# Facet-Aware Evaluation for Extractive Text 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 text summarization like ROUGE focus on the lexical similarity and are facet-agnostic. In this paper, we present a facet-aware evaluation procedure for better assessment of the information coverage in extracted summaries while still supporting automatic evaluation once annotated. Specifically, we treat *facet* instead of *token* as the basic unit for evaluation, manually annotate the *support sentences* for each facet, and directly evaluate extractive methods by comparing the indices of extracted sentences with support sentences. We demonstrate the benefits of the proposed setup by performing a thorough *quantitative* investigation on the CNN/Daily Mail dataset, which in the meantime reveals useful insights of state-of-the-art summarization methods.[1]

## 1 Introduction

Summarizing one document into a concise paragraph is what generations of researchers have been pursuing. The most common automatic evaluation metric of this task is lexical overlap (*i.e.*, ROUGE (Lin, 2004)), which takes the system and reference summaries as a sequence of tokens and measures their n-gram overlap. However, recent studies observe the limits of ROUGE and find in some cases, it fails to reach consensus with human judgment (Paulus et al., 2017; Schluter, 2017).

Since lexical overlap describes the lexical similarity, ROUGE would favor summaries that share more fragments with the reference, whereas such summaries may not always be the desired ones. For example, in Table 1, the sentence with the highest ROUGE score has more term matches but irrelevant semantics, while the manually extracted sentence is over-penalized as it involves other details. Also, the manual extract sometimes may not even have any lexical overlap with the reference. Instead of lexical overlap, we believe that information coverage in summarization can be better characterized by *facet overlap*—*i.e.*, whether the system summary covers the facets in the reference summary. Compared to treating the summary as a sequence of n-grams, *facet-aware evaluation* considers information coverage at a granularity that is semantically richer, and thus can provide a more accurate assessment on the summary quality.

Table 1: In some cases, lexical overlap (finding the sentence in the document with the highest ROUGE score) could be misleading.

**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 term matches but totally different meanings**)
**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)**

**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**)

In this paper, we propose to treat facet as the basic unit for summary evaluation and assess one method by how many facets it covers. Specifically, we focus on extractive summarization, treat each *reference sentence* as a facet, group *document sentences* that cover the information of each facet (referred to as *support sentences*) and calculate the Facet-Aware Recall (**FAR**) as the measure of information coverage. We take the *CNN/Daily Mail dataset* (Nallapati et al., 2016) as a test bed and manually create Facet-Aware Mappings (**FAMs**) from each facet (sentence) in one reference summary to its support sentences in the document. FAMs can be viewed as extractive labels indi-

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

cating which sentences should be extracted but they are organized around each facet rather than flat label sets and thus contain richer information. While FAMs look similar to Summarization Content Units (SCUs) in Pyramid (Nenkova and Passonneau, 2004), the proposed setup is different in that 1) FAMs are created using both the documents and reference summaries while SCUs only consider the references. 2) FAMs are at sentence level and can thus be used to *automatically* evaluate extractive methods once created—simply by matching the sentence indices we can know how many facets are covered, while in Pyramid systems have to be *manually* evaluated (Fig. 1).

In the following sections, we first revisit state-of-the-art extractive methods with explicit consideration of facet coverage. We then show that FAMs are also beneficial for fine-grained performance analysis of both abstractive and extractive methods. Finally, we evaluate existing approaches that *approximate* extractive labels against FAMs to see how accurate they are.

**Contributions.** 1) We propose a facet-aware evaluation setup that better assesses the information coverage for extractive summarization than ROUGE while still supporting automatic evaluation once created. 2) We perform the initial trial of building *extractive* summarization benchmarks 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 useful insights. 4) To our knowledge, this is also the first thorough *quantitative* analysis regarding the characteristics of the CNN/Daily Mail text summarization task.

## 2 Building *Extractive* CNN/Daily Mail

In this section, we describe the procedure of annotating CNN/Daily Mail. For each facet (sentence) in the reference summary, we find all its support sentences in the document that can cover its meaning. Note that the support sentences are likely to be more verbose, but we only consider if the sentences cover the semantics of the facet regardless of their length. The reason is that we believe extractive summarization should focus on information coverage and once salient sentences are extracted, one can then compress them in an abstractive way (Chen and Bansal, 2018; Hsu et al., 2018). Formally, we denote one document-summary pair as $\{d, r\}$, where $d = \{d^j\}_{j=1}^{D}$, $r = \{r^j\}_{j=1}^{R}$, and $D$, $R$ denote the number of sentences. We define one *support group* of facet $\mathcal{F}$ as a minimum set of sentences in the document that express the meaning of $\mathcal{F}$. For each $r^j$, we annotate a FAM $r^j \rightarrow \{\{d^{s^k_{j,1}}\}_{k=1}^{\mathrm{K}_1}, \{d^{s^k_{j,2}}\}_{k=1}^{\mathrm{K}_2}, ..., \{d^{s^k_{j,N}}\}_{k=1}^{\mathrm{K}_N}\}$ in which each $\{d^{s^k_{j,n}}\}_{k=1}^{\mathrm{K}_n}$ is a support group and $s^k_{j,n}$ is the index of the $k$-th support sentence in group $n$.[2]

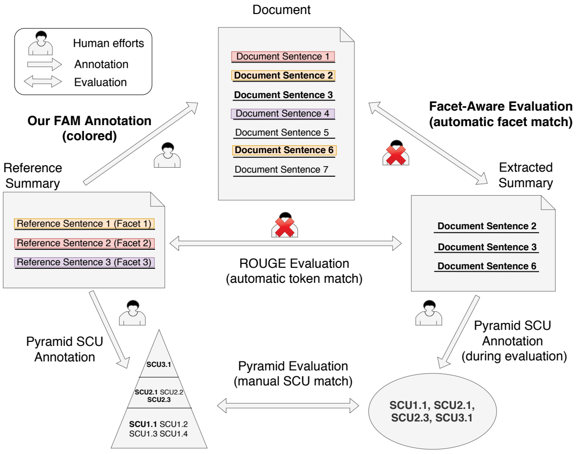
Document
Human efforts
Annotation
Evaluation
Document Sentence 1
Document Sentence 2
Document Sentence 3
Document Sentence 4
Document Sentence 5
Document Sentence 6
Document Sentence 7
Our FAM Annotation (colored)
Facet-Aware Evaluation (automatic facet match)
Reference Summary
Extracted Summary
Reference Sentence 1 (Facet 1)
Reference Sentence 2 (Facet 2)
Reference Sentence 3 (Facet 3)
ROUGE Evaluation (automatic token match)
Document Sentence 2
Document Sentence 3
Document Sentence 6
Pyramid SCU Annotation
Pyramid SCU Annotation (during evaluation)
Pyramid Evaluation (manual SCU match)
SCU3.1
SCU2.1 SCU2.2 SCU2.3
SCU1.1 SCU1.2 SCU1.3 SCU1.4
SCU1.1, SCU2.1, SCU2.3, SCU3.1

Figure 1: Comparison of summarization metrics. Support sentences share the same color as their facets. Facet 1 is covered by the (extracted) document sentences 2 and 6.

One may regard the procedure as creating extractive labels, which is widely used in extractive summarization since only *abstractive* references are available in existing datasets. The major differences are that 1) We label all the support sentences instead of just one or fixed number of sentences, *i.e.*, we do not specify $\mathrm{K}_n$. For example, we would put *two* sentences to *one* support group if they are complementary and only combining them can cover the facet. 2) We find multiple support groups ($N > 1$), as there could be more than one set of sentences that cover the same facet and extracting any one of them is acceptable. In contrast, there is no concept of support group in extractive labels as they inherently form one such group. We sampled 150 document-summary pairs from the test set of CNN/Daily Mail. 344 FAMs were created by three annotators with high agreement (pairwise Jaccard index 0.71) and further verified to reach consensus. We found that the facets can be divided into three categories based on their quality and degree

[2] 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.

Table 2: Category breakdown of the Facet-Aware Mappings (FAMs) from reference summaries to documents.

| Categories | # | % | Examples (full documents, reference summaries, and the FAMs can be found in Appendix D) |
|---|---|---|---|
| Random (**R**) | 41 | 26% | • **Reference**: “Furious 7” opens Friday. (**unimportant detail**)<br>• **Reference**: Click here for all the latest Floyd Mayweather vs Manny Pacquiao news. (**not found in the document**)<br>• **Reference**: Vin Diesel: “This movie is more than a movie”. (**random quotation**)<br>• **Reference**: “I had a small moment of awe,” she said. (**random quotation**) |
| Low Abstraction (**L**) | 344<br>K=1:<br>304<br>K=2:<br>40 | 93%<br>K=1:<br>82%<br>K=2:<br>11% | • **Reference**: Willis never trademarked her most-famous work, calling it “my gift to the city”.<br>• **Support**: Willis never trademarked her most-famous work, calling it “my gift to the city.” (**identical**)<br>• **Reference**: Thomas K. Jenkins, 49, was arrested last month by deputies with the Prince George’s County sheriff’s office, authorities said.<br>• **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**) | 25 | 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.<br>• **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. Their stance puts them in conflict with socially conservative Republicans, traditionally seen as allies. |

of abstraction as follows.

**Random: The facet is quite random**, 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” are in fact “story highlights”, which seems reasonable to contain details. We found that 41/150 (26%) samples have random facet(s), implying there are severe issues in the reference summaries of CNN/Daily Mail.

**Low Abstraction: The facet can be mapped to its support sentences**. We further divide this category by the (rounded) average number of support sentences K of $N$ support groups (K $= \frac{\sum_{n=1}^{N} |\{d^{s^k_{j,n}}\}_{k=1}^{\mathrm{K}_n}|}{N}$). As in Table 2, most facets (93%) in the reference summaries 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**, which indicates that its writing requires deep understandings of the document rather than reorganizing several sentences. The proportion of this category (7%) also indicates how often extractive methods would not work (well) on CNN/Daily Mail.

Surprisingly, we found it easier than previously believed to create the FAMs on CNN/Daily Mail, as it is uncommon ($\overline{N} = 1.56$) to detect multiple sentences with similar semantics (compared to multi-document summarization). In addition, most support groups only have one or two support sentences with large lexical overlap.

## 3 Revisit of State-of-the-art Methods

By utilizing the FAMs, we revisit extractive methods to see how well they perform on facet coverage. Specifically, we compare **Lead-3**, **Refresh** (Narayan et al., 2018b), **FastRL(E)** (E for extractive only) (Chen and Bansal, 2018), **UnifiedSum(E)** (Hsu et al., 2018), **NeuSum** (Zhou et al., 2018), and **BanditSum** (Dong et al., 2018) using both ROUGE and FAMs. As these methods are facet-agnostic (*i.e.*, their outputs are not organized by facets but flat extract sets), we consider one facet is covered as long as *one of its support groups is extracted* and measure the Facet-Aware Recall (**FAR** $= \frac{\#\text{covered}}{R}$). For a fair comparison, each method extracts three sentences since extracting all would result in a perfect FAR.

Table 3: Comparison of extractive methods using ROUGE F1 and Facet-Aware Recall (FAR).

| Method | ROUGE-1 | ROUGE-2 | ROUGE-L | FAR |
|---|---|---|---|---|
| Lead-3 | 41.9 | 19.6 | 34.9 | 47.2 |
| FastRL(E) (Chen and Bansal, 2018) | 41.8 | 20.5 | 35.7 | 48.5 |
| UnifiedSum(E) (Hsu et al., 2018) | 42.7 | 20.8 | 35.6 | **50.4** |
| Refresh (Narayan et al., 2018b) | 42.8 | 20.3 | **39.3** | 49.0 |
| NeuSum (Zhou et al., 2018) | 42.8 | **22.2** | 36.5 | 48.3 |
| BanditSum (Dong et al., 2018) | **42.9** | 20.4 | 35.9 | 42.4 |

As shown in Table 3, there is almost no discrimination among the last four methods under ROUGE-1 F1, and their rankings under ROUGE-1/2/L are quite different.[3] In contrast, FAR shows that UnifiedSum(E) covers the most facets. Although FAR is supposed to be favored as *FAMs are already manually labeled and tell exactly if one sentence should be extracted* (assuming our annotations are in agreement), to further verify that

[3] The results on ROUGE Precision and Recall are similar. We provide them as well as more method comparison using facet-aware evaluation in Appendix A due to limited space.

Table 4: Proportions of system ranking in human evaluation.

| Method | 1st | 2nd | 3rd |
|---|---|---|---|
| Lead-3 | 26.8 | 46.3 | 26.8 |
| UnifiedSum(E) (Hsu et al., 2018) | **37.8** | **52.4** | 9.8 |
| NeuSum (Zhou et al., 2018) | 29.3 | 39.0 | 31.7 |

FAR correlates with human preference, we rank UnifiedSum(E), NeuSum, and Lead-3 in Table 4. The order of the 1st rank in the human evaluation coincides with FAR. FAR also has higher Spearman's coefficient $\rho$ than ROUGE (0.457 vs. 0.44, n=30, threshold=0.362 at 95% significance).

Another benefit of the FAMs is that one can employ the *category breakdown* for fine-grained analysis under any metrics of interest. Here we consider ROUGE and additionally evaluate several abstractive methods: Pointer-Generator (**PG**) (See et al., 2017), **FastRL(E+A)**(extractive+abstractive) (Chen and Bansal, 2018), and **UnifiedSum(E+A)** (Hsu et al., 2018). As depicted in Table 5, not only extractive methods fail on high abstraction samples, but there is also a huge performance gap between low and high abstraction samples for abstractive methods, which suggests that existing methods achieve decent performance mainly by extraction rather than abstraction. We also found that all the compared methods perform much worse on the documents with "random" summaries, implying that the randomness in the reference summaries might introduce noise to both *model training* and *evaluation*. Despite the fact that the sample size is relatively small, we observed consistent results when analyzing different subsets of the data.

Table 5: ROUGE-1 F1 of various methods on random (R), low abstraction (L), high abstraction (H), and high quality (L + H) samples.

| | Method | R | L | H | L + H |
|---|---|---|---|---|---|
| Extractive | Lead-3 | 34.1 | 41.9 | 24.5 | 38.7 |
| | FastRL(E) (Chen and Bansal, 2018) | 33.4 | 41.8 | 31.0 | 39.8 |
| | UnifiedSum(E) (Hsu et al., 2018) | 34.2 | 42.7 | 31.1 | 40.5 |
| | Refresh (Narayan et al., 2018b) | **35.6** | 42.8 | 32.0 | 40.8 |
| | NeuSum (Zhou et al., 2018) | 34.8 | 42.8 | 30.3 | 40.4 |
| | BanditSum (Dong et al., 2018) | 35.2 | **42.9** | **34.1** | **41.1** |
| Abstractive | PG (See et al., 2017) | 32.6 | 40.6 | 27.1 | 38.1 |
| | FastRL(E+A) (Chen and Bansal, 2018) | 35.0 | 40.9 | 29.5 | 38.7 |
| | UnifiedSum(E+A) (Hsu et al., 2018) | 34.2 | 42.2 | 28.4 | 39.6 |

## 4 Analysis of Approximate Approaches to Mapping Generation

Table 6: Performance of approximate approaches that generate extractive labels. All approaches find one support sentence for each facet except the first two.

| Method | Precision | Recall | F1 |
|---|---|---|---|
| Lead-3 | 61.0 | 33.7 | 43.4 |
| Greedy ROUGE-1 F1 (Nallapati et al., 2017) | 58.2 | 30.8 | 40.3 |
| ROUGE-1 F1 | 89.7 | 57.4 | 70.0 |
| ROUGE-2 F1 | 87.4 | 56.0 | 68.3 |
| ROUGE-L Recall (Chen and Bansal, 2018) | 89.7 | 57.4 | 70.0 |
| ROUGE-L Precision (Zopf et al., 2018) | 76.1 | 48.8 | 59.4 |
| ROUGE-L F1 | 88.4 | 56.6 | 69.0 |
| AVG ROUGE-1/2/L F1 (Narayan et al., 2018b) | **90.6** | **58.1** | **70.8** |

Although the FAMs only need to be annotated once, we investigate whether such human efforts can be further reduced by evaluating *approximate* approaches that generate extractive labels. Approximate approaches typically transform one abstractive summary 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 the extracted and reference summaries (also approximately via ROUGE). Now that the FAMs serve as ground-truth extractive labels, we can evaluate how well each approach performs accurately. Since the approximate approaches do not have the notion of support group, we flatten all the support sentences in one FAM to a label set.

Due to limited space, we leave the details of the approximate approaches (most of them are self-evident) to Appendix B. The comparison results are shown in Table 6. On the bright side, approximate approaches perform relatively well (*e.g.*, 90.6% selected sentences of Narayan et al. (2018b) indeed contain salient information). This is explainable as ROUGE is good at capturing lexical overlap and as we have shown, there are many copy-and-paste reference summaries in CNN/Daily Mail. On the other hand, these approaches are not perfect and the low recall suggests that simply mapping each facet with one support sentence would miss plenty of salient sentences, which could worsen the performance of extractive models trained on such labels. That said, how to find more than one support group for each facet or multiple support sentences in one support group automatically and accurately remains an open question.

## 5 Conclusions and Future Work

We presented the promising results towards the facet-aware evaluation for extractive summarization. In the future, we will conduct large-scale human annotations in a crowd-sourcing way on the whole test set of CNN/Daily Mail. We will also investigate benchmark multi-document summarization datasets such as DUC (Paul and James, 2004) and TAC (Dang and Owczarzak, 2008) to see if the findings coincide and how we can leverage the multiple references provided for each document set in those datasets.

## References

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of ACL*.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 484–494.

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. *arXiv preprint arXiv:1809.09672*.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. *arXiv preprint arXiv:1805.06266*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

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, Caglar Gulcehre, 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.

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)*, volume 1, pages 2020–2030.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018b. 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)*, volume 1, pages 1747–1759.

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*.

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*, volume 2, pages 41–45.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 779–784.

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. *arXiv preprint arXiv:1807.02305*.

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)*, volume 1, pages 1782–1791.

# A Detailed Comparative Analysis of Extractive Methods

The proposed setup is beneficial for analyzing extractive methods as the FAMs contain rich information of the facets. In this section, we show how one can leverage the proposed setup and the FAMs by providing more detailed comparative analysis of state-of-the-art extractive methods.

As the outputs of extractive methods are sentence-based (facet-agnostic), we further define and calculate the Sentence-Aware Recall (**SAR**) using the flattened label set as in the evaluation of approximate approaches (Sec. 4). Noting that the results on the Sentence-Aware Precision and F1 are similar, we use SAR in the following analysis alone. One way to interpret SAR is to regard it as a measure of retrieving salient (support) sentences without the consideration of redundancy. For example, *two* sentences that cover the same facet would both be considered positive in SAR, while they only contribute to the coverage of *one* facet in FAR. For brevity, we denote ROUGE Precision and ROUGE Recall as **RP** and **RR**, respectively.

By comparing the results of extractive methods under both ROUGE and FAMs, we can discover many useful insights. For example, comparing FAR with RR (Fig. 2), one can see that the performance of Refresh, FastRL(E), NeuSum are quite close to the Lead-3 baseline under FAR (*i.e.*, they cover a similar number of facets), but they generally have higher RR. Such results imply that these extractive methods might have learned to add relevant keywords that are not from the support sentences, *i.e.*, the words that do not directly contribute to the major semantics but still have lexical overlap with reference summaries. It is also likely that they extract redundant support sentences that happen to have term matches with other facets. Overall, UnifiedSum(E) covers the most facets (high FAR) and also has decent lexical matches (high RR).

By comparing SAR with RP, one may tell 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 close SAR with Lead-3 but higher RP, which again implies that it might extract non-support sentences with term matches. Similarly, BanditSum is very good at finding term matches (high RP), but those terms may not contribute much to the major semantics (low SAR).

By comparing FAR with SAR (Fig. 3), we found that FastRL(E) and NeuSum have FARs similar to those of Lead-3 and Refresh, but higher SAR. One possible explanation is that FastRL(E) and NeuSum are actually 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.

# B Review of Existing Approaches that Approximate Extractive Labels

We briefly review recent approximate approaches that generate sentence-level extractive labels for extractive summarization 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 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 as possible but at wasting as little space as possible in every greedy step. Narayan et al. (2018b) rank sentences by the average of ROUGE-1/2/L F1. We also test other variants of ROUGE such as ROUGE-2 F1 and ROUGE-L F1 in the experiments.

There are approximate approaches that we do not test either due to reproducibility or the deviation of focus. For example, Cheng and Lapata (2016) label a number of documents manually and then train a rule-based classifier (with 80% acc) but the details are not provided. Zhang et al. (2018) argue that the approximate labels are suboptimal and train a latent variable extractive model to measure the Precision and Recall of the extraction instead. Hsu et al. (2018) extract a sentence as long as it can improve ROUGE recall but they only use those extracted sentences as the attention for abstractive summarization.

# C Practical Notes on CNN/Daily Mail

During the creation of FAMs on the CNN/Daily Mail dataset, we have noted several issues that are worth mentioning. We hope that the practitioners working on text summarization can be aware of these issues.

One issue is that sometimes titles and image captions are introduced in the main body of the document by mistake (usually captured by "-lrb-pictured -rrb-" or colons). This issue may lead

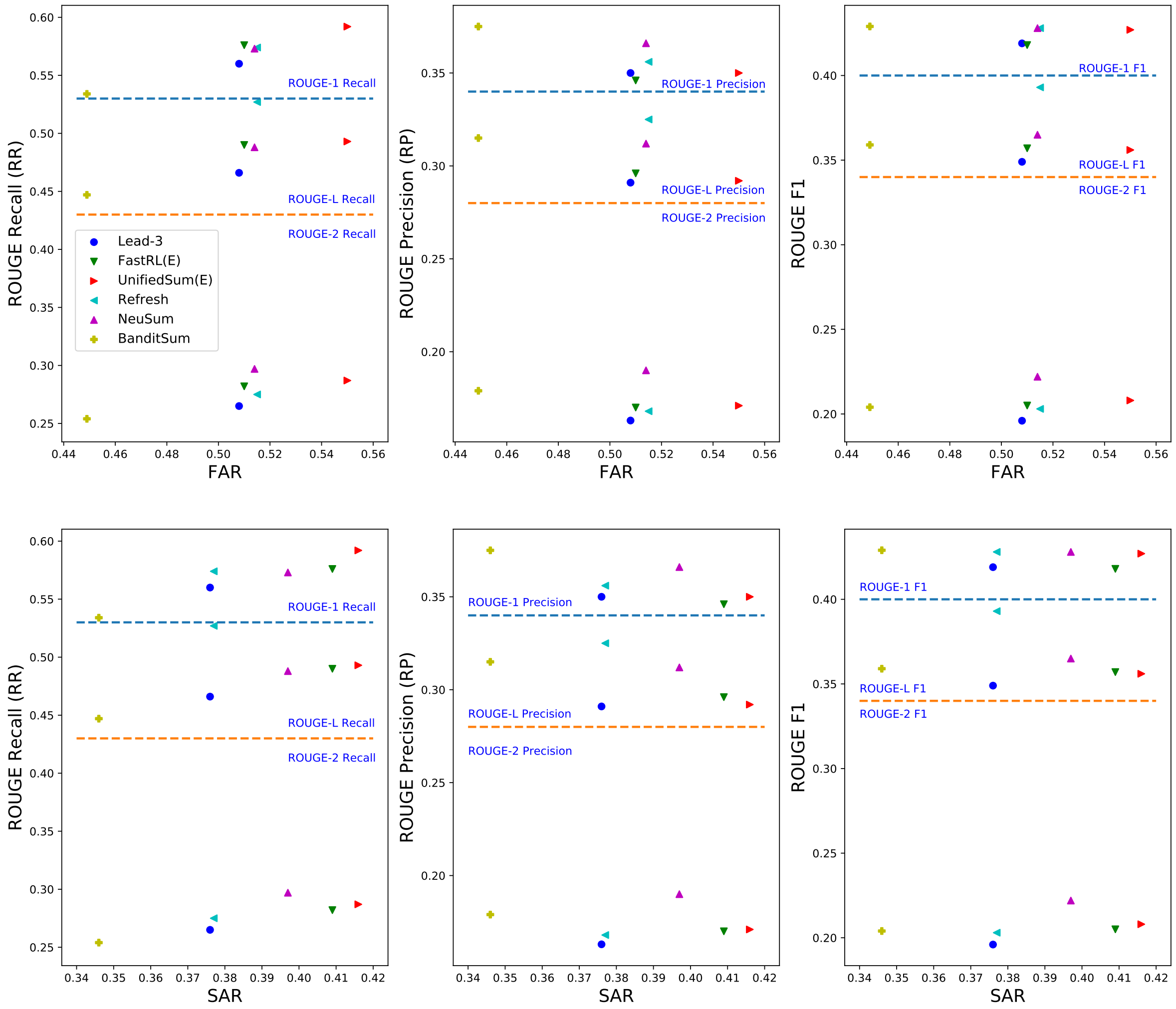

Figure 2: Comparison of various extractive methods under ROUGE and FAMs (FAR and SAR). The rankings under ROUGE-1/2/L often contradict with each other.

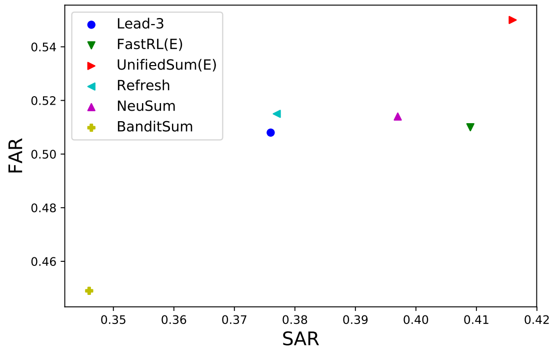

Figure 3: Comparison of extractive methods under FAR and SAR reveals their capability of extracting salient and non-redundant sentences.

to bias since the reference summary is likely to be generated by the titles and image captions (Narayan et al., 2018a). 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 examples can be found in the data file named "*low_abstraction.txt*".

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 usually be found afterward, which suggests that the first three sentences may not be from the main body of the document.

## D Detailed Examples

Here we list the full documents, reference summaries, and the corresponding FAMs of the examples shown in Table 2. In addition, we provide all of the annotated data in separate files, which have be released to the public to facilitate following studies.

Table 7: Full document, reference summary, and FAMs presented in Table 2.

**ID**: 1b2cc634e2bfc6f2595260e7ed9b42f77ecbb0ce
**Category**: Random

**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**)

**Support**:
Not applicable

Table 8: Full document, reference summary, and 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 .

**Reference Summary**:

- **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 9: Full document, reference summary, and 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

**Reference Summary**:

- **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 10: Full document, reference summary, and 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

**Support**:
Not applicable

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

| |
|---|
| **ID**: 1b2cc634e2bfc6f2595260e7ed9b42f77ecbb0ce<br>**Category**: High Abstraction |
| **Document**:<br>-LRB- CNN -RRB- He 's a blue chip college basketball recruit . She 's a high school freshman with Down syndrome .<br>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 .<br>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 .<br>So why is he taking Ellie instead ? " She 's great ... she listens and she 's easy to talk to " he said .<br>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 .<br>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 .<br>" That 's the kind of person Trey is , " she said .<br>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 .<br>Trey also got the OK from Ellie 's parents the night before via text . They were thrilled .<br>" 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 . "<br>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 .<br>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 .<br>" 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 . "<br>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<br>Trey 's mother , Shelly Moses , was also proud of her son .<br>" It 's exciting to bring awareness to a good cause , " she said . " Trey has worked pretty hard , and he 's a good son . "<br>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 .<br>As for Ellie , she ca n't stop thinking about prom .<br>" Ellie ca n't wait to go dress shopping " her mother said .<br>" Because I 've only told about a million people ! " Ellie interjected . |
| **Reference Summary**:<br>College-bound basketball star asks girl with down syndrome to high school prom. (**highly abstractive, hard to obtain by rephrasing original sentences**)<br>Pictures of the two during the "prom-posal" have gone viral. |
| **Support**:<br>Not applicable |