) and further verified to reach consensus.[4] On average, 5.44 (6.04 non-unique) document sentences are included as the support sentences in each document-summary pair.

To summarize, we found that the facets can be divided into three categories based on their quality and degree of abstraction as follows.

Noise: The facet is noisy and irrelevant to the main content, either because the document itself is too hard to summarize (*e.g.*, a report full of quotations) or the human editor was too subjective when writing the summary (See et al., 2017). Another possible reason is that the so-called "summaries" in CNN/Daily Mail are in fact "story highlights", which seems reasonable to include certain details. We found that 41/150 (27.3%) samples have noisy facet(s), indicating that the reference summaries of CNN/Daily Mail are rather noisy. We show in Sec. 3.2 that existing summarization methods perform poorly on this category, which justifies our judgment of "noisy facets" from another aspect. Also note that there would not be a "noise" category in a "clean" dataset. However, given the creation process of popular summarization datasets (Nallapati et al., 2016; Narayan et al., 2018b), it is unlikely that all of their samples are of high quality.

Low Abstraction: The facet can be mapped to its support sentences. We denote the (rounded) average number of support sentences for each facet as $\overline{M}$ ($= \frac{1}{N}\sum_{j=1}^{N} M_j$, N represents the number of support groups). As shown in Table 2, all the facets with non-empty FAMs in CNN/Daily Mail are paraphrases or compression of one to two sentences in the document without much abstraction.

High Abstraction: The facet *cannot* be mapped to its support sentences ($N = 0$) by humans, which indicates that the writing of the facet requires deep understanding of the document rather than simply reorganizing several sentences. The proportion of this category (13.3%) also indicates how often extractive methods would not work (well) on CNN/Daily Mail.

We found it easier than previously believed to create the FAMs on CNN/Daily Mail, as it is uncommon (average number of support groups $\overline{N} = 1.6$) to detect multiple sentences with similar semantics. In addition, most support groups only have one or two support sentences with large lexical overlap, which coincides with the fact that extractive methods work quite well on CNN/Daily Mail and abstractive methods are often hybrid and learn to copy words directly from the documents. That said, we try to automate the FAM creation and scale facet-aware evaluation to the whole test set of CNN/Daily Mail using machine-created FAMs (Sec. 4).

3 Facet-Aware Evaluation

In this section, we introduce the facet-aware evaluation setup (Sec. 3.1) and demonstrate its effectiveness by revisiting state-of-the-art summarization methods under this new setup (Sec. 3.2). We then illustrate the additional benefits of facet-aware evaluation, including fine-grained evaluation (Sec. 3.3) and comparative analysis (Sec. 3.4).

3.1 Proposed Metrics

As current extractive methods are facet-agnostic, *i.e.*, their output is not nested (organized by facets) but a flat set of extracted sentences, we consider one facet as being "covered" if any of its support groups can be found in the whole extracted summary. Formally, we define the Facet-Aware Recall (FAR) as follows.

$$\text{FAR} = \frac{\sum_{i=1}^{R} \mathbf{Any}(\mathbf{I}(\mathcal{S}_1^i, \mathcal{E}), ..., \mathbf{I}(\mathcal{S}_N^i, \mathcal{E}))}{R},$$

where $\mathbf{Any}(\mathcal{X})$ returns 1 if any $x \in \mathcal{X}$ is 1 and 0 otherwise, $\mathbf{I}(\mathcal{X}, \mathcal{Y})$ returns 1 if set $\mathcal{X} \subset \mathcal{Y}$ and 0 otherwise, $\mathcal{E}$ denotes the set of extracted sentences, and R is the number of facets. Intuitively, FAR does not over-penalize extractive methods for extracting long sentences as long as the extracted sentences cover the semantics of the facets. FAR

[4] One alternative way is to store multiple FAMs for each sample (like multiple reference summaries) and average their results as in ROUGE.

also treats each facet equally, whereas ROUGE weighs higher the facets with more tokens since they are more likely to incur lexical overlap.

To further measure model capability of retrieving salient (support) sentences without considering redundancy as FAR does, we merge all the support sentences of one document-summary pair to one single support set and define the Support-Aware Recall (SAR) as follows. SAR is used in Sec. 3.4 for the comparative analysis of extractive methods.

$$\text{SAR} = \frac{|\cup_{i=1}^{R}\cup_{j=1}^{N}\mathcal{S}_j^i \cap \mathcal{E}|}{|\cup_{i=1}^{R}\cup_{j=1}^{N}\mathcal{S}_j^i|}.$$

Example (Fig. 1). Assume that $R = 2$, $r_1 \rightarrow \{\{d_1\},\{d_3\},\{d_4\}\}, r_2 \rightarrow \{\{d_2, d_4\}\}$, and $\mathcal{E} = \{d_1, d_2, d_3\}$. Then FAR $= \frac{1}{2}$ as $\mathcal{E}$ covers $\{d_1\}$ (or $\{d_3\}$) for r_1 but cannot cover $\{d_2, d_4\}$ for r_2. SAR $= \frac{|\{d_1,d_2,d_3,d_4\}\cap\{d_1,d_2,d_3\}|}{|\{d_1,d_2,d_3,d_4\}|} = \frac{3}{4}$. Note that d_1 and d_3 are salient (support sentences) and both considered positive in SAR, while they only contribute to the coverage of one facet in FAR.

3.2 Automatic Evaluation with FAR

By utilizing the low abstraction category on the extractive CNN/Daily Mail dataset, we revisit extractive methods to evaluate how they perform on information coverage. Specifically, we compare Lead-3 (that extracts the first three document sentences), FastRL(E) (E for extractive only) (Chen and Bansal, 2018), BanditSum (Dong et al., 2018), NeuSum (Zhou et al., 2018), Refresh (Narayan et al., 2018c), and UnifiedSum(E) (Hsu et al., 2018) using both ROUGE and FAR. For a fair comparison, each method extracts three sentences ($|\mathcal{E}| = 3$).[5]

Results on Neural Extractive Methods. As shown in Table 3, there is almost no discrimination among the last four methods under ROUGE-1 F1, and the rankings under ROUGE-1/2/L often contradict with each other. The observations on ROUGE Precision/Recall are similar. We provide them as well as more comparative analysis under facet-aware evaluation in Sec. 3.4. For facet coverage, the upper bound of FAR by extracting 3 sentences (Oracle, given the ground-truth FAMs) is 84.8, much higher than all the compared methods. The best performing extractive method under FAR is UnifiedSum(E), which indicates that it covers the most facets semantically.

Method	ROUGE-1	ROUGE-2	ROUGE-L	FAR
Lead-3	41.9	19.6	34.8	50.6
FastRL(E)	41.6	20.3	35.5	50.8
BanditSum	42.7	20.2	35.8	44.7
NeuSum	42.7	**22.1**	36.4	51.2
Refresh	42.8	20.3	**39.3**	51.3
UnifiedSum(E)	42.6	20.7	35.5	**54.8**
Oracle	53.8	32.1	48.1	84.8

Table 3: Performance comparison of extractive methods under ROUGE F1 and Facet-Aware Recall (FAR).

FAR's Correlation w. Human Evaluation. Although FAR is supposed to be favored as the FAMs are manually labeled and indicate accurately whether one sentence should be extracted (assuming the annotations are in high quality), to further verify that FAR correlates with human preference, we ask the annotators to rank the outputs of UnifiedSum(E), NeuSum, and Lead-3 and measure ranking correlation. As listed in Table 4, we observe that the method with the most 1st ranks in the human evaluation coincides with FAR. We also find that FAR has higher Spearman's coefficient ρ than ROUGE (0.457 vs. 0.44).[6]

Method	1st	2nd	3rd
Lead-3	26.8%	46.3%	26.8%
NeuSum	29.3%	39.0%	31.7%
UnifiedSum(E)	**37.8%**	**52.4%**	9.8%

Table 4: **Proportions of system ranking in human evaluation**. FAR shows better human correlation than ROUGE and prefers UnifiedSum(E).

3.3 Fine-grained Evaluation

One benefit of facet-aware evaluation is that we can employ the *category breakdown* of FAMs for fine-grained evaluation, namely, how one method performs on noisy / low abstraction / high abstraction samples, respectively. Any metric of interest can be used for this fine-grained analysis. Here we consider ROUGE and additionally evaluate several abstractive methods: PG (Pointer-Generator) (See et al., 2017), FastRL(E+A) (extractive+abstractive) (Chen and Bansal, 2018), and UnifiedSum(E+A) (Hsu et al., 2018).

As shown in Table 5, extractive methods perform poorly on high abstraction samples, which

[5] Extracting all the sentences results in a perfect FAR, which is expected as FAR measures recall. One can also normalize FAR by the number of extracted sentences.

[6] We expect that one can observe larger gains on datasets with less lexical overlap than CNN/Daily Mail.

is somewhat expected since they cannot perform abstraction. Abstractive methods, however, also exhibit a huge performance gap between low and high abstraction samples, which suggests that existing abstractive methods achieve decent overall performance mainly by extraction rather than abstraction, *i.e.*, performing well on low abstraction samples of CNN/Daily Mail. We also found that all the compared methods perform much worse on the documents with "noisy" reference summaries, implying that the randomness in the reference summaries might introduce noise to both model training and evaluation. Note that although the sample size is relatively small, we observe consistent results when analyzing different subsets of the data.

	Method	N	L	H	L + H
Extractive	Lead-3	34.1	41.9	24.9	38.9
	FastRL(E)	33.5	41.6	31.2	39.8
	BanditSum	35.3	42.7	**34.1**	**41.2**
	NeuSum	34.9	42.7	30.7	40.6
	Refresh	**35.7**	42.8	32.2	40.9
	UnifiedSum(E)	34.2	42.6	31.3	40.6
Abstractive	PG	32.6	40.6	27.5	38.2
	FastRL(E+A)	35.1	40.8	29.9	38.8
	UnifiedSum(E+A)	34.2	42.4	29.2	40.1

Table 5: ROUGE-1 F1 of extractive and abstractive methods on noisy (N), low abstraction (L), high abstraction (H), and high quality (L + H) samples.

3.4 Comparative Analysis

Facet-aware evaluation is also beneficial for comparing extractive methods regarding their capability of extracting salient and non-redundant sentences. We show the FAR, SAR, and ROUGE scores of various extractive methods in Fig. 2. We next illustrate how one can leverage these scores under different metrics for comparative analysis. For brevity, we denote ROUGE Precision and ROUGE Recall as **RP** and **RR**, respectively.

FAR vs. ROUGE. By comparing the scores of extractive methods under FAR and ROUGE, one can discover useful insights. For example, we observe that the performance of Refresh, FastRL(E), NeuSum are quite close to Lead-3 under FAR, but they generally have higher RR. Such results imply that these methods might have learned to extract sentences that are not the support sentences, *i.e.*, sentences that do not directly contribute to the facet coverage, but still have lexical overlap with reference summaries. It is also likely that they extract redundant support sentences that happen to have token matches with other facets. Overall, UnifiedSum(E) covers the most facets (high FAR) and also has decent lexical matches (high RR).

SAR vs. ROUGE. By comparing SAR with RP, one can find that UnifiedSum(E) extracts salient but possibly redundant support sentences, as it has higher SAR but similar RP to Lead-3. On the contrary, Refresh has similar SAR with Lead-3 but higher RP, which again implies that it might extract non-support sentences that contain token matches but irrelevant semantics. Similarly, BanditSum is capable of lexical overlap (high RP), but the matched tokens may not contribute much to the major semantics (low SAR).

FAR vs. SAR. By comparing FAR with SAR (Fig. 3), we observe that FastRL(E) and NeuSum have FAR scores similar to Lead-3 and Refresh, but higher SAR scores. One possible explanation is that FastRL(E) and NeuSum are better at extracting support sentences, but they do not handle redundancy very well, *i.e.*, the extracted sentences might contain multiple support groups of the same facet (recall the example in Sec. 3.1). For instance, there are 30.3% extracted summaries of FastRL(E) that can cover more than one support group of the same facet while there are 19.1% for Lead-3.

4 Evaluation without Human Annotation

In the previous sections, we have demonstrated the effectiveness and benefits of facet-aware evaluation. One remaining issue that might prevent facet-aware evaluation from scaling is the need of human-annotated FAMs. We thus study the feasibility of automatic FAM creation with sentence regression and present a pilot study of conducting facet-aware evaluation *without any human annotation* in this section.

4.1 Sentence Regression for FAM Creation

Similar to most benchmark constructions, facet-aware evaluation requires one-time annotation — once the FAMs are annotated, we can reuse them for automatic evaluation. That said, we explore various approaches to automate this one-time process. Specifically, we investigate whether facet-aware evaluation can be conducted without any human effort by utilizing *sentence regression* (Zopf et al., 2018) to automatically create the FAMs.

Sentence regression is widely used to create extractive labels. Sentence regression approaches typ-

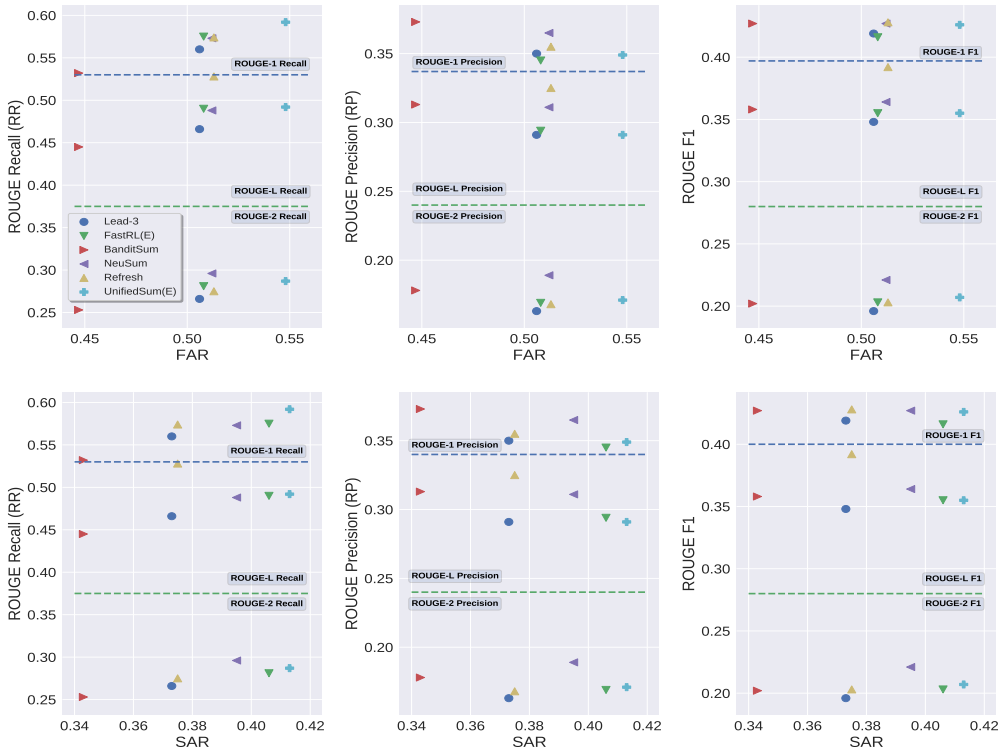

Figure 2: **Performance of extractive methods under ROUGE, FAR, and SAR**. The results under ROUGE-1/2/L often disagree with each other. UnifiedSum(E) generally performs the best in the facet-aware evaluation.

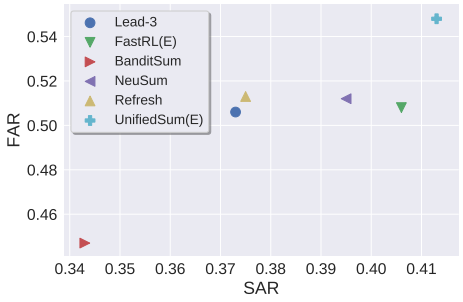

Figure 3: Comparison of extractive methods under FAR and SAR reflects their capability of extracting salient and non-redundant sentences.

ically transform abstractive reference summaries to extractive labels heuristically using ROUGE. Previously, one could only estimate the quality of these labels by evaluating the extractive models trained using such labels, *i.e.*, comparing their extracted summaries with the reference summaries (also approximately via ROUGE). Now that the human-annotated FAMs serve as ground-truth extractive labels, we can evaluate how well each approach performs accurately.

Sentence Regression Approaches. We briefly review recent sentence regression approaches as follows. Nallapati et al. (2017) greedily select sentences that maximize ROUGE-1 F1 until adding another sentence decreases it. Chen and Bansal (2018) find for each reference sentence the most similar sentence in the document by ROUGE-L recall. Zopf et al. (2018) argue that precision is a better measure than recall because it aims not at covering as much information but at wasting as little space as possible. Narayan et al. (2018c) measure sentence similarity by the average of ROUGE-1/2/L F1. We also test other variants of ROUGE and TF-IDF, which represents sentences by TF-IDF features and measures their cosine similarity.

4.2 Evaluation with Machine-Created FAMs

Results on Support Sentence Discovery. We first evaluate sentence regression with its original function, *i.e.*, creating extractive labels (finding support sentences). We merge the support groups of each sample and calculate precision and recall (*i.e.*, SAR). The performance of sentence regression approaches is shown in Table 6. The relatively low recall suggests that simply finding one support sentence for each facet as most existing approaches do would miss plenty of salient sentences, which could possibly worsen the models trained on such labels since the models would treat missed support sentences as unimportant ones. On the bright side, many sentence regression approaches achieve high

precision. For instance, 90.0% document sentences labeled positive by Narayan et al. (2018c) indeed contain salient information. This is to some extent explainable as ROUGE captures lexical overlap and as we have shown, there are many copy-and-paste reference summaries in CNN/Daily Mail.

Method	Precision	Recall	F1
Lead-3	61.0	33.7	43.4
Greedy ROUGE-1 F1	58.2	30.8	40.3
TF-IDF	83.7	51.9	64.0
ROUGE-1 F1	88.9	53.1	66.5
ROUGE-2 F1	86.6	52.3	65.2
ROUGE-L Recall	89.3	53.7	67.1
ROUGE-L Precision	77.2	45.5	57.2
ROUGE-L F1	87.8	53.5	66.5
ROUGE-AVG F1	**90.0**	**53.9**	**67.4**

Table 6: **Performance of sentence regression approaches regarding support sentence discovery**. High precision and low recall are often observed.

Correlation w. Human-Annotated FAMs. We then explore the correlation between human-annotated and machine-created FAMs by evaluating extractive methods against both of them. This time we extend to find for each facet multiple support sentences and put each support sentence into a separate support group. We measure the correlation between estimated and ground-truth FAR by Pearson's r. We measure the correlation between system rankings induced from estimated and ground-truth FAR by Spearman's ρ and Kendall's τ. The detailed correlation results of representative approaches are listed in Table 7. We observe that creating three support groups consistently shows the highest correlation for the same sentence regression approach. Also, the FAMs created by ROUGE-1 F1 and ROUGE-AVG F1 have very high correlation with human annotation, indicating the usability and reliability of machine-created FAMs for system ranking.

Method	$N=1$			$N=2$			$N=3$		
	r	ρ	τ	r	ρ	τ	r	ρ	τ
ROUGE-1 F1	70.5	37.1	33.3	72.0	71.4	60.0	**88.4**	**94.3**	**86.7**
ROUGE-2 F1	11.0	25.7	20.0	43.4	65.7	46.7	88.4	65.7	60.0
ROUGE-L F1	34.0	54.3	46.7	37.5	42.9	20.0	62.3	42.9	46.7
ROUGE-AVG F1	49.6	54.3	46.7	46.1	65.7	46.7	83.2	82.9	73.3

Table 7: Correlation between ground-truth and estimated FAR scores by Pearson's r, Spearman's ρ, and Kendall's τ. N denotes the number of support groups.

FAR Prediction. Despite the high correlation, we also find that the estimated FAR scores may vary in range compared to the ground-truth FAR.[7] Therefore, we further use the estimations of different sentence regression approaches to train a linear regression model to fit the ground-truth FAR (denoted as AutoFAR). We then calculate the estimated FAR scores on the whole test set of CNN/Daily Mail and use the trained linear regressor to predict a (supposedly) more accurate FAR score (denoted as AutoFAR-L). As shown in Table 8, the fitting of AutoFAR is very close to the ground-truth FAR, and the system ranking on the large-scale evaluation under AutoFAR-L follows a similar trend to that under FAR with Spearman's $\rho = 54.3$. On the other hand, although our preliminary analysis on AutoFAR-L shows promising results, we also note that since the human annotation on the whole test set is lacking, the reliability of such extrapolation is not guaranteed and we leave more rigorous study with a larger number of systems and samples as future work.

Method	FAR	AutoFAR	AutoFAR-L	FAR vs. AutoFAR(-L)
BanditSum	44.7	44.8	44.7	Pearson's r
Lead-3	50.6	51.3	45.6	**97.6** (42.9)
FastRL(E)	50.8	51.0	43.1	Spearman's ρ
NeuSum	51.2	49.9	44.3	77.1 (54.3)
Refresh	51.3	51.7	46.2	Kendall's τ
UnifiedSum(E)	**54.8**	**54.5**	**46.9**	60.0 (46.7)

Table 8: **FAR prediction via linear regression**. AutoFAR(-L) denotes the results on the human-annotated subset (entire CNN/Daily Mail dataset).

5 Related Work

Evaluation Metrics for Text Summarization. ROUGE (Lin, 2004) is the most widely used evaluation metric for text summarization. Extensions of ROUGE include ROUGE-WE (Ng and Abrecht, 2015) that incorporated word embedding into ROUGE, ROUGE 2.0 (Ganesan, 2018) that considered synonyms, and ROUGE-G (ShafieiBavani et al., 2018) that applied graph analysis to WordNet for lexical and semantic matching. Nevertheless, these extensions did not draw enough attention as the original ROUGE and recent advances (Gu et al., 2020; Zhang et al., 2019a) are still primarily evaluated by the vanilla ROUGE.

Another popular branch is Pyramid-based metrics (Nenkova and Passonneau, 2004; Yang et al., 2016), which annotate and compare the Summarization Content Units (SCUs) in the summaries.

[7] The raw estimated FAR scores are provided in Appendix B Fig. 5 in the interest of space.

FAR is related to Pyramid and HighRES (Hardy et al., 2019) in that Pyramid employs the summaries to annotate SCUs and HighRES highlights salient text fragments in the documents, while FAR considers both the summaries and documents.

Beyond lexical overlap, embedding-based evaluation metrics (Zhang et al., 2019b; Zhao et al., 2019; Sun and Nenkova, 2019; Xenouleas et al., 2019) are gaining more traction along with the dominance of pre-trained language models. One straightforward way to incorporate embedding-based metrics into FAR is to use them as similarity measures instead of the ROUGE-based approaches tested in Sec. 4.1 for automatic FAM creation (*i.e.*, finding support sentences for each facet by the scores of embedding-based metrics). Such similarity measures are especially beneficial when the facet and its support sentences are not similar at the lexical level.

Reflections on Text Summarization. There has been increasing attention and critique to the issues of existing summarization metrics (Schluter, 2017), methods (Kedzie et al., 2018; Shapira et al., 2018), and datasets (Jung et al., 2019). Notably, Kryscinski et al. (2019) conducted a comprehensive critical evaluation for summarization from various aspects. Zopf et al. (2018) investigated sentence regression approaches in a manner similar to ours but they could only evaluate them approximately against ROUGE as no ground-truth labels (FAMs) existed.

Annotation and Analysis. Many recent studies conduct human annotation or evaluation on text summarization and other NLP tasks to gain useful insights. Hardy et al. (2019) annotated 50 documents to demonstrate the benefits of highlight-based summarization evaluation. Recent summarization methods (Paulus et al., 2017; Narayan et al., 2018c; Chen and Bansal, 2018) generally sampled 50 to 100 documents for human evaluation in addition to ROUGE in light of its limitations. Chen et al. (2016); Yavuz et al. (2018) inspected 100 samples and analyzed their category breakdown for reading comprehension and semantic parsing, respectively. We observed similar trends when analyzing different subsets of the FAMs, indicating that our findings are relatively stable. We thus conjecture that our sample size is sufficient to verify our hypotheses and benefit future research.

6 Conclusion and Future Work

We propose a facet-aware evaluation setup for better assessment of information coverage in extractive summarization. We construct an extractive summarization dataset and demonstrate the effectiveness of facet-aware evaluation on this newly constructed dataset, including better human correlation on the assessment of information coverage, and the support for fine-grained evaluation as well as comparative analysis. We also evaluate sentence regression approaches and explore the feasibility of fully-automatic evaluation without any human annotation. In the future, we will investigate multi-document summarization datasets such as DUC (Paul and James, 2004) and TAC (Dang and Owczarzak, 2008) to see whether our findings coincide when multiple references are provided. We will also explore better sentence regression approaches for the use of both extractive summarization methods and automatic FAM creation.

Acknowledgement

We thank Woojeong Jin and Jiaming Shen for the valuable feedback on the paper draft. We thank anonymous reviewers for the constructive comments. Research was sponsored in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and SocialSim Program No. W911NF-17-C-0099, National Science Foundation IIS 16-18481, IIS 17-04532, and IIS-17-41317, and DTRA HDTRA11810026. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and should not be interpreted as necessarily representing the views, either expressed or implied, of DARPA or the U.S. Government.

References

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the CNN/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Hoa Trang Dang and Karolina Owczarzak. 2008. Overview of the tac 2008 update summarization task. In *TAC*.

Yue Dong, Yikang Shen, Eric Crawford, Herke van Hoof, and Jackie Chi Kit Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3739–3748, Brussels, Belgium. Association for Computational Linguistics.

Kavita Ganesan. 2018. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

Xiaotao Gu, Yuning Mao, Jiawei Han, Jialu Liu, Hongkun Yu, You Wu, Cong Yu, Daniel Finnie, Jiaqi Zhai, and Nicholas Zukoski. 2020. Generating representative headlines for news stories. *WWW*.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy. Association for Computational Linguistics.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 132–141, Melbourne, Australia. Association for Computational Linguistics.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in 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 3315–3326, Hong Kong, China. Association for Computational Linguistics.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

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

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*, pages 3075–3081.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Ronald Cardenas, Nikos Papasarantopoulos, Shay B. Cohen, Mirella Lapata, Jiangsheng Yu, and Yi Chang. 2018a. Document modeling with external attention for sentence extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2020–2030, Melbourne, Australia. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018c. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. 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*, pages 1925–1930, Lisbon, Portugal. Association for Computational Linguistics.

Over Paul and Yen James. 2004. An introduction to duc-2004. In *Proceedings of the 4th Document Understanding Conference (DUC 2004)*.

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

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond 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*, pages 762–767, Brussels, Belgium. Association for Computational Linguistics.

Ori Shapira, David Gabay, Hadar Ronen, Judit Bar-Ilan, Yael Amsterdamer, Ani Nenkova, and Ido Dagan. 2018. Evaluating multiple system summary lengths: A case study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 774–778, Brussels, Belgium. 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, Hong Kong, China. 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 the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6004–6010, Hong Kong, China. Association for Computational Linguistics.

Qian Yang, Rebecca J Passonneau, and Gerard De Melo. 2016. Peak: Pyramid evaluation via automated knowledge extraction. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Semih Yavuz, Izzeddin Gur, Yu Su, and Xifeng Yan. 2018. What it takes to achieve 100% condition accuracy on WikiSQL. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1702–1711, Brussels, Belgium. Association for Computational Linguistics.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.

Markus Zopf, Eneldo Loza Mencía, and Johannes Fürnkranz. 2018. Which scores to predict in sentence regression for text summarization? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1782–1791, New Orleans, Louisiana. Association for Computational Linguistics.

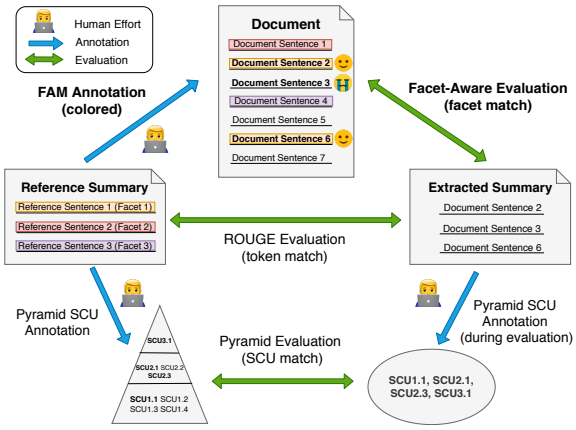

Figure 4: **Comparison of summarization metrics**. Support sentences are marked in the same color as their corresponding facets. SCUs have to be annotated for each extracted summary during evaluation, while facet-aware evaluation can be conducted automatically by comparing sentence indices.

A Practical Notes on CNN/Daily Mail

We note several issues of the CNN/Daily Mail dataset in the hope that the researchers working on this dataset are better aware of these issues.

One issue is that sometimes the titles and image captions are introduced in the main body of the document by mistake (usually captured by "-lrb-pictured -rrb-" or colons), which may lead to bias or label leaking for model training since the reference summaries are observed to be similar to the titles and image captions (Narayan et al., 2018a). For example, we found that if there is a sentence in the main body that is almost the same as one of the captions, then that sentence is very likely to be used in the reference summary. Many such cases can be found in our annotated data.

We also found that in many documents, the 4-th sentence is "*scroll down for video*". And if this sentence appears in one document, it is often the case that the first three sentences are good enough to summarize the whole document. This finding provides yet another evidence why a simple Lead-3 baseline could be rather strong on CNN/Daily Mail. In addition, sentences similar to the first three sentences can often be found afterward, which suggests that the first three sentences may not even belong to the main body of the document.

B Additional Illustration

In Fig. 4, we show the comparison of ROUGE, FAR, and Pyramid. In Fig. 5, we show the the ground-truth FAR scores, the FAR scores estimated by various sentence regression approaches, and the prediction of FAR scores by linear regression.

C Detailed Examples

We list below the full documents, reference summaries, and the corresponding FAMs of several examples shown in Table 2. In particular, Table 10 shows an example of several support groups covering the same facet. We release all of the annotated data to facilitate facet-aware evaluation and follow-up studies along this direction.

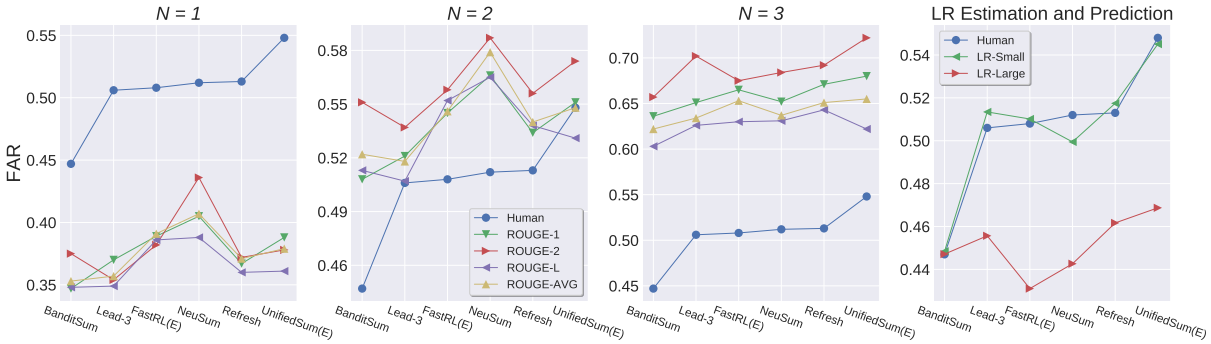

Figure 5: The first three figures show the ground-truth and estimated FAR scores via human-annotated FAMs and machine-created FAMs. The fourth figure shows the fitting of linear regression on the human-annotated samples (LR-Small) and the prediction on the whole test set of CNN/Daily Mail (LR-Large). Systems are sorted in an ascending order by the ground-truth FAR on the human-annotated samples.

ID: 1b2cc634e2bfc6f2595260e7ed9b42f77ecbb0ce
Category: Noise

Document:
-LRB- CNN -RRB- Paul Walker is hardly the first actor to die during a production .
But Walker 's death in November 2013 at the age of 40 after a car crash was especially eerie given his rise to fame in the " Fast and Furious " film franchise . The release of **" Furious 7 " on Friday** (**this is the only mention of "Friday" in the whole document**) offers the opportunity for fans to remember – and possibly grieve again – the man that so many have praised as one of the nicest guys in Hollywood .
" He was a person of humility , integrity , and compassion , " military veteran Kyle Upham said in an email to CNN . Walker secretly paid for the engagement ring Upham shopped for with his bride .
" We did n't know him personally but this was apparent in the short time we spent with him . I know that we will never forget him and he will always be someone very special to us , " said Upham .
The actor was on break from filming " Furious 7 " at the time of the fiery accident , which also claimed the life of the car 's driver , Roger Rodas . Producers said early on that they would not kill off Walker 's character , Brian O'Connor , a former cop turned road racer .
Instead , the script was rewritten and special effects were used to finish scenes , with Walker 's brothers , Cody and Caleb , serving as body doubles .
There are scenes that will resonate with the audience – including the ending , in which the filmmakers figured out a touching way to pay tribute to Walker while " retiring " his character .
At the premiere Wednesday night in Hollywood , Walker 's co-star and close friend **Vin Diesel gave a tearful speech before the screening , saying " This movie is more than a movie . "** (**random quotation, may use other quotes as well**)
" You 'll feel it when you see it , " Diesel said . " There 's something emotional that happens to you , where you walk out of this movie and you appreciate everyone you love because you just never know when the last day is you 're gon na see them . "
There have been multiple tributes to Walker leading up to the release . Diesel revealed in an interview with the " Today " show that he had named his newborn daughter after Walker . Social media has also been paying homage to the late actor .
A week after Walker 's death , about 5,000 people attended an outdoor memorial to him in Los Angeles . Most had never met him .
Marcus Coleman told CNN he spent almost $ 1,000 to truck in a banner from Bakersfield for people to sign at the memorial .
" It 's like losing a friend or a really close family member ... even though he is an actor and we never really met face to face , " Coleman said . " Sitting there , bringing his movies into your house or watching on TV , it 's like getting to know somebody . It really , really hurts . "
Walker 's younger brother Cody told People magazine that he was initially nervous about how " Furious 7 " would turn out , but he is happy with the film . " It 's bittersweet , but I think Paul would be proud , " he said .
CNN 's Paul Vercammen contributed to this report .

Reference Summary:
" Furious 7 " pays tribute to star Paul Walker , who died during filming
Vin Diesel : " This movie is more than a movie " (**random quotation**)
" Furious 7 " opens Friday (**unimportant detail**)

FAMs:
N/A

Table 9: Full document, reference summary, and the FAMs presented in Table 2.

ID: d58bf9387cd76f34bbb95fe25f8036015e5cc90a
Category: Low Abstraction

Document:
Dover police say a man they believe to be the so-called ‘ rat burglar ’ who cut holes to tunnel into buildings has been arrested in Maryland .
Authorities said in a news release Thursday that 49-year-old Thomas K. Jenkins of Capitol Heights , Maryland , was arrested last month by deputies with the Prince George ’s County Sheriff ’s Office .
‘ Rat burglar ’ : Thomas K. Jenkins , pictured is accused of robbing 18 Dover businesses
From September 2014 to February 2015 , Jenkins allegedly carried out 18 commercial robberies in Dover , Delaware , authorities there said .
‘ During the investigation it was learned that the Prince George ’s County Sheriff ’s Department had a series of burglaries that were similar in nature to the eighteen committed in Dover , ’ the release said .
Thomas Jenkins has been accused by the Dover Police Department of robbing multiple businesses .
They are :
Maple Dale Country Club
Manlove Auto Parts
Sovereign Properties
Morgan Properties
U and I Builders
AMCO Check Cashing
Colonial Investment
1st Capital Mortgage
Advantage Travel
Ancient Way Massage
Tranquil Spirit Massage/Spa
Christopher Asay Massage
Morgan Communities
Vincenzo ’s Restaurant
Happy Fortune Chinese Restaurant
Happy 13 Liquors
Del-One Credit Union
Pizza Time
Melvin ’s Auto Service
Source : Dover Police Department/The News Journal
A car was found behind a building where a robbery took place and led deputies in Maryland to consider Jenkins as a suspect , authorities said .
Law enforcement later found Jenkin ’s car and tracked where he went , Dover police said .
Police say Jenkins had cut a hole in the roof of a commercial business in Maryland on March 9 and deputies arrested him as he fled .
According to Dover police , ‘ Jenkins was found in possession of .45 - caliber handgun that was stolen from a business in Delaware State Police Troop 9 jurisdiction . A search of Jenkins vehicle revealed an additional .45 - caliber handgun stolen from the same business . ’
Jenkins is being held in Maryland and will face 72 charges involving the 18 burglaries in Dover when he is returned to Delaware .
The charges he is facing break down to : four counts of wearing a disguise during the commission of a felony , eighteen counts of third-degree burglary , eighteen counts of possession of burglary tools , fourteen counts of theft under $ 1,500 , and eighteen counts of criminal mischief , two of which are felonies , authorities said .
Cpl. Mark Hoffman with the Dover Police Department told the News Journal that Delaware State Police are planning to file charges over a 19th robbery at Melvin ’s Auto Service , which reportedly occurred in a part of Dover where jurisdiction is held by state police .
Sharon Hutchison , who works at one of the businesses Jenkins allegedly robbed , told the newspaper ‘ He cut through two layers of drywall , studs and insulation . ’
The Prince George ’s County Sheriff ’s Department did not immediately return a request for information on what charges Jenkins is facing there .

FAMs:

- **thomas k. jenkins , 49 , was arrested last month by deputies with the prince george ’s county sheriff ’s office , authorities said .**

  [Support Group0][Sent0]: authorities said in a news release thursday that 49-year-old thomas k. jenkins of capitol heights , maryland , was arrested last month by deputies with the prince george ’s county sheriff ’s office .

- **police say jenkins had cut a hole in the roof of a commercial business in maryland on march 9 and deputies arrested him as he fled .**

  [Support Group0][Sent0]: police say jenkins had cut a hole in the roof of a commercial business in maryland on march 9 and deputies arrested him as he fled .

- **jenkins is accused of carrying out multiple robberies in dover , delaware .**

  [Support Group0][Sent0]: jenkins is being held in maryland and will face 72 charges involving the 18 burglaries in dover when he is returned to delaware .

  [Support Group1][Sent0]: ‘ rat burglar ’ : thomas k. jenkins , pictured is accused of robbing 18 dover businesses .

  [Support Group2][Sent0]: thomas jenkins has been accused by the dover police department of robbing multiple businesses .

- **he is facing 72 charges from the dover police department for 18 robberies .**

  [Support Group0][Sent0]: jenkins is being held in maryland and will face 72 charges involving the 18 burglaries in dover when he is returned to delaware .

- **the delaware state police is planning to file charges over a 19th robbery , which occurred in a part of dover where jurisdiction is held by state police .**

  [Support Group0][Sent0]: mark hoffman with the dover police department told the news journal that delaware state police are planning to file charges over a 19th robbery at melvin ’s auto service , which reportedly occurred in a part of dover where jurisdiction is held by state police .

Table 10: Full document, reference summary, and the FAMs presented in Table 2.

ID: d1fa0db909ce45fe1ee32d6cbb546e9d784bcf74
Category: Low Abstraction

Document:
-LRB- CNN -RRB- You probably never knew her name , but you were familiar with her work .
Betty Whitehead Willis , the designer of the iconic " Welcome to Fabulous Las Vegas " sign , died over the weekend . She was 91 .
Willis played a major role in creating some of the most memorable neon work in the city .
The Neon Museum also credits her with designing the signs for Moulin Rouge Hotel and Blue Angel Motel
Willis visited the Neon Museum in 2013 to celebrate her 90th birthday .
Born about 50 miles outside of Las Vegas in Overton , she attended art school in Pasadena , California , before returning home .
She retired at age 77 .
Willis never trademarked her most-famous work , calling it " my gift to the city . "
Today it can be found on everything from T-shirts to refrigerator magnets .
People we 've lost in 2015

FAMs:

- **willis never trademarked her most-famous work , calling it " my gift to the city "**

  [Support Group0][Sent0]: willis never trademarked her most-famous work , calling it " my gift to the city . "

- **she created some of the city 's most famous neon work .**

  [Support Group0][Sent0]: willis played a major role in creating some of the most memorable neon work in the city .

Table 11: Full document, reference summary, and the FAMs presented in Table 2.

ID: dc833f8b55e381011ce23f89ea909b9a141b5a66
Category: High Abstraction

Document:
-LRB- CNN -RRB- As goes Walmart , so goes the nation ?
Everyone from Apple CEO Tim Cook to the head of the NCAA slammed religious freedom laws being considered in several states this week , warning that they would open the door to discrimination against gay and lesbian customers .
But it was the opposition from Walmart , the ubiquitous retailer that dots the American landscape , that perhaps resonated most deeply , providing the latest evidence of growing support for gay rights in the heartland .
Walmart 's staunch criticism of a religious freedom law in its home state of Arkansas came after the company said in February it would boost pay for about 500,000 workers well above the federal minimum wage . Taken together , the company is emerging as a bellwether for shifting public opinion on hot-button political issues that divide conservatives and liberals .
And some prominent Republicans are urging the party to take notice .
Former Minnesota Gov. Tim Pawlenty , who famously called on the GOP to " be the party of Sam 's Club , not just the country club , " told CNN that Walmart 's actions " foreshadow where the Republican Party will need to move . "
" The Republican Party will have to better stand for " ideas on helping the middle class , said Pawlenty , the head of the Financial Services Roundtable , a Washington lobbying group for the finance industry . The party 's leaders must be " willing to put forward ideas that will help modest income workers , such as a reasonable increase in the minimum wage , and prohibit discrimination in things such as jobs , housing , public accommodation against gays and lesbians . "
Walmart , which employs more than 50,000 people in Arkansas , emerged victorious on Wednesday . Hours after the company 's CEO , Doug McMillon , called on Republican Gov. Asa Hutchinson to veto the bill , the governor held a news conference and announced he would not sign the legislation unless its language was fixed .
Walmart 's opposition to the religious freedom law once again puts the company at odds with many in the Republican Party , which the company 's political action committee has tended to support .
In 2004 , the Walmart PAC gave around $ 2 million to Republicans versus less than $ 500,000 to Democrats , according to data from the Center for Responsive Politics . That gap has grown less pronounced in recent years . In 2014 , the PAC spent about $ 1.3 million to support Republicans and around $ 970,000 for Democrats .
It has been a gradual transformation for Walmart .
In 2011 , the company bulked up its nondiscrimination policies by adding protections for gender identity . Two years later , the company announced that it would start offering health insurance benefits to same-sex partners of employees starting in 2014 .
Retail experts say Walmart 's evolution on these issues over the years is partly a reflection of its diverse consumer base , as well as a recognition of the country 's increasingly progressive views of gay equality -LRB- support for same-sex marriage is at a new high of 59 % , according to a recent Wall Street Journal/NBC News poll -RRB- .
" It 's easy for someone like a Chick-fil-A to take a really polarizing position , " said Dwight Hill , a partner at the retail consulting firm McMillanDoolittle .
" But in the world of the largest retailer in the world , that 's very different . "
Hill added : Same-sex marriage , " while divisive , it 's becoming more common place here within the U.S. , and the businesses by definition have to follow the trend of their customer . "
The backlash over the religious freedom measures in Indiana and Arkansas this week is shining a bright light on the broader business community 's overwhelming support for workplace policies that promote gay equality .
After Indiana Gov. Mike Pence , a Republican , signed his state 's religious freedom bill into law , CEOs of companies big and small across the country threatened to pull out of the Hoosier state .
The resistance came from business leaders of all political persuasions , including Bill Oesterle , CEO of the business-rating website Angie 's List and a one-time campaign manager for former Indiana Gov. Mitch Daniels . Oesterle announced that his company would put plans on hold to expand its footprint in Indianapolis in light of the state 's passage of the religious freedom act .
NASCAR , scheduled to hold a race in Indianapolis this summer , also spoke out against the Indiana law .
" What we 're seeing over the past week is a tremendous amount of support from the business community who are standing up and are sending that equality is good for business and discrimination is bad for business , " said Jason Rahlan , spokesman for the Human Rights Campaign .
The debate has reached presidential politics .
National Republicans are being forced to walk the fine line of protecting religious liberties and supporting nondiscrimination .
Likely GOP presidential candidate Jeb Bush initially backed Indiana 's religious freedom law and Pence , but moderated his tone a few days later . The former Florida governor said Wednesday that Indiana could have taken a " better " and " more consensus-oriented approach . "
" By the end of the week , Indiana will be in the right place , " Bush said , a reference to Pence 's promise this week to fix his state 's law in light of the widespread backlash .
Others in the GOP field are digging in . Sen. Ted Cruz of Texas , the only officially declared Republican presidential candidate , said Wednesday that he had no interest in second-guessing Pence and lashed out at the business community for opposing the law .
" I think it is unfortunate that large companies today are listening to the extreme left wing agenda that is driven by an aggressive gay marriage agenda , " Cruz said .
Meanwhile , former Secretary of State Hillary Clinton , who previously served on Walmart 's board of directors , called on Hutchinson to veto the Arkansas bill , saying it would " permit unfair discrimination " against the LGBT community .
Jay Chesshir , CEO of the Little Rock Regional Chamber of Commerce in Arkansas , welcomed Hutchinson 's pledge on Wednesday to seek changes to his state 's bill . He said businesses are not afraid to wade into a politically controversial debate to ensure inclusive workplace policies .
" When it comes to culture and quality of life , businesses are extremely interested in engaging in debate simply because it impacts its more precious resource – and that 's its people , " Chesshir said . " Therefore , when issues arise that have negative or positive impact on those things , then the business community will again speak and speak loudly . "

Reference Summary:
While Republican Gov. Asa Hutchinson was weighing an Arkansas religious freedom bill , Walmart voiced its opposition (**highly abstractive, hard to obtain by rephrasing original sentences**)
Walmart and other high-profile businesses are showing their support for gay and lesbian rights
Their stance puts them in conflict with socially conservative Republicans , traditionally seen as allies

FAMs:
N/A

Table 12: Full document, reference summary, and the FAMs presented in Table 2.

ID: 1b2cc634e2bfc6f2595260e7ed9b42f77ecbb0ce
Category: High Abstraction

Document:
-LRB- CNN -RRB- He 's a blue chip college basketball recruit . She 's a high school freshman with Down syndrome .
At first glance Trey Moses and Ellie Meredith could n't be more different . But all that changed Thursday when Trey asked Ellie to be his prom date .
Trey – a star on Eastern High School 's basketball team in Louisville , Kentucky , who 's headed to play college ball next year at Ball State – was originally going to take his girlfriend to Eastern 's prom .
So why is he taking Ellie instead ? " She 's great ... she listens and she 's easy to talk to " he said .
Trey made the prom-posal -LRB- yes , that 's what they are calling invites to prom these days -RRB- in the gym during Ellie 's P.E. class .
Trina Helson , a teacher at Eastern , alerted the school 's newspaper staff to the prom-posal and posted photos of Trey and Ellie on Twitter that have gone viral . She was n't surpristed by Trey 's actions .
" That 's the kind of person Trey is , " she said .
To help make sure she said yes , Trey entered the gym armed with flowers and a poster that read " Let 's Party Like it 's 1989 , " a reference to the latest album by Taylor Swift , Ellie 's favorite singer .
Trey also got the OK from Ellie 's parents the night before via text . They were thrilled .
" You just feel numb to those moments raising a special needs child , " said Darla Meredith , Ellie 's mom . " You first feel the need to protect and then to overprotect . "
Darla Meredith said Ellie has struggled with friendships since elementary school , but a special program at Eastern called Best Buddies had made things easier for her .
She said Best Buddies cultivates friendships between students with and without developmental disabilities and prevents students like Ellie from feeling isolated and left out of social functions .
" I guess around middle school is when kids started to care about what others thought , " she said , but " this school , this year has been a relief . "
Trey 's future coach at Ball State , James Whitford , said he felt great about the prom-posal , noting that Trey , whom he 's known for a long time , often works with other kids
Trey 's mother , Shelly Moses , was also proud of her son .
" It 's exciting to bring awareness to a good cause , " she said . " Trey has worked pretty hard , and he 's a good son . "
Both Trey and Ellie have a lot of planning to do . Trey is looking to take up special education as a college major , in addition to playing basketball in the fall .
As for Ellie , she ca n't stop thinking about prom .
" Ellie ca n't wait to go dress shopping " her mother said .
" Because I 've only told about a million people ! " Ellie interjected .

Reference Summary:
College-bound basketball star asks girl with down syndrome to high school prom. (**highly abstractive, hard to obtain by rephrasing original sentences**)
Pictures of the two during the "prom-posal" have gone viral.

FAMs:
N/A

Table 13: Full document, reference summary, and the FAMs presented in Table 2.