OASum: Large-Scale Open Domain Aspect-based Summarization

Xianjun Yang[1*] **Kaiqiang Song**[2*] **Sangwoo Cho**[2]
Xiaoyang Wang[2] **Xiaoman Pan**[2] **Linda Petzold**[1] **Dong Yu**[2]
{xianjunyang,petzold}@ucsb.edu
{riversong,swcho,shawnxywang,xiaomanpan,dyu}@tencent.com
[1] University of California, Santa Barbara [2] Tencent AI Lab, Seattle

Abstract

Aspect or query-based summarization has recently caught more attention, as it can generate differentiated summaries based on users' interests. However, the current dataset for aspect or query-based summarization either focuses on specific domains, contains relatively small-scale instances, or includes only a few aspect types. Such limitations hinder further explorations in this direction. In this work, we take advantage of crowd-sourcing knowledge on `Wikipedia.org` and automatically create a high-quality, large-scale open-domain aspect-based summarization dataset named ***OASum***, which contains more than 3.7 million instances with around 1 million different aspects on 2 million Wikipedia pages. We provide benchmark results on *OASum* and demonstrate its ability for diverse aspect-based summarization generation. To overcome the data scarcity problem on specific domains, we also perform zero-shot, few-shot, and fine-tuning on seven downstream datasets. Specifically, zero/few-shot and fine-tuning results show that the model pre-trained on our corpus demonstrates a strong aspect or query-focused generation ability compared with the backbone model. Our dataset and pre-trained checkpoints are publicly available.[1]

1 Introduction

Text summarization aims to provide accurate, concise, and useful information about the original inputs for users to fast browse. Existing generic summarization or aspect agnostic summarization methods (See et al., 2017; Narayan et al., 2018; Liu, 2019; Zhang et al., 2020; Liu et al., 2022; Wang et al., 2022b) typically generate only one summary for all different requests which is not optimal for diverse demands. It could fail to preserve the required information that the user needs or miss important details (Woodsend and Lapata, 2012; Angelidis and Lapata, 2018). By contrast, the aspect or query-based summarization methods (Xu and Lapata, 2020; Zhong et al., 2021; Ahuja et al., 2022) provide the flexibility of generating summaries for differentiated demands.

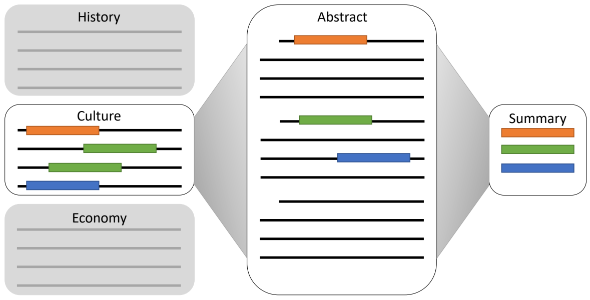

Figure 1: The left section titles are naturally adopted from the Wikipedia page to serve as different aspects, while the middle abstract is the head section serving as an overall summary of the article. The right part is the corresponding aspect-based summary.

However, existing datasets for aspect-based summarization are either on a small scale (Wang et al., 2022a; Bahrainian et al., 2022a; Kulkarni et al., 2020), only focusing on a specific domain (Zhong et al., 2021; Zhan et al., 2022), or with limited aspects (Fremann and Klementiev, 2019; Hayashi et al., 2021). To the best of our knowledge, there is no existing dataset with millions of aspects and instances for large-scale open-domain aspect-based summarization. Models trained in a small-scale dataset with limited instances or aspects may fail to adapt to other aspects or domains in realistic open-domain scenarios.

To tackle the limitations of the existing aspect-based summarization datasets, we propose a large-scale open-domain aspect-based summarization dataset named ***OASum***. Table 1 compares ***OASum*** with seven existing datasets for aspect or query-based summarization.

To create the data, as illustrated in Fig. 1, we take advantage of crowd-sourcing knowledge in

[*]Work done during Xianjun Yang's internship at Tencent AI Lab Seattle. The first two authors contributed equally.
[1]https://github.com/tencent-ailab/OASum

Findings of the Association for Computational Linguistics: ACL 2023, pages 4381–4401
July 9-14, 2023 ©2023 Association for Computational Linguistics

Type	Dataset	Domain	#Instances	#Input Tk.	#Output Tk.	#Asp. Type	Method
Query	AQualMuse	General	7,168	9,764	106	7,160	**A**
	QMSum	Meeting	1,808	9,070	70	1,566	**M**
	SQuALITY	Sci-fi	2,540	6,052	252	437	**M**
Aspect	CovidET	Reddit	7,112	192	27	7	**M**
	MA-News	News	286,701	1,350	54	6	**A**
	NEWTS	News	6,000	602	74	50	**M**
	Wikiasp	Wikipedia	399,696	13,672	214	200	**A**
	ASPECTNews	News	400	248	115	4	M
Ours	***OASum***	Wikipedia	**3,747,569**	1,612	40	**1,045,895**	**A**

Table 1: Statistics of query/aspect-based summarization datasets. The last column contains the methods of dataset creation. **A** stands for "Automatic", **M** stands for "Manual". **#Input Tk.** and **#Output Tk.** represent the number of input and output token lengths, respectively. **#Asp. Type** is the number of all aspect types. **#Instances** stands for the total number of $(article, summary)$ pairs in the corresponding dataset.

English Wikipedia pages and parse them to collect the information on each page including the title of each section and its contents. On the one hand, the head section is a natural abstract of each Wikipedia page. On the other hand, the remaining sections describe different aspects of that page. Therefore, we use the section titles as the aspect inputs and apply a rule base process to automatically select sentences in the abstract section as the matched summary for different aspects.

Specifically, we use the Wikipedia dump on 2022/06/21. It contains around 6.3 million pages after parsing. After preprocessing, we keep approximately 2 million pages that contain around 3.7 million instances in total. Our dataset includes 1,045,895 different aspects on 32,956 different domains (categorized with the original Wikipedia pages), providing plenty of useful information for open-domain aspect-based summarization. It also provides abstractive summaries that are not directly extracted from the original inputs. To ensure the quality, we perform a manual evaluation with randomly selected 66 pages, and the overall satisfaction score is 3.13 out of 5. Based on our curated million-level aspect-based summarization corpus, we pretrain Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) model on ***OASum*** in an end-to-end way. Compared with the backbone model, our pretrained model achieves better performance on six out of seven downstream tasks for the fine-tuning and zero-shot settings and all six downstream tasks for the few-shot setting.

The contributions of our work are in two folds:

- We create the first large-scale open-domain aspect-based summarization dataset namely ***OASum***. The statistic shows ***OASum*** contains a variant of input lengths, highly abstractive summaries, and contents in a large number of aspects and domains. Overall, it contains more than 3.7M instances and 1M different aspect types.
- We further pre-train the backbone model on ***OASum*** and test the pretrained model with zero-shot, few-shot, and fine-tuning settings on seven downstream datasets. The results illustrate ***OASum*** provides useful information that can further benefit other query/aspect-based summarization tasks.

2 Related Works

Aspect / Query based Summarization. Aspect-based summarization was proposed to generate summaries based on different aspects for opinions and reviews (Kansal and Toshniwal, 2014; Wu et al., 2016; Akhtar et al., 2017; Angelidis and Lapata, 2018; Coavoux et al., 2019; Tan et al., 2020). Recent researches attempt to summarize different aspects for news (Fremann and Klementiev, 2019; Bahrainian et al., 2022a; Ahuja et al., 2022) and other domains (Hayashi et al., 2021; Zhan et al., 2022). Similarly, query-based summarization (Kulkarni et al., 2020; Zhong et al., 2021; Wang et al., 2022a) takes finer-grained questions as input for summarization. As our ***OASum*** contains even finer-grained aspects, we believe it can benefit both tasks.

Wikipedia as data. Wikipedia has been widely used as a rich source for many NLP tasks, including Language Modeling (Guo et al., 2020), Question answering (Yang et al., 2015; Rajpurkar et al., 2018), Information Extraction (Wu and Weld, 2010), Dialogue (Dinan et al., 2018), and Summarization (Liu et al., 2018; Ghalandari et al., 2020; Sun et al., 2021; Iv et al., 2022). WikiAsp (Hayashi et al., 2021) directly uses external documents to generate the corresponding section contents with limited aspect types. Comparatively, ***OASum*** employs a matching method to obtain the aspect-based summaries from the head section of a Wikipedia page according to their similarities to the remaining page, resulting in more than one million aspect types.

Long document summarization. The summarization task typically has long inputs (Shen et al., 2022; Kryściński et al., 2021; Song et al., 2022; Cho et al., 2022). Recent Transformer-based models (Radford et al., 2018; Devlin et al., 2019; Lewis et al., 2020) with full attention require a huge amount of GPU memories during training. Efficient transformers (Beltagy et al., 2020; Zaheer et al., 2020; Guo et al., 2022) are proposed for handling long sequences with simplified attention. Extract-then-generate strategies (Zhong et al., 2020; Pilault et al., 2020; Song et al., 2020; Zheng et al., 2020) have been used for such issues. As ***OASum*** has a large number of instances containing more than 4096 input tokens, we thus use LED (Beltagy et al., 2020) as our backbone model.

3 *OASum* Dataset

3.1 Dataset Construction

This dataset is built upon the observation that the abstract section is a natural summary for the later sections, and sentences in the abstract section may present one or more aspects described in the later sections. We use the English Wikipedia dump from 2022-06-20 for creating our dataset. Originally, there are over 6.33 million pages.

Data Cleaning. Each Wikipedia page is written in a special markup language. We first adopt a tool[2] (Pan et al., 2017) to remove all undesired markups (e.g., templates, internal/external links, and HTML tags) and keep section boundaries. Next, we discard structural sections including *References*, *See also*, *External links*, *Further reading*, and *Bibliography*. We further remove structural contents such as item lists in other sections. Finally, we split sentences using Spacy[3]. We collect 3.75 million non-empty pages after data cleaning.

Aspect Summaries Construction. An abstract sentence should be considered as a summary sentence of the specific aspect iff it has enough content overlap with the corresponding section. Shown in Algorithm 1, we first use a greedy method to map each abstract sentence to a list of sentences in the later sections. Then, we assign a matching score $\mathcal{S}(x, \alpha)$ for each abstract sentence x and a potential aspect α. We use the *ROUGE-1-recall* between the abstract sentence x and the intersection of its mapped sentences $\mathcal{M}(x)$ and the sentences in the aspect section Y_a.

Algorithm 1 Greedy Mapping

Input: sentence x, set of sentences Y
Output: set of mapped sentences S

$S \leftarrow \varnothing$; // Set of mapped Sentences
Score $\leftarrow 0$; // Current ROUGE-1-Recall
while $Y \setminus S \neq \varnothing$ **do**
  $\delta \leftarrow 0$; // Best Improvements
  $\eta \leftarrow null$; // Best Candidate
  for $y \in Y \setminus S$ **do**
    $S' \leftarrow S \cup \{y\}$
    if *ROUGE-1-Recall*(x, S') − *Score* $> \delta$ **then**
      $\delta \leftarrow$ *ROGUE-1-Recall*(x, S') − *Score*;
      $\eta \leftarrow y$;
    end if
  end for
  if $\delta \leq 0$ **then**
    Break;
  end if
  Score $\leftarrow$ *Score* $+ \delta$;
  $S \leftarrow S \cup \{\eta\}$;
end while
return S

$$\mathcal{S}(x, a) = \textit{ROUGE-1-recall}(x, Y_a \cap \mathcal{M}(x)). \quad (1)$$

This score indicates the content overlap between the abstract sentence and the aspect section. To filter out sentences with limited content overlap, an aspect-based summary includes only abstract sentences with a matching score $\mathcal{S}(x, a)$ greater or equal to a pre-defined threshold λ. To determine the exact value of the threshold, we try $\lambda \in [0.3, 0.4, 0.5, 0.6, 0.7]$ and evaluate them manually. Specifically, we randomly pick 66 Wikipedia pages consisting of 103 aspect-summary pairs for each threshold, and assigned them to 5 experts for evaluating the dataset quality. The Cohen's kappa between annotators is calculated to be 0.43, showing moderate agreement. The results are shown in Table 2. We then choose to use $\lambda = 0.5$.

Data Splitting. We split the data into train/validation/test sets with 94%/3%/3% of the Wikipedia pages after data cleaning. After filtering out the instances where the summary is longer than the input text, we obtain 3,523,986/111,578/112,005 instances for the train/validation/test set. In Ta-

[2] https://github.com/panx27/wikiann
[3] model "en-core-web-sm", version 3.0.0

$\lambda =$	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**
avg Score	2.61	2.85	3.13	3.05	2.75

Table 2: Summary quality with different thresholds. The scores are in the range of 1-5, representing *very bad*, *bad*, *fair*, *good*, and *excellent*, respectively.

ble 3, we demonstrate the aspect-based summaries constructed from the "Seattle" Wikipedia Page[4].

3.2 Data Statistics and Analysis

In this section, we demonstrate the properties of our dataset from different perspectives including the statistics of input and output length, abstractiveness, aspect distribution, and page ontology. **Length**. On average, the input documents have 1,856.09 tokens or 62.23 sentences, and the output summary contains 48.61 output tokens or 1.77 sentences. In Fig. 2, we further plot the length

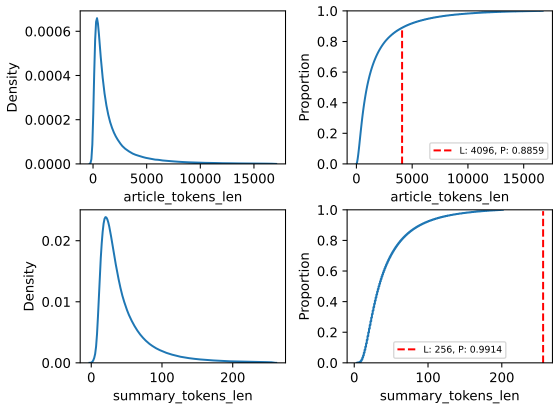

Figure 2: Input (Top) and output (Bottom) length in terms of tokens with Probability Density Functions (Left) and Cumulative Distribution Functions (Right). The red dashed lines represent the truncation we used for model training. L and P represent the token length and cumulative probability, respectively.

distribution functions for both inputs and outputs. We find ***OASum*** contains a variety of lengths for both inputs and outputs. The inputs can range from 4 tokens to 78,498 tokens, while the outputs can range from 3 to 9,792. This creates a playground suitable for tackling long-tail problems that involve both lengthy inputs and extended summaries. In addition, the compression ratios of ***OASum*** are distributed widely from 0.68[5] to 32,148, which may promote the research of generating summaries with different granularity.

Abstractiveness. We use novel n-gram ratios between the article and summary for measuring the abstractiveness of the summary. More than 15.96/59.45/81.00/89.68 percent of unique 1/2/3/4-grams have not appeared in the original input. This indicates the summary is highly abstractive. More-

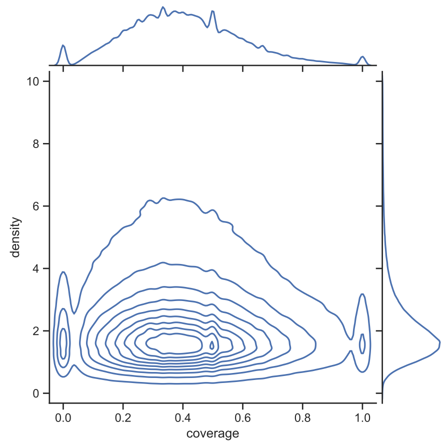

Figure 3: Normalized bi-variate density plot of bi-gram coverage vs. density for 95% of the data.

over, we follow (Grusky et al., 2018) and visualize the bi-variate distribution of bi-gram[6] coverage and density over ***OASum*** in Fig. 3. It shows that ***OASum*** covers a large range of summarization abstractiveness styles in terms of coverage and density.

Aspects. In Table 1, we compare ***OASum*** with other query/aspect-based summarization datasets. ***OASum*** contains a significantly larger amount of aspect types. On average, there are 1.82 aspects per article and 99% articles have less than 9 aspects per single document. As shown in Fig. 4,

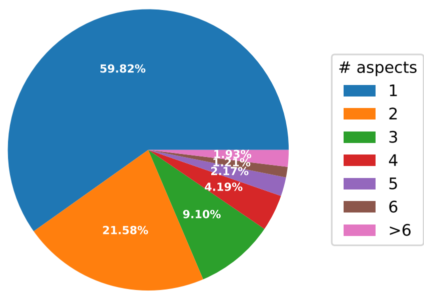

Figure 4: The pie chart for aspects per article.

although 59.82% articles only have one aspect, there are around 40% articles that have multiple aspects ranging from 2 to more than 6. In total,

[4] https://en.wikipedia.org/wiki/Seattle

[5] We filtered out cases in which the summary is longer than the input document in terms of words. However, this compression ratio is calculated based on its tokens.

[6] We explained the reason for using bi-gram coverage and density instead of uni-gram in Appendix A.3

History: Seattle is a seaport city on the West Coast of the United States. It is the seat of King County, Washington. The Seattle area was inhabited by Native Americans for at least 4,000 years before the first permanent European settlers. Arthur A. Denny and his group of travelers, subsequently known as the Denny Party, arrived from Illinois via Portland, Oregon, on the schooner "Exact" at Alki Point on November 13, 1851. The settlement was moved to the eastern shore of Elliott Bay and named "Seattle" in 1852, in honor of Chief Siáhl of the local Duwamish and Suquamish tribes. Growth after World War II was partially due to the local Boeing company, which established Seattle as a center for aircraft manufacturing. The Seattle area developed into a technology center from the 1980s onwards with companies like Microsoft becoming established in the region; Microsoft founder Bill Gates is a Seattleite by birth. The stream of new software, biotechnology, and Internet companies led to an economic revival, which increased the cityś population by almost 50,000 between 1990 and 2000. Seattle also has a significant musical history.
History ; Founding: It is the seat of King County, Washington. The Seattle area was inhabited by Native Americans for at least 4,000 years before the first permanent European settlers. Arthur A. Denny and his group of travelers, subsequently known as the Denny Party, arrived from Illinois via Portland, Oregon, on the schooner "Exact" at Alki Point on November 13, 1851. The settlement was moved to the eastern shore of Elliott Bay and named "Seattle" in 1852, in honor of Chief Siáhl of the local Duwamish and Suquamish tribes.
History ; Post-war years: aircraft and software: Growth after World War II was partially due to the local Boeing company, which established Seattle as a center for aircraft manufacturing. The stream of new software, biotechnology, and Internet companies led to an economic revival, which increased the city's population by almost 50,000 between 1990 and 2000. Seattle also has a significant musical history.
Geography: Seattle is situated on an isthmus between Puget Sound (an inlet of the Pacific Ocean) and Lake Washington.
Economy: A major gateway for trade with East Asia, Seattle is the fourth-largest port in North America in terms of container handling . Internet retailer Amazon was founded in Seattle in 1994, and major airline Alaska Airlines is based in SeaTac, Washington, serving Seattle's international airport, Seattle–Tacoma International Airport.
Culture: Between 1918 and 1951, nearly two dozen jazz nightclubs existed along Jackson Street, from the current Chinatown/International District to the Central District. The jazz scene nurtured the early careers of Ray Charles, Quincy Jones, Ernestine Anderson, and others. Seattle is also the birthplace of rock musician Jimi Hendrix, as well as the origin of the bands Nirvana, Pearl Jam, Soundgarden, Heart, Alice in Chains, Foo Fighters, and the alternative rock movement grunge.
Demographics: Today, Seattle has high populations of Native, Scandinavian, Asian American and African American people, as well as a thriving LGBT community that ranks sixth in the United States by population.

Table 3: Example of aspect-based summaries constructed from "Seattle". We only show part of the aspect summaries.

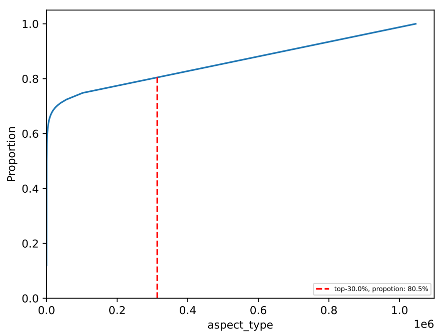

Figure 5: Cumulative proportion of aspect distribution. The horizontal axis represents the sorted aspects from high frequency to low frequency.

Figure 6: Word cloud based on the top 400 categories drawn from the first-level category names in ***OASum***. Word size is proportional to the word count. The size of the dominant category *human* is reduced 10 times in corresponding to the whole category set.

OASum contains 1,045,895 different types of aspects. The top-10 common aspects are *History*, *Career*, *Background*, *Geography*, *Life*, *Reception*, *Description*, *Early life*, *Demographics* and *Production*, containing 447,589, 171,447, 69,266, 45,134, 43,398, 42,664, 36,199, 34,663, 34,057 and 33,424 instances, respectively. As shown in Fig. 5, we find that the top 30% aspect types cover 80.5% of all the cases, while the remaining 19.5% cases come from the other 70% aspects. This naturally provides open-domain and diverse multiple-aspects knowledge for aspect-based summarization.

Ontology. We analyze the domain distribution of our dataset using the ontology information provided by Wikidata's `instance of (P31)` property. In Fig. 6, we show the word cloud of the top 400 first-level categories of Wikipedia pages in ***OASum***. In total, we cover 32,956 out of 45,042 first-level categories among Wikidata, suggesting ***OASum*** contains text information in a large number of different domains. To conclude, ***OASum*** is a large-scale open-domain aspect-based summarization dataset containing varieties of input/output lengths and abstractive summaries with human-verified qualities.

4 Baselines and Analysis

4.1 Metrics and Models

In this section, we investigate the baseline models' performance over ***OASum***. It includes heuristic methods(Heu), unsupervised methods, aspect-agnostic extractive methods(Ext), and aspect-based abstractive methods(Abs). Our results are reported with ROUGE metrics (Lin, 2004), including ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum. We compare our system with extractive and abstractive summarization baselines.

ORACLE is generated by comparing the reference summary and each sentence in the document and obtaining the sentences with the best ROUGE scores in a greedy method (Liu and Lapata, 2019).

RANDOM-N Random sentences are selected for the summary. We choose the same number of sentences in the reference summary.

LEAD-N The leading sentences are known to be a good summary, especially in the news domain. We select the first N sentences as the summary.

SumBasic (Vanderwende et al., 2007) This method takes the frequently occurring words in a document cluster for the summary.

TextRank (Barrios et al., 2016) is a graph-based approach that computes connections between sentence importance based on significant words.

KLSum (Haghighi and Vanderwende, 2009) is a greedy approach to adding a sentence to the summary by minimizing KL divergence.

LEXRANK (Erkan and Radev, 2011) is similar to the *TextRank* but tries to alleviate the redundant information by reranking selected sentences.

Longformer-(base/large) is a supervised extractive method. As ***OASum*** contains long documents, we utilize the Longformer model to efficiently process the long sequence and the sentence-level Transformer layers for the sentence-level interactions. The oracle sentences are used as labels for predicting the best summary sentences.

LED-(base/large)-*OASum*. We adapt LED (Beltagy et al., 2020) for the aspect-based summarization task. We directly format the problem into an end-to-end sequence-to-sequence task and fine-tune the corresponding model over ***OASum***. We prepend the *aspect* to the input document with a *[BOS]* token between them as the sequence input and use the corresponding summary as the sequence output.

4.2 Experiment Settings

We implement our code using pytorch-lightning[7] and Huggingface Transformers[8]. The inputs and outputs are truncated to a maximum of 4096/256 tokens. In Fig. 2, the selected maximum lengths can cover 88.6% of the entire input sequences and 99.1% of the entire output sequences. Since the input length is very long, we can only feed 4 instances to a single GPU for the base model and 2 instances for the large model. For speeding up the training, Distributed Data-Parallel and Automatic Mixed Precision (FP16) are used. Specifically, we utilize 64 NVIDIA V100 GPUs for base models and 128 NVIDIA V100 GPUs for large models for training both aspect-agnostic extractive models and aspect-based abstractive models. The gradient accumulation step is set to 8 for reducing the communication bandwidth. Therefore, the actual batch size is 2048. We use Fused-Adam (Kingma and Ba, 2015) implemented by NVIDIA-apex[9] for the optimization. The initial learning rate is $1e-4$, and it linearly decreases to 0. The betas are 0.9 and 0.999 respectively. We do not apply warm-up for ***OASum*** training. Weight decay is 0.01. We evaluate the model 5 times per epoch on the validation set and pick the checkpoint with the highest average ROUGE-1/2/Lsum scores for testing.

4.3 Results & Analysis

In Table 4, we show the baseline model performance on ***OASum***. The oracle performs the strong baseline and is used for the labels of *Longformer* models. It outperforms all extractive and abstractive methods except for the ROUGE-2 and ROUGE-L of the *LED-large* model. This indicates that the reference summary of ***OASum*** is more abstractive than extractive. The lead sentences perform similarly to the unsupervised baselines meaning that the important information is distributed to the beginning part of the documents but are not necessarily the best sentences as they under-perform the supervised methods. Random selection is the worst choice for the summary. For the supervised models, the extractive method outperforms the unsupervised methods but is outperformed by the abstractive methods by a large margin. We also include some generated good and bad examples as case studies in Appendix C.1.

[7] https://www.pytorchlightning.ai/
[8] https://github.com/huggingface/transformers
[9] https://github.com/NVIDIA/apex

Baselines	Type	Aspect	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
Oracle	Heu	Y	44.97	22.74	32.98	39.17
Random-N	Heu	N	21.03	4.37	14.92	17.45
LEAD-N	Heu	N	23.93	6.02	17.44	19.98
SumBasic	Ext	N	22.79	5.63	16.55	19.26
TexRank	Ext	N	23.09	5.90	15.99	18.62
LexRank	Ext	N	23.95	6.00	16.81	19.64
KLSum	Ext	N	22.80	5.59	15.81	18.40
Longformer-base (4K)	Ext	N	30.06	10.75	22.08	25.35
Longformer-large (4K)	Ext	N	30.76	11.21	22.23	25.78
LED-base (4K)	Abs	Y	37.26	20.84	31.97	33.71
LED-large (4K)	Abs	Y	**39.61**	**22.17**	**33.34**	**35.46**

Table 4: Baseline results on ***OASum*** test set. Y and N mean including aspect or not.

Datasets	Models	R-1	R-2	R-Lsum
AQuaMuse	*L*	49.34	33.26	46.42
	O	**49.98**	**34.12**	**47.09**
CovidET	*L*	**26.19**	**6.85**	**20.82**
	O	25.61	6.58	20.45
MA-News	*L*	37.8	**17.43**	35.3
	O	**38.12**	17.41	**35.51**
NEWTS	*L*	31.96	10.75	28.72
	O	**32.45**	**11.64**	**29.14**
QMSum	*L*	29.52	7.00	25.68
	O	**30.30**	**7.56**	**26.67**
SQuaLITY	*L*	36.78	8.31	34.47
	O	**37.6**	**8.81**	**35.14**
Wikiasp	*L*	22.18	8.21	20.48
	O	**22.69**	**8.29**	**20.92**

Table 5: Fine-tuning results on downstream tasks. Wikiasp results are the average number of all 20 domains. *L* represents LED-base, *O* represents LED-OASum-base.

5 Downstreams

To verify the knowledge inside ***OASum*** provides transfer ability, we further use the model pre-trained on ***OASum*** for seven abstractive downstream datasets (see Appendix B.1) including three query-based summarization datasets and four aspect-based summarization datasets, across different domains. We test our model with zero-shot, few-shot, and fine-tuning abilities on these 7 datasets to see whether ***OASum*** can benefit the downstream tasks. In general, the model pre-trained on the ***OASum*** outperforms the backbone model on 6 out of 7 tasks in the fine-tuning and zero-shot setting, 6 out of 6 tasks (w/o WikiAsp) in the few-shot setting.[10]

[10]WikiAsp has 20 different subsets on different domains, we only perform the results for zero-shot and fine-tuning setting.

5.1 Experiment Settings

For all downstream tasks, we only test the base model to demonstrate the ability of our pretrained checkpoint in an end-to-end setting. We experiment with different decoding hyper-parameters and find the $length\_penalty = 1.0$, $num\_beams = 4$, and $no\_repeat\_ngram_size = 3$ consistently achieve optimal performance on multiple datasets in the zero-shot setting. Thus, we keep these parameters for all downstream task experiments. For the backbone **LED-base** model (denoted as **L**), we initialize the model using the checkpoint provided by (Beltagy et al., 2020) on Huggingface[11]. On top of the backbone model, our model checkpoint is further fine-tuned on **LED-*OASum*** (denoted as **O**) for 20 epochs. Notice that for fine-tuning and zero-shot scenarios, the Wikiasp results are reported on an average of 20 domains tested independently.

5.2 Fine-tuning Settings

For fine-tuning experiments, we directly fine-tune the model on the whole training set and report the ROUGE scores on the test set by selecting the best-performing checkpoint on the validation set. We present all the fine-tuning results in Table 5 with ROUGE-1, ROUGE-2, and ROUGE-Lsum scores. In general, models fine-tuned on our checkpoint consistently perform better and demonstrate a strong advantage in ROUGE scores. Appendix B.3 shows the complete results of the 20 domains of Wikiasp. We find that our fine-tuned models outperform the backbone model on most of the domains, with only a few exceptions. Overall, our experiments demonstrate that fine-tuning the backbone model on ***OASum*** is an effective approach for improving performance on a variety of aspects or query-based summarization tasks.

[11]https://huggingface.co/allenai/led-base-16384

Datasets	Models	Few-shot 0.3%			Few-shot 1%			Few-shot 3%		
		R-1	R-2	R-Lsum	R-1	R-2	R-Lsum	R-1	R-2	R-Lsum
AQuaMuse	*L*	$32.44_{\pm0.91}$	$13.89_{\pm0.95}$	$29.41_{\pm0.91}$	$35.29_{\pm1.39}$	$16.91_{\pm1.30}$	$32.16_{\pm1.17}$	$37.55_{\pm0.57}$	$19.88_{\pm0.83}$	$34.40_{\pm0.71}$
	O	$38.77^{\uparrow6.33}_{\pm0.53}$	$20.61^{\uparrow6.72}_{\pm0.80}$	$35.61^{\uparrow6.20}_{\pm0.70}$	$40.63^{\uparrow5.34}_{\pm0.20}$	$22.81^{\uparrow5.90}_{\pm0.80}$	$37.50^{\uparrow5.34}_{\pm0.39}$	$41.50^{\uparrow3.95}_{\pm1.04}$	$24.25^{\uparrow4.37}_{\pm0.92}$	$38.55^{\uparrow4.15}_{\pm1.01}$
CovidET	*L*	$20.33_{\pm0.01}$	$3.75_{\pm0.17}$	$16.53_{\pm0.15}$	$21.19_{\pm0.30}$	$4.40_{\pm0.02}$	$17.39_{\pm0.10}$	$22.56_{\pm0.18}$	$5.07_{\pm0.07}$	$18.37_{\pm0.15}$
	O	$22.00^{\uparrow1.67}_{\pm0.12}$	$4.58^{\uparrow0.83}_{\pm0.14}$	$17.90^{\uparrow1.37}_{\pm0.12}$	$22.16^{\uparrow0.97}_{\pm0.02}$	$4.58^{\uparrow0.18}_{\pm0.02}$	$18.02^{\uparrow0.64}_{\pm0.03}$	$22.73^{\uparrow0.17}_{\pm0.34}$	$5.02^{\downarrow0.05}_{\pm0.15}$	$18.40^{\uparrow0.03}_{\pm0.24}$
MA-News *	*L*	$20.12_{\pm0.03}$	$5.08_{\pm0.01}$	$18.52_{\pm0.03}$	$20.41_{\pm0.05}$	$5.40_{\pm0.02}$	$18.84_{\pm0.03}$	$22.07_{\pm0.02}$	$6.63_{\pm0.02}$	$20.41_{\pm0.01}$
	O	$24.15^{\uparrow4.03}_{\pm0.17}$	$7.37^{\uparrow2.28}_{\pm0.05}$	$22.22^{\uparrow3.69}_{\pm0.13}$	$25.12^{\uparrow4.71}_{\pm0.01}$	$7.98^{\uparrow2.58}_{\pm0.01}$	$23.11^{\uparrow4.27}_{\pm0.01}$	$27.58^{\uparrow5.51}_{\pm0.08}$	$9.67^{\uparrow3.03}_{\pm0.02}$	$25.49^{\uparrow5.08}_{\pm0.07}$
NEWTS	*L*	$26.24_{\pm0.03}$	$7.35_{\pm0.11}$	$23.47_{\pm0.05}$	$26.77_{\pm0.53}$	$8.16_{\pm0.23}$	$24.63_{\pm0.66}$	$27.92_{\pm0.02}$	$8.47_{\pm0.28}$	$25.06_{\pm0.04}$
	O	$27.75^{\uparrow1.50}_{\pm0.49}$	$8.10^{\uparrow0.75}_{\pm0.02}$	$24.77^{\uparrow1.30}_{\pm0.40}$	$28.59^{\uparrow1.82}_{\pm0.09}$	$8.66^{\uparrow0.50}_{\pm0.03}$	$25.50^{\uparrow0.88}_{\pm0.06}$	$28.15^{\uparrow0.24}_{\pm0.67}$	$8.80^{\uparrow0.33}_{\pm0.02}$	$25.27^{\uparrow0.20}_{\pm0.54}$
QMSum	*L*	$19.80_{\pm0.63}$	$3.23_{\pm0.01}$	$17.28_{\pm0.14}$	$22.58_{\pm3.58}$	$3.80_{\pm0.31}$	$19.70_{\pm2.08}$	$24.52_{\pm0.70}$	$4.64_{\pm0.38}$	$21.39_{\pm0.50}$
	O	$22.98^{\uparrow3.17}_{\pm2.06}$	$4.40^{\uparrow1.17}_{\pm0.29}$	$19.88^{\uparrow2.60}_{\pm0.87}$	$24.51^{\uparrow1.93}_{\pm1.88}$	$4.53^{\uparrow0.73}_{\pm0.26}$	$21.38^{\uparrow1.68}_{\pm1.06}$	$25.48^{\uparrow0.96}_{\pm0.01}$	$5.30^{\uparrow0.67}_{\pm0.16}$	$22.21^{\uparrow0.82}_{\pm0.05}$
SQuaLITY	*L*	$26.27_{\pm0.68}$	$4.39_{\pm0.01}$	$24.72_{\pm0.69}$	$26.79_{\pm1.24}$	$4.58_{\pm0.14}$	$25.27_{\pm1.16}$	$31.52_{\pm0.66}$	$5.79_{\pm0.22}$	$29.55_{\pm0.63}$
	O	$29.05^{\uparrow2.78}_{\pm0.23}$	$5.19^{\uparrow0.80}_{\pm0.05}$	$27.18^{\uparrow2.45}_{\pm0.28}$	$30.72^{\uparrow3.93}_{\pm0.20}$	$5.75^{\uparrow1.17}_{\pm0.06}$	$28.80^{\uparrow3.53}_{\pm0.20}$	$33.05^{\uparrow1.53}_{\pm0.68}$	$6.71^{\uparrow0.92}_{\pm0.01}$	$31.04^{\uparrow1.49}_{\pm0.63}$

Table 6: Few-shot performance. MA-News results are under 0.03%, 0.1%, and 0.3%. *L* represents LED-base, *O* represents LED-OASum-base.

Datasets	Models	R-1	R-2	R-Lsum
AQuaMuse	*L*	24.98	9.22	22.93
	O	**36.80**	**18.18**	**33.50**
CovidET	*L*	14.61	**3.08**	12.37
	O	**15.75**	2.01	**12.72**
MA-News	*L*	17.01	5.56	15.82
	O	**20.06**	**5.81**	**18.41**
NEWTS	*L*	**26.71**	**8.49**	**22.14**
	O	24.06	6.91	21.04
QMSum	*L*	13.96	2.29	12.70
	O	**22.51**	**3.27**	**19.95**
SQuaLITY	*L*	26.87	3.69	25.41
	O	**30.54**	**5.72**	**28.86**
Wikiasp	*L*	8.90	1.06	8.04
	O	**15.61**	**2.75**	**13.91**

Table 7: Zero-shot results on downstream tasks. Wikiasp results are the average number on all 20 domains. *L* represents LED-base, *O* represents LED-OASum-base.

5.3 Few-Shot Settings

For few-shot experiments, we randomly pick 0.3%, 1%, and 3% of the training data, then perform 60-epoch training on the picked low-resource samples. To compensate randomness, we conduct all experiments for three times using different random seeds to pick the training data. The results are reported based on the average and variance of ROUGE scores over experiments with different random seeds. Table 6 includes the few-shot performance of the backbone model and our model. An obvious superiority is demonstrated based on our checkpoint models for almost all R-1, R-2, and R-Lsum scores under 0.3%, 1%, and 3% settings for every aspect or query-based summarization dataset. For all the datasets, we can achieve substantial advancements of around 1 to 7 points improvements under ROUGE evaluation. Besides, the greatest improvements almost always happen in extremely low-resource(0.3%) scenarios, demonstrating the great adaptability of our model for various domains. Given the difficulty of gathering such data, we think our findings are beneficial across many disciplines. In Table 16, we also show some typical examples.

5.4 Zero-shot Settings

For the zero-shot experiments, we only test the models on the whole test set without any optimization of the training data. The zero-shot evaluation results are demonstrated in Table 7. The complete results on 20 domains of Wikiasp are also shown in Table 13. As we can see, except for NEWTS datasets, our LED-*OASum* consistently achieves significantly better results in almost all evaluation metrics. We believe this improvement comes from the rich knowledge contained in the large corpus learned during the pre-training. The performance almost doubles on Wikiasp and AQuaMuse, validating that the knowledge is successfully transferred into the generation process. More case studies can be found in Table 15 and Table 16.

6 Conclusions

In summary, we contribute the first large-scale open-domain aspect-based summarization corpus collected using Wikipedia section titles as aspects by rules with good quality. Detailed statistics reveal many different aspects of the corpus, confirming its broader coverage. We also outline the methods we use for pre-training the generative language models and present abstractive and extractive results as a baseline for future work. Furthermore, we prove that our pre-trained model can consistently improve seven widely-used downstream tasks, especially in few-shot and zero-shot settings. We hope our data and pre-trained models can further foster relevant research in this area.

7 Limitations

First of all, our *OASum* inevitably contains inappropriate summaries not strongly correlated with certain aspects since it is automatically curated. The model trained on it could furthermore hold such misinformation and affect other downstream tasks. But we hope the large-scale training can alleviate such effects to a minimum. At the current stage, we are not responsible for any products directly built on our results. In the future, a potential denoising mechanism could be designed to further reduce the noisy summaries.

Secondly, we only opt for end-to-end extraction, which requires large computational memory and cost that may not be afforded by everyone. Thus, a meaningful direction would be investigating other extract-then-summarize two-step methods for dealing with long document summarization. Besides, our vanilla dataset contains millions of summaries that are difficult for certain researchers with limited computational resources to directly reproduce results on. We recommend using a small subset of our corpus if enough computational capability is not immediately available.

Finally, we only explore a simple strategy for controlling the summarization based on input aspects. However, we find it can not always guarantee aspect-focused generation. How to efficiently and accurately generate specific summaries by confining aspects is not only challenging for model design but also difficult for humans to evaluate. We leave these issues for future work.

8 Acknowledgments

This work was done when Xianjun Yang was doing an internship at Tencent AI Lab Seattle. Xianjun Yang was supported in part by the UC Santa Barbara NSF Quantum Foundry funded via the Q-AMASEi program under NSF award DMR1906325.

References

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. ASPECTNEWS: Aspect-oriented summarization of news documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6494–6506, Dublin, Ireland. Association for Computational Linguistics.

Nadeem Akhtar, Nashez Zubair, Abhishek Kumar, and Tameem Ahmad. 2017. Aspect based sentiment oriented summarization of hotel reviews. *Procedia computer science*, 115:563–571.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. 2022a. Newts: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503.

Seyed Ali Bahrainian, Sheridan Feucht, and Carsten Eickhoff. 2022b. NEWTS: A corpus for news topic-focused summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 493–503, Dublin, Ireland. Association for Computational Linguistics.

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. *CoRR*, abs/1602.03606.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Sangwoo Cho, Kaiqiang Song, Chen Li, Dong Yu, Hassan Foroosh, and Fei Liu. 2020. Better highlighting: Creating sub-sentence summary highlights. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6282–6300, Online. Association for Computational Linguistics.

Sangwoo Cho, Kaiqiang Song, Xiaoyang Wang, Fei Liu, and Dong Yu. 2022. Toward unifying text segmentation and long document summarization.

Maximin Coavoux, Hady Elsahar, and Matthias Gallé. 2019. Unsupervised aspect-based multi-document abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Günes Erkan and Dragomir R. Radev. 2011. Lexrank: Graph-based lexical centrality as salience in text summarization. *CoRR*, abs/1109.2128.

Lea Fremann and Alexandre Klementiev. 2019. Inducing document structure for aspect-based summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6263–6273, Florence, Italy. Association for Computational Linguistics.

Demian Gholipour Ghalandari, Chris Hokamp, John Glover, Georgiana Ifrim, et al. 2020. A large-scale multi-document summarization dataset from the wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308.

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.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Boulder, Colorado. Association for Computational Linguistics.

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. Wikiasp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Robert Iv, Alexandre Passos, Sameer Singh, and Ming-Wei Chang. 2022. FRUIT: Faithfully reflecting updated information in text. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3670–3686, Seattle, United States. Association for Computational Linguistics.

Hitesh Kansal and Durga Toshniwal. 2014. Aspect based summarization of context dependent opinion words. *Procedia Computer Science*, 35:166–175.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

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

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *arXiv preprint arXiv:2010.12694*.

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.

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

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. Brio: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903.

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

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.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Jonathan Pilault, Raymond Li, Sandeep Subramanian, and Chris Pal. 2020. On extractive and abstractive neural document summarization with transformer language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9308–9319, Online. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

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.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *arXiv preprint arXiv:2206.10883*.

Kaiqiang Song, Chen Li, Xiaoyang Wang, Dong Yu, and Fei Liu. 2020. Automatic summarization of open-domain podcast episodes. In *Proceedings of the 29th Text REtrieval Conference (TREC)*.

Kaiqiang Song, Chen Li, Xiaoyang Wang, Dong Yu, and Fei Liu. 2022. Towards abstractive grounded summarization of podcast transcripts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4407–4418, Dublin, Ireland. Association for Computational Linguistics.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020. Summarizing text on any aspects: A knowledge-informed weakly-supervised approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6301–6309, Online. Association for Computational Linguistics.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Inf. Process. Manag.*, 43(6):1606–1618.

Jesse Vig, Alexander Fabbri, Wojciech Kryscinski, Chien-Sheng Wu, and Wenhao Liu. 2022. Exploring neural models for query-focused summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1455–1468, Seattle, United States. Association for Computational Linguistics.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R Bowman. 2022a. Squality: Building a long-document summarization dataset the hard way. *arXiv preprint arXiv:2205.11465*.

Fei Wang, Kaiqiang Song, Hongming Zhang, Lifeng Jin, Sangwoo Cho, Wenlin Yao, Xiaoyang Wang, Muhao Chen, and Dong Yu. 2022b. Salience allocation as guidance for abstractive summarization.

Kristian Woodsend and Mirella Lapata. 2012. Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 233–243, Jeju Island, Korea. Association for Computational Linguistics.

Fei Wu and Daniel S Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 118–127.

Haibing Wu, Yiwei Gu, Shangdi Sun, and Xiaodong Gu. 2016. Aspect-based opinion summarization with convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3157–3163. IEEE.

Yumo Xu and Mirella Lapata. 2020. Coarse-to-fine query focused multi-document summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3632–3645, Online. Association for Computational Linguistics.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Hongli Zhan, Tiberiu Sosea, Cornelia Caragea, and Junyi Jessy Li. 2022. Why do you feel this way? summarizing triggers of emotions in social media posts. *arXiv preprint arXiv:2210.12531*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Chujie Zheng, Kunpeng Zhang, Harry Jiannan Wang, and Ling Fan. 2020. A two-phase approach for abstractive podcast summarization. *arXiv preprint arXiv:2011.08291*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

A Details in Data Statistics

A.1 Top 50 Aspects

In Table 8, We show the most common 50 aspects in ***OASum*** and their frequencies. As we can see, those aspects naturally cover many perspectives of an article, serving as good and diverse aspects to be summarized with.

A.2 Top 50 Categories

In Table 9, We show the most common 50 categories of Wikipedia pages in ***OASum*** and their frequencies. In general, the top-50 and top-10% categories take up around 57.84%, and 93.51% of all the categories, respectively.

A.3 Bi-gram coverage and density

We notice that uni-gram coverage and density presented in the (Grusky et al., 2018) could only represent the token level extractiveness. However, summarizers typically extract self-contained (Cho et al., 2020) text spans to construct a summary. It usually works on sentence-level or sub-sentence level. In such cases, the token-level extractiveness cannot well represent how extractiveness the instance is. It becomes worse when the input document is long enough, containing different pieces of summary tokens in different places of the document. On the country, bi-gram coverage and density reduce the chance of wrongly representing the extractiveness of the instances. Thus, in this work, we choose to use bi-gram coverage and density for presenting the extractiveness / abstractiveness of instances.

B Details in Experiments

B.1 Datasets

We list the 7 downstream datasets below, their statistics are shown in Table 1:

AQuaMuse (Kulkarni et al., 2020) is a Query-based multi-document summarization(qMDS) dataset built by automatically mining qMDS examples from question-answering datasets and large document corpora. We follow the preprocessing steps in (Vig et al., 2022) to build the AQuaMuse based on Version 3 and get a train/validation/test split of 5,784/637/747. For multiple documents, we directly concatenate them together as inputs in a natural order.

CovidET (Zhan et al., 2022) includes abstractive summaries of seven emotion triggers related to COVID-19 Reddit posts written by humans. Following their public repository[12], we successfully build 4,419/1,077/1,616 instances for train/validation/test. Notice that in their dataset, one instance may have several different reference summaries. We follow their evaluation considering the average ROUGE scores if multiple references exist.

[12] https://github.com/honglizhan/CovidET

Aspect	Count	Aspect	Count	Aspect	Count
History	447,589	*Career*	171,447	*Background*	69,266
Geography	45,134	*Life*	43,398	*Reception*	42,664
Description	36,199	*Early life*	34,663	*Demographics*	34,057
Production	33,424	*Plot*	32,331	*Overview*	23,465
Professional career	21,237	*Political career*	20,232	*Club career*	18,520
Release	17,867	*Playing career*	17,735	*Life and career*	17,672
Personal life	15,786	*Development*	14,253	*Early life and education*	13,056
Critical reception	12,889	*Track listing*	11,176	*Route description*	10,065
Legacy	9,982	*International career*	9,582	*Gameplay*	8,533
Location	8,379	*Coaching career*	7,860	*Aftermath*	7,672
Taxonomy	7,570	*College career*	7,302	*Synopsis*	7,063
Design	6,651	*Demographics ; 2010 census*	6,541	*Education*	6,461
Distribution and habitat	6,344	*Early life and career*	6,328	*Description and history*	5,711
Death	5,709	*Early years*	5,664	*Awards*	5,657
Structure	5,541	*Composition*	5,535	*Music video*	5,513
Politics	5,191	*Function*	5,061	*Distribution*	5,034
Origins	4,942	*Publication history*	4,809		

Table 8: The most common 50 aspects and their frequencies.

Wikidata ID	Category	Count	Wikidata ID	Category	Count
Q5	*human*	572,975	Q11424	*film*	45,427
Q16521	*taxon*	40,182	Q482994	*album*	35,055
Q4830453	*business*	33,406	Q134556	*single*	21,768
Q215380	*musical group*	20,615	Q27020041	*sports season*	18,045
Q13406463	*Wikimedia list article*	17,150	Q7889	*video game*	14,776
Q486972	*human settlement*	14,744	Q7725634	*literary work*	14,710
Q5398426	*television series*	13,969	Q34442	*road*	13,224
Q43229	*organization*	13,085	Q7366	*song*	12,516
Q55488	*railway station*	11,662	Q476028	*association.*	10,652
Q14350	*radio station*	10,120	Q532	*village*	9,794
Q9826	*high school*	9,212	Q11446	*ship*	9,060
Q1093829	*city.*	8,403	Q16970	*church building*	8,376
Q176799	*military unit*	7,845	Q47461344	*written work*	6,917
Q21191270	*television.*	6,870	Q41176	*building*	6,543
Q4022	*river*	6,502	Q498162	*census.*	6,402
Q3918	*university*	5,833	Q3914	*school*	5,769
Q15127012	*town.*	5,674	Q3957	*town*	5,576
Q6881511	*enterprise*	5,542	Q15632617	*fictional human*	5,502
Q11173	*chemical compound*	5,404	Q7278	*political party*	5,284
Q178561	*battle*	5,159	Q891723	*public company*	4,916
Q17343829	*unincorporated.*	4,792	Q1115575	*civil parish*	4,672
Q163740	*nonprofit organization*	4,418	Q123705	*neighborhood*	4,413
Q515	*city*	3,900	Q15416	*television program*	3,864
Q3231690	*automobile model*	3,811	Q41710	*ethnic group*	3,747
Q7187	*gene*	3,724	Q74817647	*aspect.*	3,719

Table 9: The most common 50 categories, the corresponding Wikidata IDs, and their frequencies. *unincorporated.*, *aspect.*, *city.*, *town.*, *association.*, *television.* and *census.* are short for *unincorporated community in the United States*, *aspect in a geographic region*, *city/town of the United States*, *association football club*, *television series episode*, *census-designated place*, respectively.

MA-News (Frermann and Klementiev, 2019) synthesize multi-aspect summaries by interleaving paragraphs of n_d documents belonging to different aspects and pairing the document with each of its n_d components' reference summaries. It includes 284,700/14,589/12,800 train/validation/test summarization pairs.

NEWTS (Bahrainian et al., 2022b) contains 4800 training and 1,200 testing aspect-based abstractive summaries annotated by humans derived from the well-known CNN/Dailymail (Hermann et al., 2015; Nallapati et al., 2016) dataset. Each article contains two general aspects, such as economics and politics. We randomly split the original 1,200 testing samples into 300 instances for validation and 900 for the test.

QMSum (Zhong et al., 2021) select and summarize relevant spans of meetings in response to a specific query. It contains 1,257, 272, and 279 training, validation, and test instances, respectively. The query is usually a general question such as *summarize the whole meeting* . or a specific query like *how did marketing design the product evaluation ?*.

SQuALITY (Wang et al., 2022a) is a dataset of question-focused long-document summaries built on the public-domain short stories by hiring highly-qualified contractors to read stories and write original summaries from scratch. Documents are an average of 5,199.4 tokens long, while responses and plot summaries are 237.1, and 441.9 tokens long on average, respectively.

Wikiasp (Hayashi et al., 2021) provides multi-domain aspect-based summarization by using the section titles and boundaries of each Wikipedia article for aspect annotation and all available references as source with an average length of 13,672. It contains 20 different domains and 200 aspects, we present the averaged results on all 20 domains.

B.2 Hyper-parameters

Fine-tuning. For downstream tasks, we fine-tune the model with 20 epochs on WikiAsp and 50 epochs on the remaining datasets. We then pick the checkpoint with the best validation average ROUGE performance to test its final performance on the testing data. In Table 10, we show the hyper-parameters used in the fine-tuning setting of different datasets. For decoding, we keep no_repeat_ngrams as 3, the beam size is set to 4, and the length penalty is set to 1.0. We use a linearly decreasing learning rate schedule on all tasks without any warm-up. The weight decay is set to 0.01.

Dataset	#Mai	#Mio	#Mao	Bs	lr
AQuaMuse	16,384	64	256	32	5e−5
CovidET	512	25	256	32	5e−5
MA-News	2,048	64	256	64	5e−5
NEWTS	2,048	25	256	32	5e−5
QMSum	16,384	30	256	32	2e−5
SQuaLITY	16,384	256	512	32	1e−4
Wikiasp	16,384	10	256	32	1e−5

Table 10: Finetuning hyper-paramters parameters. #Mai, #Mio, #Mao, Bs and lr represent Max input length, Min output length, Max output length, batch size and learning rate, respectively.

Dataset	#Mai	#Mio	#Mao
AQuaMuse	16,384	64	256
CovidET	512	25	256
MA-News	2,048	64	256
NEWTS	2,048	25	256
QMSum	16,384	25	256
SQuaLITY	16,384	256	512
Wikiasp	16,384	128	256

Table 11: Zero-shot hyper-parameters parameters. #Mai, #Mio and #Mao represent Max input length, Min output length and Max output length, respectively.

Dataset	0.3 %	1%	3%
AQuaMuse	17	57	173
CovidET	13	44	132
*MA-News**	85	285	854
NEWTS	14	48	144
QMSum	3	12	137
SQuaLITY	3	10	30

Table 12: Number of training instances under different few-shot settings. MA-News results are under 0.03%, 0.1%, and 0.3%.

Zero/Few-shot. In Table 11, we show the hyper-parameters used for zero/few-shot settings where no_repeat_ngrams is kept at 3/0, the beam size is 4/1, and the length penalty is always set to 1.0. We only use early_stopping for zero-shot. The epochs and learning rate for few-shot training are always 60 and 2e-5 respectively. warm-up rates are set to 0.05, while weight decay is 0.01. Batch size is 2 for 0.3%, while 4 for 1% and 3% scenarios. In Table 12, we also show the exact number of instances used for few-shot training. The total number of picked training instances ranges from less than ten to several hundred.

B.3 WikiAsp Full Results

In Table 13, we present all WikiAsp (Hayashi et al., 2021) 20 domains results with fine-tuning and zero-shot settings. It is obvious that our **LED-*OASum*** consistently achieves near-double performance for all domains under almost all ROUGE metrics. The improvements over finetuning results are less substantial but still preserve more than 0.5 points improvements. We attribute this advance comes from the rich knowledge contained in our *OASum* corpus. It is worth noting that the inputs of our *OASum* are close to the outputs of Wikiasp, but we are not sure whether the information seen during our training in encoding has direct help for tuning wikiasp in the decoding stage.

C Case Study

C.1 *OASum* Examples

Here we show two examples of Wikipedia pages *Pokémon*[13] and *Shanghai*[14] from *OASum* test set in Table 14. The aspect-based summary results are generated by our trained LED-*OASum* checkpoint. It is clear that 4 aspects for *Pokémon* and 7 aspects for *Shanghai* indeed produce strongly relevant and coherent aspect-based summaries. But it still fails for generating correct summaries for aspect *Cultural influence* and *Demographics* highlighted in red. We attribute such errors to coming from two perspectives: the model fails to focus on a certain aspect or it can not generate correct summaries. For example, for *Cultural influence* in *Pokémon*, the generated summary is coherent, fluent, and "correct", but not related to this specific aspect at all. For *Demographics* in *Shanghai*, the first half of the summary is focused on *Demographics*, but the remaining description *the capital of the province of Zhejiang* is both unrelated and inaccurate.

C.2 LED-*OASum* Examples

Zero-shot. Here we show three examples from downstream AQuaMuse, QMSum, and NEWTS datasets under the zero-shot setting in Table 15. As we can see from the results of AQuaMuse and QMSum, LED-*OASum* can produce much better summaries. For another example from NEWTS, although LED-base achieves higher rouge scores, the summary is actually redundant and repetitive. On the contrary, the LED-*OASum* generated summary(highlighted in green) preserves the summary towards the chosen aspect and demonstrates good quality.

Few-shot. Besides, we also show one example from SQuALITY dataset under few/zero-shot setting in Table 16. Under the zero-shot conditions, our LED-*OASum* can generate a much better query-based summary than the original LED-base model, which can also be observed from ROUGE scores. When the models are furthermore tuned on a small amount of 3% (30) of training instances, the improvements mainly come from ROUGE-L and ROUGE-LSum.

[13] https://en.wikipedia.org/wiki/Pok%C3%A9mon

[14] https://en.wikipedia.org/wiki/Shanghai

Domain	Models	Finetune			Zero-shot		
		R-1	R-2	R-Lsum	R-1	R-2	R-Lsum
Album	LED	19.02	7.56	17.28	7.64	0.79	6.83
	LED-OASum	**19.83**	**7.72**	**18.04**	**15.01**	**2.33**	**13.17**
Animal	LED	23.16	9.19	21.52	6.83	0.73	6.17
	LED-OASum	**24.16**	**9.44**	**22.41**	**12.97**	**2.07**	**11.58**
Artist	LED	21.12	6.76	19.42	7.56	0.89	6.91
	LED-OASum	**21.52**	**6.77**	**19.78**	**14.31**	**2.17**	**12.81**
Building	LED	22.94	**7.19**	21.30	13.27	1.88	12.09
	LED-OASum	**23.18**	7.16	**21.49**	**19.75**	**3.90**	**17.73**
Company	LED	18.44	4.97	16.87	9.26	1.09	8.22
	LED-OASum	**19.12**	**5.07**	**17.50**	**15.77**	**2.66**	**13.99**
EducationalInstitution	LED	21.12	7.46	18.96	8.72	1.17	7.66
	LED-OASum	**21.37**	**7.85**	**19.22**	**15.96**	**3.05**	**13.98**
Event	LED	19.33	5.56	17.57	10.90	1.20	10.02
	LED-OASum	**20.62**	**5.90**	**18.80**	**17.35**	**3.19**	**15.65**
Film	LED	19.53	6.77	17.85	7.45	0.78	6.84
	LED-OASum	**20.09**	**6.98**	**18.38**	**15.21**	**2.80**	**13.68**
Group	LED	**18.25**	**5.21**	**16.89**	8.23	1.13	7.50
	LED-OASum	18.22	4.96	16.74	**12.84**	**2.27**	**11.56**
HistoricPlace	LED	**27.49**	**9.96**	**26.02**	13.53	1.77	12.44
	LED-OASum	26.96	9.64	25.39	**19.34**	**3.59**	**17.70**
Infrastructure	LED	23.68	9.72	21.96	9.49	1.06	8.44
	LED-OASum	**23.96**	**9.75**	**22.15**	**14.97**	**2.34**	**13.08**
MeanOfTransportation	LED	22.74	7.79	20.98	9.49	1.06	8.44
	LED-OASum	**23.96**	**8.28**	**22.11**	**14.97**	**2.34**	**13.08**
OfficeHolder	LED	24.13	9.30	22.18	8.65	1.24	7.77
	LED-OASum	**24.58**	**9.33**	**22.61**	**16.41**	**3.24**	**14.42**
Plant	LED	23.87	8.13	22.17	7.57	0.82	6.83
	LED-OASum	**24.59**	**8.15**	**22.70**	**14.61**	**2.22**	**13.10**
Single	LED	20.66	**8.04**	18.95	8.39	0.99	7.52
	LED-OASum	**21.63**	7.98	**19.68**	**16.83**	**3.79**	**14.85**
SoccerPlayer	LED	16.93	5.52	15.10	6.79	0.83	5.98
	LED-OASum	**17.73**	**5.88**	**15.89**	**10.3**	**1.49**	**8.80**
Software	LED	20.29	5.70	18.55	9.32	0.97	8.35
	LED-OASum	**20.67**	**5.82**	**18.95**	**17.99**	**3.39**	**16.18**
TelevisionShow	LED	17.53	**4.77**	15.75	7.65	0.83	6.93
	LED-OASum	**17.53**	4.68	**15.83**	**15.02**	**2.45**	**13.29**
Town	LED	43.34	**29.37**	42.33	7.86	0.85	7.27
	LED-OASum	**43.52**	29.20	**42.34**	**14.35**	**2.66**	**13.00**
WrittenWork	LED	19.98	5.16	18.04	9.06	0.86	8.33
	LED-OASum	**20.45**	**5.22**	**18.48**	**17.14**	**2.72**	**15.27**

Table 13: Finetuing and zero-shot performance on Wikiasp datasets.

Title	Aspect	Summary
Pokémon	*Name*	"Pokémon" is a media franchise based on the "Pokémon" video game series created by Satoshi Tajiri and published by Nintendo.
	Concept	Pokémon is a media franchise based on the "Pokémon" video game franchise created by Satoshi Tajiri and published by Nintendo.
	Criticism and controversy	"Pokémon" has been criticized by some fundamentalist Christians over perceived occult and violent themes and the concept of "Pokémon evolution", which they feel goes against the Biblical creation account.
	In other media	"Pokémon" has also been the subject of numerous manga series, anime, and video games, as well as the "Pokémon Trading Card Game", a collectible card game.
	Cultural influence	Pokémon is a media franchise based on the video game series of the same name, created by Satoshi Tajiri and published by Nintendo.
Shanghai	*History ; Imperialism*	Shanghai was founded in 746 as a trading port on the Yangtze River, and became an important sea port during the Ming and Qing dynasties. In the 19th century, international attention to Shanghai grew due to European recognition of its economic and trade potential at the river. In 1842, the Treaty of Nanking opened Shanghai as one of the five treaty ports for international trade.
	History ; Japanese invasion	In 1937, the city was occupied by the Japanese during the Battle of Shanghai, which resulted in the occupation of the city by the occupying forces.
	Politics ; Administrative divisions	Shanghai is the capital and largest city of the People's Republic of China. Shanghai is one of the four municipalities under the direct administration of the Government of China, and is divided into 16 county-level districts.
	Economy	Shanghai is a global financial center, ranking first in the whole of Asia & Oceania region and third globally (after New York and London) in the 28th edition of the Global Financial Centres Index, published in September 2020 by ZYen and China Development Institute.
	Education and research	Shanghai is home to many universities and colleges, including Fudan University, Shanghai Jiao Tong University, Tongji University, East China University of Science and Technology, Donghua University, and Shanghai International Studies University.
	Geography	Shanghai is located on the east coast of the Yangtze River estuary, which separates it from the provinces of Zhejiang to the south and Jiangsu to the west and north.
	Culture ; Arts	Shanghai is also known for its art, culture, and architecture, and is home to many museums, including the Shanghai Conservatory of Music, the Shanghai Museum, and the Shanghai Art Museum.
	Demographics	Shanghai is the second-largest city in mainland China and the capital of the province of Zhejiang.

Table 14: Case study: finetuning results on our *OASum* test set.

Query in AQuaMuse test set: *what do you mean by analysis of variance*
LED-*OASum*: The analysis of variance (ANOVA) is a collection of statistical tests used to determine whether the variance of a variable in a particular variable is partitioned into components attributable to different sources of variation. For this reason, ANOVA is useful in comparing two, three, or more means. For the purpose of ANOVA, the observed variance in a specific variable (or group of means) is divided by means to determine the probability that the means of several groups are equal, and therefore generalizes the t-test to more than two groups. 'rouge1': 64.35, 'rouge2': 42.11, 'rougeL': 42.61, 'rougeLsum': 61.74
LED-base: This is a summary of the results of the ANOVA. The results of ANOVA are summarized in the following table. This is the summary of all the results from ANOVA: The results are summarized as follows:The ANOVA is a collection of statistical methods used to determine the degree of variance in a given variable. The result is summarized in terms of the number of variance and the types of variance that can be determined. The examples are:The results of AnOVA are summarised as follows. The analysis of variance is summarised in the above table. The example is summarized in the below table.The Example is summarizing the results. The Examples are summarizing all the findings. The Example Is summarizing these results.The Examples are summarized by the following Table.The examples are summarized by the preceding table. They are summarized with the following examples:The Examples of ANOVAs are summarising all the examples. The Methods are summarization of the following:The Example of anOVA is summarising the results and summarizing them in the form of a summary. The method is summarizable by the fact that the results are summarizable. The methods are summarisable by the means of the various methods. The Method is summarization by the form. 'rouge1': 26.93, 'rouge2': 8.07, 'rougeL': 16.62, 'rougeLsum': 25.79
GOLD: Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher. In the ANOVA setting, the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether the population means of several groups are equal, and therefore generalizes the t-test to more than two groups. ANOVA is useful for comparing (testing) three or more group means for statistical significance. It is conceptually similar to multiple two-sample t-tests, but is more conservative, resulting in fewer type I errors, and is therefore suited to a wide range of practical problems.
Query in QMSum test set: *summarize the discussion about the vcr / vhr / dvd capability button .*
LED-*OASum*: The vocalsound button is a button on the vocalsound synthesizer that can be used to control the volume of the synthesizer. The vocalsound can also be used as a remote control. 'rouge1': 25.29, 'rouge2': 2.35, 'rougeL': 16.09, 'rougeLsum': 20.69
LED-base: summarize the discussion about the vcr / vhr / dvd capability button. marketing: yeah. project manager: okay. marketing : yeah. user interface: yeah, i think it's a good idea to have a button on the tv, i mean it's a button that you can put on your tv, and then you can use the button to put it on your TV, and you can also use the buttons to put the buttons on your television, and if you want to do that, you can do that. marketing, okay. product manager: oh. marketing, okay. user interfaces: yeah. marketing: okay, i'm sure we can do this, we're gonna do it, we're going to do it, we're gonna do this, we 'll do it. we'll do it. we've got a lot of stuff to do, we'll be doing this, and we's going to try to do this. we're going to be doing it, and i 'll be doing the same thing we did last time, we'll do the same, we will do it in the next time. we 'd do it with a button, we could do it on our tv. 'rouge1': 17.39, 'rouge2': 3.19, 'rougeL': 10.28, 'rougeLsum': 15.81
GOLD: project manager mentioned that the group needed to decide on the vcr capabilities and answered industrial designer 's question on whether vcrs were the same across the world . so industrial designer suggested a whole different set of buttons for individual vcrs . project manager suggested dvd type if considering the target group as young professionals and teenagers .
Aspect in NEWTS test set: *attack, attacks, killed, attacked, bomb, people, explosion, bombing, injured, blast,Cairo*
LED-*OASum*: At least 12 people were killed and more than 40 were injured in the attacks, according to the Egyptian government. The attacks occurred during the ongoing insurgency in Egypt's northern Sinai Peninsula. 'rouge1': 23.91, 'rouge2': 8.89, 'rougeL': 15.22, 'rougeLsum': 19.57
LED-base: attack, attacks, killed, attacked, bomb, people, explosion, bombing, injured, blast,Cairo (CNN)At least 12 people were killed Sunday, and more injured, in separate attacks on a police station, a checkpoint and along a highway in Egypt's northern Sinai, authorities said. Six people, including one civilian, were killed when a car bomb exploded near the police station in Al-Arish, capital of North Sinai, Health Ministry spokesman Hossam Abdel-Ghafar told Ahram Online. He said 40 people were injured. Ansar Beit Al-Maqdis, an ISIS affiliate, claimed responsibility for the attack, which came hours after another operation that the group also claimed. In that earlier attack, a first lieutenant, a sergeant and four conscripts were killed in an armored vehicle. Two other soldiers were injured and taken to a military hospital. Another soldier was killed in the second attack, the military said in a statement. The military said the attack was carried out by an unknown assailant, who was not identified by name, but was believed to be a member of the Islamic State of Iraq and the Levant (ISIL), a militant group affiliated with al-Qaida in the Islamic Maghreb. The army said the. 'rouge1': 36.08, 'rouge2': 20.55, 'rougeL': 25.88, 'rougeLsum': 32.94
GOLD: Twelve people in total were killed at a police station, a checkpoint, and along a highway in Egypt. Six people were killed in a car bomb near the police station and forty were injured. Six other people were killed when their armored car was attacked and two were injured. Three security personnel were also injured after being shot at.

Table 15: Examples of aspect-based summaries under zero-shot setting.

Zero-shot: Query in SQuALITY test set: *What is the significance of Ghost Ships in the story?*
LED-*OASum*: Space Man Willard is a short story by American writer James Willard, published in 1942. The story is about a space ship that Willard had never seen before. The ship was lost for twenty years, and Willard was the only survivor of the Mary Lou, which had been destroyed by a meteor. Willard's ship, the "Mary Lou II", was the first space ship to sail beyond the Solar System, and the only ship to have been lost for over a thousand years. The Mary Lou was named in honor of his father, who died when the ship was still in space. It was the last ship Willard would see before his ship would be able to return to Earth. The ghost of the ship would come back to him, and he would never see it again. It would be the last time he would see Earth again, and it would be too late for him to see the ship again. He would spend the next twenty years of his life in space alone, and would never return to the Earth again. In the story, Willard imagines that the ghost ship will come to Earth and bring him home. He imagines the ship will be the home of spacemen who can never go home again. And this is what happened to. 'rouge1': 45.78, 'rouge2': 10.71, 'rougeL': 16.32, 'rougeLsum': 35.52
LED-base: What is the significance of Ghost Ships in the story? Willard and Larry Dobbin left the Rocket Port for the exploration of an alleged planetoid beyond Pluto. The ship has not been seen or heard from since. J. Willard, II, son of the lost explorer, is planning the manufacture of a super-size exploration ship to be called Mary Lou II, in memory of his father." Chorus TextColor TextColor He awoke with a start and a cry of alarm ran through him as he thought that perhaps he might still be in the Mary Lou. The warm, smiling face of a man quickly reassured him.Chorus FontSize TextColor "I'll call the captain," the space man said. "He said to let him know when you came to."Chorus of TextColor The ghost ship was a ghost ship. It was the ghost ship, the ghost of a ghost. The ghost ships were ghost ships, ghost ships and ghost ships. The Ghost Ship was ghost ships of ghost ships in the Ghost Ship in his youth.chorus escription TextColor It was ghost ship of ghost ship TextColor When Willard awoke, he could not sleep. And so it was that each night—for Willard did not give up the Earth-habit of keeping time—Willard dreamed of the days he had known on Earth. chorus TextColor A ghost ship in the sky. "Nonsense!" Willard broke in hurriedly, hoping that the dying man would not see through the lie. "We've got the sun's gravity helping us drift back to Earth! We'll be there soon! You'll get well soon and we'll start to work again on a new idea of mine...." His voice trailed helplessly away and the words were lost. He was no longer able to sleep. Chorus of the Ghost ship in space TextColor Chorus Of the ghost ships In the sky, he thought, "It's all right," Willard whispered. The sick man did not hear him. Two tears rolled down his cheeks. His face contorted as he tried to withhold a sob.Chrome TextColor ITextColor "How do you feel, Space Man Willard?" Chorus, chorus of ghosts TextColor chorus chorus "Oh, you know me?" Willard looked at him in surprise, and then smiled, "I don't know you."chorus Of TextColor Darkness TextColor There was nothing to see. Darkness - Darkness Archdemon Darkness Gleaming Darkness 'rouge1': 37.12, 'rouge2': 9.31, 'rougeL': 16.32, 'rougeLsum': 35.52
GOLD: The Ghost Ship is a legend that sailors and space travelers alike have claimed people see in the moments before they die at sea or in space. In the seconds before Dobbin dies in Willard's arms, he looks out the window of the "Mary Lou" and claims to see the Ghost Ship himself. Throughout Willard's long periods of solitude aboard the "Mary Lou", he thinks he sees the Ghost Ship several times. First, from a distance, as a blinking light advancing closer and closer before turning back and sailing off into dark space; later, he thinks he sees the ship return, only this time it passes nearer before turning back and leaving again. With each return of the Ghost Ship, Willard believes he sees it clearer than he had before. After decades adrift in space, Willard believes a ship has finally come to rescue him. He does not think it is the Ghost Ship because it is solid, and he is greeted by a crew of people. However, the captain explains that the longer a vessel spends lost in space, the more it loses itself and slips into a kind of un-reality, along with those aboard. The more the "Mary Lou" drifted into this space, the more real the Ghost Ship became to Willard. Willard realizes that the "Mary Lou" has become a "ghost ship" herself.
Few-shot: Query in SQuALITY test set: *What is the significance of Ghost Ships in the story?*
LED-*OASum*: Ghost Ships are ghost-like ships that haunt the Solar System that gave it birth. They are the home of spacemen who can never go home again. When your last measure of fuel is burnt and your ship becomes a lifeless hulk—the Ghost will come—for you. Space Man Willard has a friend who died in space, J. Willard II, the son of the lost explorer. He is planning to manufacture a super-size exploration ship to be called Mary Lou II, in memory of his father. Each night he sees the sight of the ship and thinks he will see it again. However, after several dreams and visions of the ghost ship haunting him, it is clear that the ship is not a phantom ship. Instead, a life and blood rocket ship from Earth is tormenting him once again. Starlight shines on it and not through it! Its lines, window, vents were all solid and had none of the spirit-like quality he remembered in the Ghost Ship in his youth. He yawns and sighs as he remembers the days he had spent on Earth. Although Willard believed that the ghost ships were superstitions, he saw them from his ship, and he tried to rip the memories of Earth out. 'rouge1': 45.80, 'rouge2': 10.48, 'rougeL': 21.31, 'rougeLsum': 42.63
LED-base: What is the significance of Ghost Ships in the story?GALACTIC GHOSTS IN SPACEWillard is the captain of the Mary Lou ship. He is also a friend of Dobbin, a scientist who had come to explore the Solar System. He knows that Dobbin has died a horrible death, and the stories that he tells about the ghost ship that he has read about in the newspaper. He also knows that the unnamed unnamed ship which came to capture Dobbin's attention is a ghost ship. However, despite these beliefs, Willard does not believe that the ship has actually landed on the planet and is actually traveling to see Dobbin again. For a moment the reason in his mind tottered on a balance. Doubt assailed him. Was this the Ghost Ship come to torment him again? But no, it does not. He is greeted by a small crowd of people and his room is filled with empty rooms. The ghost ship seems to have abandoned its passengers and crew members, leaving Willard alone in his room. He does not understand that his friends have just landed their ship here and are experiencing the land for the first time. perspectemaleHe is taken aback by the large crowd that has gathered around him, and he begins to imagine that his ship has landed on Earth. He would see Earth again! That single thought runs through his mind constantly. The tapping of the space-telegrapher reassured him. He hears a message from the ship that says "CALLING SPACE SHIP MARY LOU," the message rapped out, "Yes, that is it!" With trembling fingers that he could scarcely control, old Willard sent the answering message. It is considered to be the most important message of the story. 'rouge1': 44.23, 'rouge2': 13.13, 'rougeL': 20.77, 'rougeLsum': 41.93

Table 16: Examples of aspect-based summaries under zero/few-shot setting. Few-shot means the model is finetuned on randomly chosen 3% samples from the training set.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
abstract and 1

☑ A4. Have you used AI writing assistants when working on this paper?
Grammarly for grammar correction

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
3

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We will follow Wikipedia's license CC BY-SA 3.0.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
3

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
3

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3

C ☑ Did you run computational experiments?

4.2

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4.2 and B.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4.2

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3

☒ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We only annotate a very small amount of examples by our own authors.

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
We only annotate a very small amount of examples by our own authors.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
3

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
3

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We only annotate a very small amount of examples by our own authors.