Analyzing Multi-Sentence Aggregation in Abstractive Summarization via the Shapley Value

Jingyi He[1]* **Meng Cao**[2] **Jackie Chi Kit Cheung**[2]

Cohere AI[1] Mila / McGill University[2]

{jingyi.he@mail, meng.cao@mail, jcheung@cs}.mcgill.ca

Abstract

Abstractive summarization systems aim to write concise summaries capturing the most essential information of the input document in their own words. One of the ways to achieve this is to gather and combine multiple pieces of information from the source document, a process we call *aggregation*. Despite its importance, the extent to which both reference summaries in benchmark datasets and system-generated summaries require aggregation is yet unknown. In this work, we propose **AGGSHAP**, a measure of the degree of aggregation in a summary sentence. We show that AGGSHAP distinguishes multi-sentence aggregation from single-sentence extraction or paraphrasing through automatic and human evaluations. We find that few reference or model-generated summary sentences have a high degree of aggregation measured by the proposed metric. We also demonstrate negative correlations between AGGSHAP and other quality scores of system summaries. These findings suggest the need to develop new tasks and datasets to encourage multi-sentence aggregation in summarization.

1 Introduction

Abstractive summarization aims to gather important information from some source text and to synthesize this information into a brief, informative, and factually correct summary. Summary-worthy information on a topic can be located in multiple parts of the document or even in different documents in the multi-document summarization case. They may appear in multiple sentences with either overlapping content or complementary information that is related in discourse. Therefore, *aggregation*, the process of combining multiple related pieces of information, is necessary to generate more useful and concise abstractive summaries.

Multi-sentence aggregation or fusion has been studied as a way to perform abstractive summarization (Barzilay and McKeown, 2005; Thadani and McKeown, 2013; Brook Weiss et al., 2022). A good summary can be written by fusing a set of salient sentences on the same topic. Therefore, the capability of aggregating information is extremely important in many summarization settings, such as long document summarization, multi-document summarization and timeline summarization. Moreover, from the theoretical perspective, multi-sentence aggregation motivates future studies of more fine-grained semantic operations (e.g. modelling contradictions and synthesizing common information across texts).

Previous studies compute proxies of abstractiveness that are closely related to the aggregation of a summary. They quantify how a summary uses words and phrases that are not found in the document, such as the percentage of novel n-grams as one of the ways to achieve highly condensed and abstractive summaries. Note that higher abstractiveness can be achieved by a broader set of rewriting operations (e.g. paraphrasing, sentence fusion, synthesizing and external knowledge). In this paper, we are particularly interested in measuring summary sentences formed by multi-sentence aggregation.

As an illustration of the difference between aggregation and abstractiveness, all three summary sentences in Table 1 contain a similar percentage of novel uni-, bi- and tri-grams, but they are formed by using different types of rewriting techniques. Novel n-grams are not able to distinguish instances that require information from multiple sentences, or that require external knowledge to infer the summary sentence, from those only formed by single-sentence compression and paraphrasing.

Another reason that aggregation is under-explored is that some popular summarization benchmark datasets are nearly extractive. As a

*Work done at Mila/McGill University

Rewriting Type	Source Document	Summary Sentence Novel [uni-, bi- tri-]grams	AGGSHAP [LM, ROUGE]
Paraphrase	(1) (CNN)Recently, a New York judge issued an opinion authorizing service of divorce papers on a husband completely via Facebook.	A court allowed a wife to serve divorce papers via Facebook . [0.25, 0.72, 0.9]	[0.449, 0.651]
Multi-sentence Fusion	(1) (CNN) Five years ago , Rebecca Francis posed for a photo while lying next to a dead giraffe . (2) The trouble started Monday , when comedian Ricky Gervais tweeted the photo with a question .	Rebecca Francis' photo with a giraffe was shared by Ricky Gervais. [0.153, 0.666, 0.909]	[0.823, 0.856]
External Knowledge	(1)The Masters 2015 is almost here. (2) To help get you in the mood for the first major of the year, [golfers' names] give the lowdown on every hole at the world-famous Augusta National Golf Club. (3) Click on the graphic below to get a closer look at what the biggest names in the game will face when they tee off on Thursday.	The 79th Masters Tournament gets underway at Augusta National on Thursday . [0.33,0.72,0.93]	[0.951, 0.896]

Table 1: Examples from CNN/DM test set show summary sentences formed by diverse types of rewriting techniques with a similar level of novel n-grams. The source sentences are highlighted based on the magnitude of their Shapley values from AGGSHAP-LM. We use three shades to indicate the relative contributions of the individual source sentence, namely [40%, 100%] , [20%, 40%) and [0, 20%) .

result, systems are not rewarded for performing aggregation. For example, Lebanoff et al. (2019b) show that only 30% of the summary sentences in the CNN/DM (Nallapati et al., 2016) are generated by fusing two or more sentences. Only relatively recently have datasets been proposed which are less extractive in terms of novel n-grams (Hermann et al., 2015; Narayan et al., 2018; Grusky et al., 2018; Koupaee and Wang, 2018; Fabbri et al., 2019). Some specifically encourage multi-sentence aggregation with summary-worthy content evenly distributed in the source (Sharma et al., 2019b). These datasets are designed to encourage systems to learn information aggregation in dispersed source document sentences, but automatically measuring this property is not yet available.

In this work, we propose a novel measure of aggregation AGGSHAP by computing a measure of many-to-one dependency between source and summary sentences. Specifically, we focus on multi-sentence aggregation where supporting information is present in the source document. Our measure uses the Shapley value (Shapley, 1953) from cooperative game theory by treating the coverage of information in a summary sentence as a coalition game played by source sentences. We compute the contribution of each source sentence using the Shapley value. Finally, the degree of aggregation of a summary sentence is characterized by the dispersion of their contributions. This measure helps us quantify intuitions about summarization datasets and the types of semantic operations that we can hope to train systems to perform using them. It also allows us to examine the phenomenon of aggregation in existing abstractive summarizers.

We validate the proposed AGGSHAP by using it to distinguish between sentences that require fusing information from multiple sentences and sentences that do not. More importantly, we show that AGGSHAP has a stronger correlation with direct human ratings of aggregation than other abstractiveness measures such as novel n-grams. Next, we apply our measure to examine the need for aggregation in existing summarization datasets and in the output of recent neural abstractive summarization models trained on these datasets. Finally, we demonstrate a negative correlation between the degree of aggregation and existing summary quality measures. This suggests that multi-sentence aggregation remains largely beyond the capability of current abstractive summarizers.

2 Related Work

2.1 Aggregation in Text Summarization

Aggregation, broadly defined, has long been a research area in NLG (Reape and Mellish, 1999; Dalianis and Hovy, 1996; Di Eugenio et al., 2005). In summarization, Jing and McKeown (1999) showed some human-written summary sentences are formed by aggregating information from multiple text spans through manual inspection. Sentence fusion is one of the most studied aggregation behaviors in the literature (Barzilay and McKeown, 2005; Elsner and Santhanam, 2011; Cheung and Penn, 2014; Yuan et al., 2021; Brook Weiss et al., 2022). Lebanoff et al. (2019a, 2020) studied sentence fusion by leveraging the syntactic cues. Much work in sentence fusion literature focuses on the syntactic dependency between similar sentences without understanding the semantic dependency between disparate sentences. As a step towards understanding semantic abstraction, Jumel et al. (2020) introduced a task of generalization and semantic aggregations of entities which is useful for performing

higher-level aggregation across sentences. Ernst et al. (2021) proposed a task of aligning summary sentences and document sentences in summarization, where aligned document sentences can be viewed as the source of aggregation.

Humans write summaries at different levels of granularity using aggregation operations beyond sentence fusion. For example, in the news domain (Hermann et al., 2015; Grusky et al., 2018), summaries are usually formed by copying and are affected by strong layout biases (Grenander et al., 2019). On the other hand, salient content may be distributed evenly throughout the text in scientific documents (Sharma et al., 2019b). Datasets for summarizing dialog (Chen et al., 2021), fiction (Kryściński et al., 2021) and meetings (Liu and Liu, 2013) show varying types of aggregation and amount of reused text (Song et al., 2020).

2.2 Measuring Aggregation

Previous work reported the percentage of novel n-grams or the notion of *Coverage* (Grusky et al., 2018) as a proxy for abstractiveness. These metrics have been adopted in other areas such as dialog (Dziri et al., 2022) to inspect the qualities and characteristics of datasets. Despite being convenient, these measures do not enable fine-grained analyses of multi-sentence aggregation.

Cheung and Penn (2013) proposed a quantitative measure of the degree of sentence aggregation at the shallow semantic level of caseframes. However, their method only accounts for limited types of aggregation and cannot be used to analyze aggregation in sentences with substantial rewriting. Wolhandler et al. (2022) proposed a metric to measure how information in a summary is dispersed in source documents in the multi-document summarization setting. They found that most summaries in certain datasets can be generated using information from only one source document.

The aggregation metric proposed in this work is inspired by the Shapley value, which is used to measure the contributions of individual players in a cooperative game (Shapley, 1953). Shapley values have been applied to settings such as feature attribution (Lundberg and Lee, 2017; Dhamdhere et al., 2019) and explaining training data contribution (Parvez and Chang, 2021).

3 Method

In this section, we propose **AGGSHAP**, an automatic metric to quantify the degree of aggregation of a summary sentence using the Shapley value. Shapley value is a concept from cooperative game theory used to determine the contributions of individual players to the outcome of a coalition game. We consider how much information in a summary sentence is covered as a coalition game in which source sentences are players. The Shapley value of a source sentence can be interpreted as its contribution to covering information in the summary sentence. The AGGSHAP score of a summary sentence captures the dispersion of source sentences' Shapley values. The degree of aggregation of a multi-sentence summary is the mean AGGSHAP scores of summary sentences.

3.1 Shapley Value Formulation

Let $D = \{d^1, ..., d^{|D|}\}$ denote a source document with $|D|$ source sentences and $S = \{s^1, ..., s^{|S|}\}$ denote a corresponding summary with $|S|$ summary sentences.

We formulate the contribution of a source sentence d_i to a summary sentence s using the Shapley value. We first define a score function $v(s, C)$ that maps a subset of source sentences $C \subseteq D$ and a summary sentence s to a real value. This represents how much information in a summary sentence is covered by the subset of source sentences. We will specify different possible instantiations of the score function $v(s, C)$ in Section 3.2 using ROUGE scores or probabilities from a conditional language model.

The Shapley value of source sentence d_i with respect to the summary sentence s is defined as

$$\phi_i(v(s,\cdot)) = \sum_{C \subset D \setminus \{i\}} \frac{|C|!(|D| - |C| - 1)!}{|D|!} [v(s, C \cup \{i\}) - v(s, C)] \quad (1)$$

where $\frac{1}{|D|!}$ is a normalization factor equal to the number of all permutations formed by $|D|$ source sentences. Given a source sentence subset C, $|C|!(|D| - |C| - 1)!$ is the number of orders in which sentences in C appear before d_i and sentences in $D \setminus (C \cup \{i\})$ can appear after d_i. We multiply the marginal gain of d_i entering into C by this factor because the marginal gains are the same for all such orders. We present a working example in the Appendix.

The time complexity for computing exact Shapley values is exponential in the number of source

sentences. Therefore, we use a Monte-Carlo method to sample subsets of source sentences and get an unbiased estimator of $\tilde{\phi}_i(v(s,\cdot))$

Measure Aggregation as Information Dispersion We define the final AGGSHAP score based on the dispersion of source sentences' Shapley values. We choose the **coefficient of variation**[1] ($\text{CV}(s) := \frac{\sigma(\phi_i(v(\cdot))}{\mu(\phi_i(v(\cdot)))}$) as the dispersion metric as it is scale-invariant. Next, we normalize the CV such that AGGSHAP $\in [0,1]$.[2]

$$\text{AGGSHAP}(s) := -\frac{\text{CV}(s)}{\sqrt{k-1}} + 1 \in [0,1] \quad (2)$$

The AGGSHAP of a summary sentence is maximized when only one of the source sentences has a non-negative Shapley value. Conversely, AGGSHAP is minimized when source sentences' Shapley values are at the same level (i.e. variance is close to 0).

3.2 Score Function Instantiations

We experimented with two methods of specifying $v(s,C)$, one based on lexical overlaps and another based on language model probabilities.

Measuring support using lexical overlap. Lexical overlap between a source and a summary sentence is one way to measure the information of s covered by a subset of source sentences C: $v^{\text{ROUGE}}(s,C) = \text{avg}(\text{ROUGE}_1(s,C) + \text{ROUGE}_2(s,C) + \text{ROUGE}_L(s,C))$. We use ROUGE recall scores in these calculations. One potential issue with lexical overlap is that it is a crude proxy of semantic relatedness, and does not account for issues such as paraphrasing.

Measuring support using LM predictions. Given a sequence-to-sequence (seq2seq) conditional text generation model, $\mathcal{M}$, parametrized by $\theta_{\mathcal{M}}$, the probability of a target sequence of n tokens $s = (s_1, ..., s_n)$ conditioning on the source text with m tokens $d = (d_1, ..., d_m)$ reflects how likely the target sequence s is to be generated. The log-likelihood of the target sequence is:

$$\mathcal{L}(s|d;\theta_{\mathcal{M}}) = \sum_{i=1}^{n} \log p(s_i|s_{<i}, d; \theta_{\mathcal{M}}) \quad (3)$$

We use the normalized log probability of the summary sentence as the value function:

$$v^{\text{LM}}(s,C) = \frac{1}{n}\mathcal{L}(s|d;\theta_{\mathcal{M}}) \quad (4)$$

$v^{\text{LM}}(\emptyset, s)$ is the (unconditional) score from a language model with no input document.

We call the two versions of the metric AGGSHAP-ROUGE and AGGSHAP-LM, respectively.

4 Evaluations of AGGSHAP

We validate the effectiveness of AGGSHAP through two experiments. First, we show fusional sentences can be distinguished from extractive ones. Next, we compute correlations between AGGSHAP or word overlap metrics on the one hand and direct assessment of aggregation on the other.

4.1 Validating AGGSHAP in Sentence Fusion

In this section, we will show that AGGSHAP effectively distinguishes instances that are a fusion of a pair of source sentences from sentences that do not require aggregation. Our assumption is that fusional instances require a higher level of aggregation and thus should be ranked higher in terms of a measure of aggregation compared to extractive instances.

Dataset. The POC (Points of Correspondence) dataset introduced by Lebanoff et al. (2019a) consists of 1,599 summary sentences and their supporting source sentence pairs from the validation and test set of the CNN/DM. The data points are deemed fusional by human annotators. The fusional instances are constructed as follows: first, the two source sentences most similar to the summary sentence based on ROUGE are selected as candidate sentences. Next, human annotators judge if the summary sentence is the fusion of these two sentences. Additionally, we extract 1,599 highly extractive summary sentences that are unlikely to require aggregation from the CNN/DM test set. These sentences are those that have at least 90% trigram overlap with a source sentence. We call this CNN/DM-EXTRACTIVE.

AGGSHAP Implementation. For all experiments and analyses in this work, we take the 30 most similar source sentences, based on ROUGE-1 F-score, to the summary sentence as the source document. We use `NLTK` for sentence tokenization

[1] σ and μ are standard deviation and mean. The coefficient of variation goes to infinity when the mean is close to zero. To avoid this, we take $\phi_i(v(\cdot) = \max(\phi_i(v(\cdot)), 0)$.

[2] Proof in the Appendix.

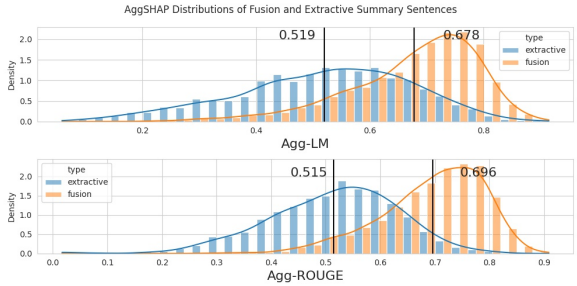

Figure 1: AGGSHAP score distributions of CNN/DM extractive and PoC fusional instances using the two variants. Gaussian kernel density estimators are fitted for each group. Mean AGGSHAP scores are annotated.

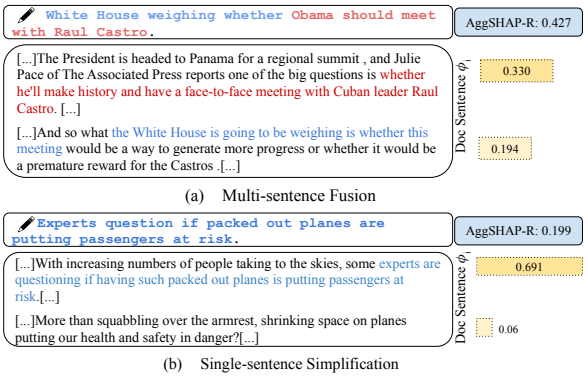

Figure 2: Eamples of multi-sentence fusion and single-sentence simplification from PoC. Document sentences with the two highest Shapley values are shown.

unless they are provided in the dataset. We use PEGASUS (Zhang et al., 2019)[3], a state-of-the-art encoder-decoder model for abstractive summarization, for computing AGGSHAP-LM. We sample 15 subsets of source sentences to compute the source sentence's Shapley value. For this particular experiment, we use PEGASUS fine-tuned on CNN/DM.[4]

Results. Figure 1 shows the distributions of AGGSHAP of fusional and extractive sentences measured by AGGSHAP-ROUGE and AGGSHAP-LM. Extractive instances and fusional instances have mean AGGSHAP-LM 0.519 and 0.678, respectively and mean AGGSHAP-ROUGE 0.515 and 0.696. The two groups are statistically significantly different with $p < 0.05$ according to the Student's t-test. AGGSHAP are effective automatic metrics capturing the difference between sentence fusion and single sentence extraction.

It is expected that novel n-grams can also separate the two groups of sentences because they are used as selection criteria for the dataset curation. Therefore they have an unfair advantage in this dataset in particular. As a strong baseline, the novel bigrams are 0.143 for fusional instances and 0.579 for extractive instances. Moreover, one should note that novel n-grams only offer an overview of how different the summary is written compared to the document. They do not provide information about the source of the supporting information.

AGGSHAP on the other hand allows fine-grained analysis of the contributions from each source sentence, which is not trivial for novel n-grams. Since computing Shapley values of source sentences is an intermediate step of AGGSHAP, we can see if the magnitude of a source sentence's Shapley value aligns with human judgments. That is, whether sentences with higher Shapley values are indeed supporting sentences. We find for 95% (1,520/1,599) of the fusional summary sentences, the highest Shapley value is assigned to one of the PoC supporting source sentences. For 50% (802/1,599) of the fusional sentences, the sentences with the top-2 highest Shapley values are the same as the pair of supporting sentences in PoC. In Figure 2, the sentence fusion example shows that the distribution of source sentences' Shapley values is flatter. In contrast, the extractive case results in a distribution with a narrow spike.

4.2 Human Evaluation of AGGSHAP

To the best of our knowledge, there is no direct assessment of multi-sentence aggregation. In order to measure how AGGSHAP aligns with human intuition about aggregation, we compute the correlations between human ratings of aggregation and AGGSHAP.

We designed the annotation procedure to directly quantify the degree of aggregation as the number of source sentences which cover all information in a summary sentence. Specifically, two of the authors of the paper are presented with 100 instances randomly sampled from the CNN/DM test set. To avoid trivial extractive cases, we filtered out summary sentences that have less than 0.3 novel bi-grams. Each instance consists of a summary sentence and the 10 most similar source sentences sorted in the decreasing order of the percentage of extractive bi-grams. We asked annotators to select the source sentence if it covers information in the summary sentence and does not cover the same

[3] `google/pegasus-cnn_dailymail` from huggingface (Wolf et al., 2020).

[4] In fact, AGGSHAP is flexible in the choice of similarity measure and language model.

piece of information as previously chosen source sentences. Finally, the number of supporting source sentences is the human rating of aggregation. We include the detailed protocol and additional analysis in the appendix.

The inter-annotator agreement measured by Krippendorff's alpha (Krippendorff, 2011) is 0.604. We used bootstrapping method with bootstrap sample size of 50 to get the 95% confidence interval of $[0.431, 0.743]$.

We compare AGGSHAP to other metrics quantifying the level of abstractiveness in summarization. **Novel n-grams** is the percentage of novel words or n-grams in a summary that is not present in the source document. **Abstractivity** (Bommasani and Cardie, 2020) derives from the notion of *coverage*, a measure of extractiveness, proposed by Grusky et al. (2018). $\text{ABS}(D, S) = 1 - \frac{\sum_{f \in \mathcal{F}(D,S)} |f|}{|S|}$, where $\mathcal{F}(D, S)$ is the set of extractive fragments in a summary extracted by greedily matching text spans shared between D and S. $|f|$ is the number of tokens in extractive fragment f.

NN-2	Abs.	AggSHAP-LM	AggSHAP-R
0.354	0.360	0.375	0.554

Table 2: Spearman correlation of various metrics and human ratings of aggregation. Abs. stands for abstractivity. All correlations have p-value $< 1.0 \times 10^{-5}$.

Table 2 presents the Spearman correlation between measures of aggregation and abstractivity and direct measure of aggregation by human annotators. AGGSHAP-ROUGE demonstrates the strongest correlation with human judgment among all measures whereas AGGSHAP-LM shows a similar level of correlation to novel n-grams and abstractivity. We speculate that the CNN/DM dataset is more extractive, thus quantifying supporting information with lexical overlaps in AGGSHAP-ROUGE is more effective than that using language model prediction.

5 Analysis

Given our automatic tool for measuring aggregation, we can use it to investigate the current state of multi-sentence aggregation in abstractive summarization. First, we study whether widely used datasets have sufficient signals to train summarization systems to perform multi-sentence aggregation (Sec. 5.1). Next, Sec. 5.2 presents how well summarizers that are trained or fine-tuned on one of these datasets (CNN/DM) perform aggregation. Finally, we are interested in whether the quality of a summary is affected by its degree of aggregation. (Sec. 5.3)

5.1 Aggregation in Summarization Datasets

In this section, we first apply AGGSHAP to measure the degree of aggregation in datasets from various genres. We are interested in the following questions in frequently used datasets: **Q1.** What is the level of aggregation exhibited by reference summaries in abstractive summarization datasets? **Q2.** Previous work reported word overlaps as intrinsic characteristics of a dataset. What is the relationship between aggregation and lexical overlaps? We answer these questions based on observation of the Table 3.

The implementation of AGGSHAP is described in Sec. 4.1. We use PEGASUS (Zhang et al., 2019) fine-tuned on corresponding datasets for AGGSHAP-LM.

Datasets. We conduct analysis on aggregation in human-written summaries of six abstractive summarization datasets. From the news domain, we analyze single-document summarization datasets CNN/Dailymail (Hermann et al., 2015), XSUM (Narayan et al., 2018), Newsroom (Grusky et al., 2018) and a multi-document summarization dataset Multinews (Fabbri et al., 2019). We also report results on PubMed (Cohan et al., 2018), a long-document dataset of scientific papers, and WikiHow (Koupaee and Wang, 2018), a dataset of articles describing a procedural task.

A1. Datasets examined show a different level of aggregation as measured by AGGSHAP, but datasets in the news domain share a similarly low level of aggregation except XSUM, as expected. Kryscinski et al. (2020) characterize CNN/DM as a benchmark dataset for the field. We show that CNN/DM has a rather low level of aggregation, novel n-grams and abstractivity. MultiNews and Newsroom display a similar level of aggregation and percentage of novel words as CNN/DM. Despite being a multi-document summarization dataset, we find that there is a substantial portion of summaries that rely on extraction from only one of the source documents in Multi-News.

The XSUM dataset has significantly higher AGGSHAP scores and novel n-grams compared to other datasets. Models that are trained on this

	CNN/DM	XSUM	Multi-News	Newsroom	PubMed	Wikihow	PoC (Fus.)	CNN/DM (Ext.)
AGGSHAP-LM	0.677	**0.800**	0.588	0.557	0.688	0.732	0.678	0.519
AGGSHAP-R	0.678	**0.828**	0.674	0.560	0.737	0.686	0.696	0.515
ABS	0.217	**0.319**	0.173	0.176	0.109	0.211	0.117	0.036
NN-1	0.203	0.356	0.277	0.202	0.171	**0.359**	0.143	0.044
NN-2	0.548	**0.816**	0.604	0.499	0.494	0.723	0.571	0.143
NN-3	0.738	**0.956**	0.764	0.615	0.696	0.908	0.802	0.254

Table 3: **Top section**: Mean aggregation scores in the test sets. **Bottom section**: Measures based on lexical overlap. Higher Novel n-grams (NN) and Abstractivity (ABS) suggest more novel phrases are used in summaries, which potentially indicates aggregation. Datasets that have the **highest** value on the measured dimension are boldfaced and the lowest values are underlined. The right section shows statistics of the PoC dataset.

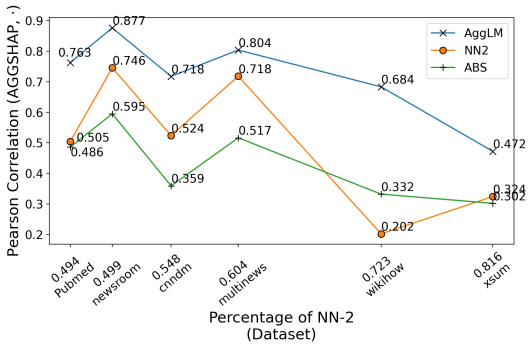

Figure 3: Pearson correlations between AGGSHAP-ROUGE and other measures as a function of dataset's percentage of novel bigrams.

dataset may be more likely to perform abstraction, multi-sentence aggregation, and utilize external knowledge. We inspected examples from the XSUM dataset with high AGGSHAP scores and their source sentences' Shapley values. We find that contributions to the summary sentence are shared among multiple source sentences.

A2. AGGSHAP and abstractiveness show strong correlations in the near-extractive datasets. The correlation between the two decreases in more abstractive datasets. We observe from Figure 3 that correlations between AGGSHAP-ROUGE and lexical overlap-based abstractiveness measures and AGGSHAP-LM decrease for datasets that have a higher proportion of novel bigrams. Since AGGSHAP-ROUGE has a moderately strong Spearman correlation with human ratings of aggregation (Table 2), decreases in correlations between AGGSHAP-ROUGE and other measures suggest that using semantic similarity measures beyond lexical overlap is necessary for investigating higher-level aggregation in more abstractive datasets.

We also notice that low novel n-gram does not necessarily imply the dataset is extractive. For example, PubMed summaries have a low proportion of novel n-grams, but they display a similarly high level of aggregation in terms of AGGSHAP as Wikihow. We speculate that mentioning proper nouns of studies in summaries is common in scientific papers, which contributes to low level of novel n-grams.

5.2 Aggregation in Current Models

We analyze the level of aggregation of summaries generated by recent abstractive summarization models trained or fine-tuned on the CNN/DM dataset. Similar analysis can be conducted on other datasets, but we focus on CNN/DM as it is one of the most frequently used datasets by models proposed over the years. We can then analyze how systems improve in aggregation. The systems-generated summaries are provided by the authors of the model and collected by Fabbri et al. (2021) under the MIT License. We follow the implementation details described in Section 4.1.

Table 4 shows the performance of summaries according to ROUGE and various measures of aggregation level and abstractiveness[5]. Overall, we find that recent abstractive summarizers display a lower level of aggregation and novel n-grams than human-written summaries across the board. Some recent models such as BART and PEGASUS use fewer novel words on average to achieve higher ROUGE scores compared to previous models, and they display a wider range of AGGSHAP scores.

We manually inspected some summaries from systems with high AGGSHAP scores. We find that BART and PEGASUS summaries often involve simple rewriting operations such as paraphrasing and concatenating text spans from multiple sentences, which may explain how they achieve higher AGGSHAP scores despite a lower proportion of

[5]Full results in the Appendix.

	ROUGE-1/2/3/L	AGGSHAP-LM	AGGSHAP-R	NN-1	NN-2	NN-3	ABS
Reference	-	**0.678**	**0.677**	**0.203**	**0.548**	**0.738**	**0.217**
M10 - **Bottom-Up**	0.412 / 0.187 / 0.106 / 0.382	0.553	**0.579**	0.155	**0.355**	**0.497**	0.019
M11 - **Improve-abs**	0.399 / 0.172 / 0.093 / 0.373	0.527	**0.580**	0.153	0.328	0.458	**0.025**
M17 - **T5**	0.448 / 0.221 / 0.134 / 0.417	0.543	0.557	**0.171**	**0.364**	**0.486**	0.011
M18 - **NeuralTD**	0.400 / 0.176 / 0.100 / 0.372	0.520	0.540	**0.173**	**0.369**	**0.497**	0.018
M21 - **UniLM**	0.431 / 0.204 / 0.122 / 0.401	0.554	0.559	0.032	0.164	0.284	**0.023**
M22 - **BART**	0.442 / 0.213 / 0.129 / 0.410	**0.555**	0.554	0.022	0.125	0.225	0.015
M23 - **Pegasus** (huge news)	0.441 / 0.215 / 0.130 / 0.410	**0.580**	0.570	0.029	0.176	0.303	0.018

Table 4: Models' ROUGE scores (partial, adapted from SummEval (Fabbri et al., 2021)) and aggregation statistics. The highest aggregation scores and percentages of novel n-grams are bolded.

novel n-grams. Bottom-Up (M10) and Improve-abs(M11) have aggregation scores on par with PEGASUS and BART. However, the quality of the generated text is significantly lower as shown by ROUGE. Enabling multi-sentence aggregation in abstractive summarization is a promising open research area, since there is still a large gap in aggregation between system-generated summaries and reference summaries.

5.3 Aggregation Versus Summary Quality

We are interested in whether systems can perform aggregation as well as generate high-quality summaries. Kryściński et al. (2018) reported a negative result where novel n-grams negatively correlate with ROUGE scores. Inspired by this, we inspect if there is a similar trade-off between aggregation and summaries' quality.

We use the human annotations from SummEval (Fabbri et al., 2019) (11 abstractive models evaluated on *Coherence*, *Factuality*, *Fluency* and *Relevance*) and NeR18 (Grusky et al., 2018) (7 systems evaluated on *Coherence*, *Fluency*, *Informativeness*, *Relevance*). We compute system-level correlations between AGGSHAP and human judgement scores. We follow the definition of system-level correlation in (Louis and Nenkova, 2013), as follows: first, we compute a system-level score of the system by averaging the scores of interest over all instances in the dataset. Next, we compute Kendall's τ between the rankings of the systems.

	SummEval				Newsroom			
	COH	FAC	FLU	REL	COH	FLU	INF	REL
NN-1	0.090	-0.310	-0.270	-0.240	-0.520	-0.520	-0.520	-0.430
NN-2	-0.050	-0.380	-0.270	-0.310	-0.900	-0.900	-0.710	-0.810
NN-3	-0.020	-0.420	-0.310	-0.350	-0.810	-0.810	-0.620	-0.710
ABS	0.050	-0.490	-0.160	-0.270	-0.330	-0.330	-0.330	-0.240
AGG-LM	-0.117	-0.450	-0.243	-0.283	-0.810	-0.810	-0.619	-0.714
AGG-R	-0.133	-0.467	-0.259	-0.267	-0.714	-0.714	-0.714	-0.619

Table 5: System-level Kendall's tau correlation coefficients between metrics of interest (AGGSHAP and novel n-grams) and human judgments. AGG are the abbreviated version of AGGSHAP.

Table 5 shows the results of these correlation computations. Both AGGSHAP and abstractiveness measures have consistent negative correlations with human ratings of quality. AGGSHAP show moderate negative correlations in factuality and weak negative correlations in relevance, indicating systems that attempt to aggregate are likely to introduce factual error into the summary. Weak correlations are shown in coherence and fluency dimensions because neither abstractiveness metrics nor AGGSHAP measure the inter-sentence connections of a summary.

One of the findings from SummEval is that reference summaries have lower scores than extractive systems (e.g. lead-3) across all four dimensions. This indicates that human judges prefer nearly extractive summaries in this dataset. Therefore, systems that are able to perform multi-sentence aggregation might not be rewarded by current evaluation schemes. To track the progress of aggregation in summarization systems, human annotators should directly assess the degree of aggregation.

6 Conclusion

In this paper, we propose AGGSHAP to quantify aggregation operations in abstractive summarization. Our metric effectively distinguishes sentences that require multiple points of dependencies from those that do not in a dataset containing fusional summary sentences. Moreover, it has a stronger correlation with human ratings of aggregation than existing n-grams overlap measures. We use AGGSHAP to compare the levels of aggregation in summarization datasets and conclude that most recent summarization datasets from the news domain contain limited instances of reference summaries that require aggregation. We show that abstractive summarization models rarely perform semantic aggregation beyond simple concatenation of text units. Finally, we find improvements in the dimension of aggregation may not be rewarded by current evalu-

ation schemes of general summarization qualities. Future evaluations should thus focus specifically on the issue of aggregation, ideally in a domain or setting whether aggregation is necessary to derive a reference summary or a useful conclusion.

Acknowledgements The authors acknowledge the material support of NVIDIA in the form of computational resources. The first author is supported by the Fonds de recherche du Quebec – Nature et technologies. The last author is supported in part by the Canada CIFAR AI Chair program.

Limitations

Computation Efficiency. As noted in the method section 3.2, computation of the Shapley value has exponential time complexity. We address this issue by using Monte-Carlo sampling method but it is still computationally expensive to conduct analysis of aggregation at a large scale. For each sentence evaluated, it requires (#Number of source sentences $\times$ #Shapley value sample) times of forward pass to compute the estimated Shapley values of source sentences. We only conducted analysis on the test set of the datasets and, for example, it took 24 hours on a single V100 GPU with 16GB of memory to evaluate AGGSHAP-LM of CNN/DM (11490 summaries with 3 sentences per summary on average).

Interpretation. In this work, we mainly focused on analyzing how summary sentences aggregate information that is faithful to the source document, and we did not address cases where information has to be drawn from external knowledge. AGGSHAP is not suitable for interpreting low-quality examples as the fundamental assumption of AGGSHAP is to quantify the degree of aggregation by how well the summary sentence is supported by the source. AGGSHAP may fail to find any supporting information from the source and consider the low-quality example to have a high level of aggregation.

Potential Risks. All scientific artifacts in this study have been made publicly available and consistent with their intended use and access conditions.

References

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence Fusion for Multidocument News Summarization. *Computational Linguistics*, 31(3):297–328.

Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. 2019. Better rewards yield better summaries: Learning to summarise without references. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3110–3120, Hong Kong, China. Association for Computational Linguistics.

Rishi Bommasani and Claire Cardie. 2020. Intrinsic evaluation of summarization datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.

Daniela Brook Weiss, Paul Roit, Ori Ernst, and Ido Dagan. 2022. Extending multi-text sentence fusion resources via pyramid annotations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1854–1860, Seattle, United States. Association for Computational Linguistics.

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

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Jackie Chi Kit Cheung and Gerald Penn. 2013. Towards robust abstractive multi-document summarization: A caseframe analysis of centrality and domain. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1233–1242, Sofia, Bulgaria. Association for Computational Linguistics.

Jackie Chi Kit Cheung and Gerald Penn. 2014. Unsupervised sentence enhancement for automatic summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 775–786, Doha, Qatar. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

Volume 2 (Short Papers), pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Hercules Dalianis and Eduard Hovy. 1996. On lexical aggregation and ordering. In *Eighth International Natural Language Generation Workshop (Posters and Demonstrations)*.

Kedar Dhamdhere, Ashish Agarwal, and Mukund Sundararajan. 2019. The shapley taylor interaction index.

Barbara Di Eugenio, Davide Fossati, Dan Yu, Susan Haller, and Michael Glass. 2005. Aggregation improves learning: Experiments in natural language generation for intelligent tutoring systems. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 50–57, Ann Arbor, Michigan. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo Ponti, and Siva Reddy. 2022. Faithdial: A faithful benchmark for information-seeking dialogue. *arXiv preprint, arXiv:2204.10757*.

Micha Elsner and Deepak Santhanam. 2011. Learning to fuse disparate sentences. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 54–63, Portland, Oregon. Association for Computational Linguistics.

Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. 2021. Summary-source proposition-level alignment: Task, datasets and supervised baseline. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 310–322. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. 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 6019–6024, Hong Kong, China. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. 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 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–697, Melbourne, Australia. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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.

Yichen Jiang and Mohit Bansal. 2018. Closed-book training to improve summarization encoder memory. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4067–4077, Brussels, Belgium. Association for Computational Linguistics.

Hongyan Jing and Kathleen R. McKeown. 1999. The decomposition of human-written summary sentences. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 129–136, New York, NY, USA. Association for Computing Machinery.

Clément Jumel, Annie Louis, and Jackie Chi Kit Cheung. 2020. TESA: A Task in Entity Semantic Aggregation for abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8031–8050, Online. Association for Computational Linguistics.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *CoRR*, abs/1810.09305.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. Association for Computational Linguistics.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization.

Logan Lebanoff, Franck Dernoncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu. 2020. Learning to fuse sentences with transformers for summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4136–4142, Online. Association for Computational Linguistics.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019a. Analyzing sentence fusion in abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 104–110, Hong Kong, China. Association for Computational Linguistics.

Logan Lebanoff, Kaiqiang Song, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 2019b. Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Fei Liu and Yang Liu. 2013. Towards abstractive speech summarization: Exploring unsupervised and supervised approaches for spoken utterance compression. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1469–1480.

Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistics*, 39(2):267–300.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

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, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. 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.

Md Rizwan Parvez and Kai-Wei Chang. 2021. Evaluating the values of sources in transfer learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5084–5116, Online. Association for Computational Linguistics.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 646–653, New Orleans, Louisiana. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Mike Reape and Chris Mellish. 1999. Just what is aggregation anyway? In *ENLG 1999)*, pages 20–29.

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.

L. S. Shapley. 1953. *17. A Value for n-Person Games*, pages 307–318. Princeton University Press.

Eva Sharma, Luyang Huang, Zhe Hu, and Lu Wang. 2019a. An entity-driven framework for abstractive 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 3280–3291, Hong Kong, China. Association for Computational Linguistics.

Eva Sharma, Chen Li, and Lu Wang. 2019b. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Kaiqiang Song, Bingqing Wang, Zhe Feng, Liu Ren, and Fei Liu. 2020. Controlling the amount of verbatim copying in abstractive summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Kapil Thadani and Kathleen McKeown. 2013. Supervised sentence fusion with single-stage inference. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1410–1418, Nagoya, Japan. Asian Federation of Natural Language Processing.

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

Ruben Wolhandler, Arie Cattan, Ori Ernst, and Ido Dagan. 2022. How "multi" is multi-document summarization? *ArXiv*, abs/2210.12688.

Ruifeng Yuan, Zili Wang, and Wenjie Li. 2021. Event graph based sentence fusion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4075–4084, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences.

A Appendix

A.1 Human Annotation Details

We give the following instructions to the annotators:

1. Read the summary sentence
2. Read the supporting sentences in the order presented in the spreadsheet (supporting sentence_0 to supporting sentence_9)
    - If the supporting sentence covers the information in the summary sentence and this piece of information has not been covered by previous supporting sentences then highlight it.
    - If two sentences are identical or very similar in content, highlight both.
    - If no single supporting sentence covers information in the summary sentence, enter *missing*

To evaluate the inter-annotator agreement of the selected supporting sentence (i.e. how well annotators agree on which source sentences are supporting sentences), we computed the Krippendorff alpha of annotated instances. The Krippendorff alpha is 0.714 with a 95% confidence interval of $[0.579, 0.835]$ from bootstrapping with bootstrap sample size of 100.

We manually inspected some instances where two annotators do not agree on the number of supporting sentences. We found that most ambiguities came from judging whether two supporting sentences are very similar in content or not.

A.2 Full results of Table 4

See Table 6.

A.3 Examples of Human-written Summaries

Table 7-10 shows randomly sampled reference summary sentences from the six datasets we evaluated. We sampled examples that are extractive AGGSHAP-LM < 0.45 and are of higher level of aggregation AGGSHAP-LM > 0.7

A.4 Examples of System-generated Summaries

Table 11 shows an example in which the reference summary contains aggregations of information from multiple points in the source text. The summary generated by PEGASUS (Zhang et al.,

	ROUGE-1/2/3/L	AGGSHAP-LM	AGGSHAP-R	NN-1	NN-2	NN-3	ABS
Reference	-	**0.678**	**0.677**	**0.203**	**0.548**	**0.738**	**0.217**
M8 - **Pointer Generator**	0.392 / 0.172 / 0.100 / 0.360	0.486	0.517	0.129	0.250	0.344	0.002
M9 - **Fast-abs-rl**	0.406 / 0.177 / 0.098 / 0.381	0.515	0.524	0.149	0.347	0.482	0.014
M10 - **Bottom-Up**	0.412 / 0.187 / 0.106 / 0.382	0.553	**0.579**	0.155	**0.355**	**0.497**	0.019
M11 - **Improve-abs**	0.399 / 0.172 / 0.093 / 0.373	0.527	**0.580**	0.153	0.328	0.458	**0.025**
M12 - **Unified-ext-abs**	0.404 / 0.179 / 0.104 / 0.368	0.470	0.502	0.138	0.258	0.351	0.013
M13 - **ROUGESal**	0.402 / 0.180 / 0.105 / 0.368	0.488	0.515	0.149	0.285	0.387	0.018
M14 - **Multi-task (Ent + QG)**	0.395 / 0.176 / 0.104 / 0.363	0.492	0.520	0.141	0.275	0.373	0.015
M15 - **Closed book decoder**	0.398 / 0.176 / 0.103 / 0.364	0.484	0.512	0.137	0.261	0.355	0.013
M16 - **SENECA**	0.415 / 0.184 / 0.105 / 0.381	0.521	0.568	0.161	0.340	0.453	0.013
M17 - **T5**	0.448 / 0.221 / 0.134 / 0.417	0.543	0.557	**0.171**	**0.364**	**0.486**	0.011
M18 - **NeuralTD**	0.400 / 0.176 / 0.100 / 0.372	0.520	0.540	**0.173**	**0.369**	**0.497**	0.018
M20 - **GPT-2** (supervised)	0.398 / 0.176 / 0.099 / 0.367	0.470	0.506	0.010	0.043	0.063	0.010
M21 - **UniLM**	0.431 / 0.204 / 0.122 / 0.401	**0.554**	0.559	0.032	0.164	0.284	**0.023**
M22 - **BART**	0.442 / 0.213 / 0.129 / 0.410	**0.555**	0.554	0.022	0.125	0.225	0.015
M23 - **Pegasus** (huge news)	0.441 / 0.215 / 0.130 / 0.410	0.580	0.570	0.029	0.176	0.303	0.018

Table 6: Models' ROUGE scores (Adapted from SummEval (Fabbri et al., 2021)) and aggregation statistics. The highest aggregation scores and percentages of novel N-grams are bolded. We remove M19 BertSum-abs from the analysis as no punctuation at the end of sentences resulting in misleadingly high aggregation scores.

Aggregation Candidate	Summary [Agg-LM, Agg-ROUGE]
Example 1: (1) Investigators found that a number of flavors were labeled 'healthy' - brimming with fiber, protein and antioxidants, while being low in fat and sodium.	FDA Investigators found that a number of flavors were labeled 'healthy' - brimming with fiber and antioxidants, while being low in fat and sodium . [0.431, 0.414]
Example 2: (1) '(CNN) Five years ago , Rebecca Francis posed for a photo while lying next to a dead giraffe . (2) The trouble started Monday , when comedian Ricky Gervais tweeted the photo with a question .	Rebecca Francis ' photo with a giraffe was shared by Ricky Gervais . [0.759, 0.801]

Table 7: CNN/DM extractive (Top) and higher-level aggregation (Bottom)

2019) contains aggregations from three source sentences, as does the human summary, while the summary generated by improve-abs (Kryściński et al., 2018) is produced by compressing a single sentence.

A.5 Abstractive Models in Section 5.2

Here we cite the list of abstractive summarization models we evaluated for aggregation. We evaluated the summaries generated by the following systems: (M8) Pointer Generator (See et al., 2017), (M9) Fast-abs-rl (Chen and Bansal, 2018), (M10) Bottom-up (Gehrmann et al., 2018), (M11) Improve-abs (Kryściński et al., 2018), (M12) Unified-ext-abs (Hsu et al., 2018), (M13) ROUGE-Sal (Pasunuru and Bansal, 2018), (M14) Multi-task(Ent+QG) (Guo et al., 2018), (M15) Closed book decoder (Jiang and Bansal, 2018), (M16) SENECA (Sharma et al., 2019a), (M17) T5 (Raffel et al., 2020), (M18) NeuralTD (Böhm et al., 2019), (M20) GPT-2 (supervised) (Ziegler et al., 2019), (M21) UniLM (Dong et al., 2019), (M22) BART (Lewis et al., 2020) and (M23) PEGASUS (Zhang et al., 2019).

Aggregation Candidate	Summary [Agg-LM, Agg-ROUGE]
Example 1: (1) Joseph Fox photographed the mudlarkers who comb the shore of London's River Thames.	All photographs taken by Joseph Fox. [0.192, 0.300]
Example 2: (1) <n> Protesters allege Edir Frederico Da Costa, 25, was "brutally beaten" by Met Police officers earlier this month. (2) <n> The Independent Police Complaints Commission (IPCC) is investigating the treatment of Mr Da Costa, who died six days after he was stopped by police. (3) <n> Mr Da Costa, known by friends as Edson, died on 21 June, six days after being stopped in a car in Woodcocks, Beckton, in Newham, east London.	Protesters have faced off with police in a demonstration over the death of a man after a traffic stop. [0.799, 0.854]

Table 8: XSUM extractive (Top) and higher-level aggregation (Bottom) instances

Aggregation Candidate	Summary [Agg-LM, Agg-ROUGE]
Example 1: (1) source sent: (PHOTOS: Scenes from Eric Cantor HQ) Asked about his future plans, Cantor replied: "That's probably between my wife and me." Addressing his colleagues earlier, Cantor's words drove Speaker John Boehner (R-Ohio) to tears.	"That's probably between my wife and me," he said. [0.182, 0.382]
Example 2: (1) More than 90% of the parts needed to restore a 1967 Mustang convertible are available new as Ford-licensed reproduction components, allowing enthusiasts to basically build from scratch a new Mustang of that era. (2) To build up a Mustang using the body shell, the powertrain, suspension and brakes, the electrical systems, the interior and trim can either be bought new or transferred from an existing car to the new body.	Just in time for classic car buffs' Christmas, Ford has added a brand-new shell for the '67 Mustang convertible to its Ford Restoration Parts line, giving enthusiasts a chance to build their own from scratch, the Los Angeles Times reports. [0.818, 0.834]

Table 9: Multi-news extractive (Top) and higher-level aggregation (Bottom) instances

Aggregation Candidate	Summary [Agg-LM, Agg-ROUGE]
Example 1: (1) source sent: If you don't have a water bottle or hot compress pad, you can pour warm water (104-108 degrees Fahrenheit) into a basin and immerse the injured area in the water for 30-45 minutes.It's normal to feel severe pain as the tissue begins to warm up, so do not be alarmed about this.	Pour warm water into a basin. [0.323, 0.306]
Example 2: (1) Disney Parks park maps aren't just written in English and Spanish. (2) , (3) Parade routes differ between the different parks. (4) The parade route will be marked on the map by some type of dotted or broken line.	Visit the Disney Park that the parade will be shown in. [0.818, 0.856]

Table 10: WikiHow extractive (Top) and higher-level aggregation (Bottom) instances

Source: British jihadis have posted pictures of junk food and drinks such as Burger King, Pringles and mojitos which they have had carried across the Turkish border into Syria. [...] It's not the first time ISIS fighters have been caught with fast food sneaked across the border. Last month a delighted fighter known only as Ghareeb posted a picture of a McDonalds bag on his social media page.
Reference summary: ISIS fighters have posted pictures on social media of Western junk food .
Pegasus generated: ISIS fighters have been posting pictures of food and drinks smuggled in .
Improve-abs generated: british jihadis have posted pictures of junk food and mojitos .

Table 11: Example of summary sentences aggregating information from three source sentences in CNN/DM dataset. Human editor aggregates *British jihadis* and *a delighted fighter know as Ghareeb* as *ISIS fighters*, and aggregates *junk food such as Burger King, Pringles* and *McDonalds* as *Western junk food*.