ASPECTNEWS: Aspect-Oriented Summarization of News Documents

Ojas Ahuja[1], **Jiacheng Xu**[1], **Akshay Gupta**[1], **Kevin Horecka**[2], **Greg Durrett**[1]
[1]The University of Texas at Austin
[2]Walmart NexTech
{ojas,jcxu}@utexas.edu, gdurrett@cs.utexas.edu

Abstract

Generic summaries try to cover an entire document and query-based summaries try to answer document-specific questions. But real users' needs often fall in between these extremes and correspond to aspects, high-level topics discussed among similar types of documents. In this paper, we collect a dataset of realistic aspect-oriented summaries, ASPECTNEWS, which covers different subtopics about articles in news sub-domains. We annotate data across two domains of articles, earthquakes and fraud investigations, where each article is annotated with two distinct summaries focusing on different aspects for each domain. A system producing a single generic summary cannot concisely satisfy both aspects. Our focus in evaluation is how well existing techniques can generalize to these domains without seeing in-domain training data, so we turn to techniques to construct synthetic training data that have been used in query-focused summarization work. We compare several training schemes that differ in how strongly keywords are used and how oracle summaries are extracted. Our evaluation shows that our final approach yields (a) focused summaries, better than those from a generic summarization system or from keyword matching; (b) a system sensitive to the choice of keywords.[1]

1 Introduction

Recent progress in text summarization (See et al., 2017; Liu and Lapata, 2019; Zhang et al., 2020a; Lewis et al., 2020) has been supported by the availability of large amounts of supervised data, such as the CNN/Daily Mail and XSum datasets (Hermann et al., 2015; Narayan et al., 2018), which provide a single, generic, topic-agnostic summary. However, a document often contains different *aspects* (Titov and McDonald, 2008; Woodsend and Lapata, 2012) that might be relevant to different users. For example, a political science researcher studying responses to earthquakes may want a summary with information about government-led recovery efforts and broader social impacts, not a high-level generic summary of what happened. Systems should be able to produce summaries tailored to the diverse information needs of different users. Crucially, these systems should be usable in realistic settings where a user is interested in vague *aspects* of the document, instead of a highly focused query.

In this work, we present a new dataset for evaluating single-document *aspect-oriented* extractive summarization which we call ASPECTNEWS. We derive subsets of examples from CNN/Daily Mail following certain topics, namely earthquakes and fraud reports. These domains are special in that the articles within them have several aspects which are repeatedly mentioned across articles and form coherent topics, e.g., impact on human lives of an earthquake. We ask annotators to select sentences relevant to such information needs, which correspond to imagined use cases. Interannotator agreement on full summaries is low due to the inherent subjectivity of the task, so rather than coming up with a consensus summary, we instead primarily evaluate against soft labels based on the fraction of annotators selecting a given sentence.

To benchmark performance on this dataset, we build a system that can summarize a document conditioned on certain aspect-level keywords without assuming annotated training data for those aspects. Since there are no large-scale supervised training sets suitable for this purpose, we explore methods to generate aspect-oriented training data from generic summaries. We compare these with past approaches (Frermann and Klementiev, 2019) on their ability to adapt to our aspect-oriented setting, which requires taking aspectual keyword inputs (as opposed to specific entities or queries) and being appropriately sensitive to these keywords.

Our experiments on our ASPECTNEWS dataset

[1]Code is available at https://github.com/oja/aosumm

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6494 - 6506

	Generic	Geo	Recovery
1	✓	✓	✓
2		✓	
3	✓		✓
4			✓
7			
8	✓		
9			✓
12			
15			

1. At least 42 people have died with hundreds more injured after a 6.2-magnitude earthquake hit Indonesia's Sulawesi island early Friday, according to Indonesia's Disaster Management Agency.
2. The epicenter of the quake, which struck at 1:28 a.m. Jakarta time, was 6 kilometers (3.7 miles) northeast of the city of Majene, at a depth of 10 kilometers (6.2 miles), according to Indonesia's Meteorology, Climatology and Geophysics Agency.
3. Thirty-four people died in the city of Mamuju, to the north of the epicenter, while another eight died in Majene.
4. In Majene, at least 637 were injured and 15,000 residents have been displaced, according to […]
7. Many people are still trapped under collapsed buildings, according to local search and rescue teams.
8. Rescuers search for survivors at a collapsed building in Mamuju city in Indonesia.
9. "Our priority is saving victims who are still buried under the buildings," Safaruddin Sanusi, head of West Sulawesi's Communications and Information Department, told CNN Friday. […]
12. "Most…of the people in Mamuju city are now displaced. They are afraid to stay at their houses."
15. "We need more extrication equipment and more personnel to work fast on saving victims trapped under the building."

Figure 1: Examples of an earthquake-related article paired with extractive summaries from the CNN/DM dataset. "Generic" represents the selection of a general purpose summarization model. "Geo(graphy)" (colored in green) and "Recovery" (colored in orange) indicate our aspects of interest for the summary. We highlight aspect-relevant phrases in the document.

and the SPACE dataset (Angelidis et al., 2021) find that our model produces summaries that score higher on agreement with human aspect-oriented annotations than generic summarization models, previous aspect-oriented models, and baselines such as keyword matching. Second, we find that the summaries our model generates are sensitive to the choice of keywords. Third, we find that our model performs competitively with leading models on the SPACE dataset in the multi-document setting. Finally, we find that abstractive query-focused systems (He et al., 2020) hallucinate significantly in this setting, justifying our choice of an extractive framework here.

2 Related Work

Relatively little recent work has focused on aspect-oriented summarization. One line of research focuses on summarization of documents with respect to specific queries (Baumel et al., 2014; Krishna and Srinivasan, 2018; Frermann and Klementiev, 2019; He et al., 2020; Xu and Lapata, 2020a). However, a query such as "*What facilities were damaged in the Oaxacan region?*" is a document specific query, which cannot be applied to other earthquake news articles and bears more resemblance to the task of long-form question answering (Fan et al., 2019). Our focus is closer to work on attribute extraction from opinions or reviews (Dong et al., 2017; Angelidis and Lapata, 2018), as factors like geographic details and recovery efforts are usually mentioned in many earthquake stories. Recent work has also begun to study summarization from an interactive perspective (Shapira et al., 2021); our approach could be naturally extended in this direction.

Methods Historically, most work on query-focused summarization has addressed the multi-document setting. You et al. (2011) apply regression models to this task, and Wei et al. (2008) approach the problem from the perspective of ranking sentences by their similarity to the query. These classic methods rely integrally on the multi-document setting, and so cannot be easily adapted to our setup. More recently, Xu and Lapata (2020b) focus on multi-document summarization by modeling the applicability of candidate spans to both the query and their suitability in a summary. Angelidis et al. (2021) explore a method using quantized transformers for aspect-oriented summarization, which we compare to.

Datasets There are several differences between ASPECTNEWS and other existing aspect-oriented summarization datasets. Firstly, ASPECTNEWS focuses on single-document summarization, while similar aspect-oriented datasets such as the SPACE dataset of reviews (Angelidis et al., 2021) and other attribute extraction settings (Dong et al., 2017; Angelidis and Lapata, 2018) are multi-document. Second, our dataset focuses on generalization *to new aspect types*, rather than assuming we've trained on data with those same aspects; that is, how can we produce appropriate aspect-oriented summaries of earthquake articles even if we have not trained on any? Third, compared to query-focused settings, our aspect-oriented dataset is closer to the actual information needs of users, since users are often interested in summaries about broad subtopics rather than specific queries.

The TAC 2010/2011 summarization datasets[2]

[2] https://tac.nist.gov/2011/Summarization

Domain	Aspect	Prompt	Keywords
Earthquake	GEO	geography, region, or location	region, location, country, geography, miles
	RECV	recovery and aid efforts (death toll and injuries, foreign/domestic government assistance, impact on survivors)	recovery, aid, survivor, injury, death
Fraud	PEN	penalty or consequences for the fraudster, or for others	penalty, consequences, jailed, fined, court
	NATURE	nature of the fraud: the amount of money taken, benefits for the fraudster, and how the fraud worked	amount, money, bank, stolen, time

Table 1: Prompts and keywords used for each of our two domains: Earthquake and Fraud. These represent prominent topics that users might be interested in.

propose *guided summarization* tasks that involve similar aspects. However, each article cluster in TAC has a single, fixed set of aspects that don't differ substantially from what a generic summary should capture. The DUC 2005/2006 task (Dang, 2005) does not have aspects but rather can accept a "granularity" level at which to produce the summary. Christensen et al. (2014) produce a hierarchy of relatively short summaries among multiple documents.

Other previous work (He et al., 2020; Xu and Lapata, 2020a; Tan et al., 2020) proposes constructing keyword sets for each individual document for training. Krishna and Srinivasan (2018); Frermann and Klementiev (2019) condition on topic tokens referring to the topic tags in metadata. Compared to these other approaches, we focus more on evaluation of aspects, as opposed to a purely keyword- and query-driven view.

3 Aspect-Oriented Data Collection

We begin by considering our target application: users who have specific information needs that they want to be satisfied. This consideration broadly falls under the category of **purpose factors** defined by Jones (1998) and should be accounted for in the summarization process.

Our data collection process involves the following steps: (1) Identifying clusters of articles in our **target domains** from a large corpus of news summaries. (2) Manually specifying multiple **user intents** per target domain, representing the *aspect* of the summarization process. (3) **Crowdsourcing** annotation of extractive summaries in these domains based on the user intents.

3.1 Target Domains

We draw our datasets from the English-language CNN/Daily Mail summarization dataset (Hermann et al., 2015). We manually identified two domains, *earthquakes* and *fraud*, based on inspecting clusters of articles in these domains. These two domains are ideal for two reasons. First, they contain a significant number of on-topic articles (over 200) after careful filtering. Second, the articles in these domains are reasonably homogeneous: each article would often feature at least broadly similar information about an event, making aspect-based summarization well-defined in these cases.[3] Although not completely universal, most earthquake articles refer to some information about each of two aspects here: *geography* (GEO) and *recovery* (RECV). Figure 1 shows an example of an earthquake-related article. Similarly, most fraud articles include information about the *penalty* (PEN) imposed for the fraud, and the *nature* (NATURE) of the fraud.

To retrieve our examples from these two domains, we first encode each article in CNN/DM corpus $\mathcal{C}$ with a text encoder E. We adopt the Universal Sentence Encoder (Cer et al., 2018) for its efficiency and robustness. We create an exemplar sentence for each domain to serve as the target to retrieve the most relevant content. We describe the choice of exemplar sentences in Section A.2. We measure the similarity of each candidate article c and the exemplar sentence s as the average of the cosine similarity between each of the candidate article's sentences c_i and the exemplar, $sim(c, s) = \frac{1}{n}\sum_{i=1}^{n} \cos(E(c_i), E(s))$.

We found this procedure to be more robust than simple keyword matching for retrieving articles with coherent aspects; for example, keyword matching for "earthquakes" resulted in returning articles primarily about tsunamis due to the imbalanced data distribution.

[3] By contrast, other domains like legislation were too heterogeneous: articles about passing a bill may focus on different aspects of a bill's journey, comments or quotes by elected officials, impact of the legislation, or other factors. We could not come up with a plausible unified information need for the sorts of articles available in this dataset, although our eventual system can be applied to such documents if given appropriate guidance.

3.2 Specifying User Intents

With these two domains, we examine our dataset to derive aspects that simulate realistic information needs of users.

Table 1 describes the domain, aspect, annotation prompt and keywords used for evaluation. For each domain, we establish two aspects. Each aspect must be well-represented in the corpus and easy to understand by both readers and annotators. The authors annotated these aspects based on inspection of the articles and brainstorming about user intents based on scenarios. For example, the *penalty* scenario was motivated by a real use case derived from the authors' colleagues investigating reporting of wrongdoing in news articles at scale, where summarization can be used to triage information.

3.3 Crowdsourcing

Finally, to construct actual extractive summaries for evaluation in these domains, we presented the user intents to annotators on Amazon Mechanical Turk. An annotator is shown a description of intent from Table 1 along with an article and is asked to identify a few sentences from the article that constitute a summary. They can rate each sentence on a scale from 0 to 3 to account for some sentences being more relevant than others. Their final summary, which they are shown to confirm before submitting, consists of all sentences rated with a score of at least 1. The exact prompt is shown in the Appendix.

Each article was truncated to 10 sentences for ease of annotation. This assumption was reasonable for the two domains we considered, and the truncation approach has been used in See et al. (2017) without much performance degradation. We found that annotators were unlikely to read a full length article due to the inherent lead bias in news articles, so this also helped simplify the task. In order to maintain a high quality of annotations, we discard annotations that do not have at least a single selected sentence in common with at least a single other annotator on that sample. In practice, this only discards a handful of isolated annotations.

3.4 Data Analysis & Annotator Agreement

In Table 2, we show the basic statistics of the collected dataset. We show the distribution of the number of sentences agreed upon by the annotators in Table 3. We see that annotators somewhat agree in most cases, but relatively few sentences are uniformly agreed upon by all annotators. Our initial pilot studies also showed that annotators are often unsure where the cutoff is for information to be notable enough to include in a summary. We therefore view this disagreement as inherent to the task, and preserve these disagreements in evaluation rather than computing a consensus summary.

	# articles	# sent	# words
PEN	100	2.90	30.5
NATURE	100	2.79	29.9
GEO	100	2.53	28.4
RECV	100	2.76	27.0

Table 2: Statistics for the collected datasets. For each aspect we collect 100 articles and each article is annotated by 5 Turkers. #sent and #words are the average number of sentences selected and average number of words in each sentence.

Agreement	1	2	3	4	5
Freq (%)	19.61	29.26	25.16	19.16	6.80

Table 3: Majority agreement distribution of 5 annotators on filtered collected data.

We also compare the overlap between aspect-oriented annotation and generic extractive oracle derived from reference summaries from CNN/DM. In Table 4, the similarity and exact match[4] between generic oracle summaries and the top 3 annotated sentences are fairly low, which means the annotated aspect driven summaries significantly differ from the standard extractive oracle.

4 Building an Aspect-Oriented System

Our aspect-oriented data collection works well to create labeled evaluation data, but it is difficult to scale to produce a large training set. Identifying suitable domains and specifying user intents requires significant human effort, and collecting real test cases at scale would require a more involved user study.

We build an aspect-oriented model without gold-labeled aspect-oriented training data. We do this by generating keywords for each article in CNN/DM, and training the model to learn the relationship between these keywords and a summary. Our system follows broadly similar principles to He et al. (2020), but in an extractive setting.

[4]The number of annotated examples for each aspect is 100, so the EM is an integer.

StdRef vs.	Jaccard Sim.	EM (%)
Pen	0.247	1.0
Nature	0.249	2.0
Geo	0.265	2.0
Recv	0.201	1.0

Table 4: Comparison of annotation labels and the non-query focused extractive oracle derived from reference summaries. We take the top-3 most common selected sentences from each aspect-oriented dataset and compute Jaccard similarity between the sets and the percentage of exact matches (EM).

Article: 1. Justine Greening has called for a major shake-up in the EU aid budget – as it emerged more than half the cash is squandered on relatively rich countries.
2. The International Development Secretary challenged the basis of the £10-billion-a-year budget, which channels cash to countries such as Turkey, Iceland and Brazil.
3. She is pressing for a major shift in policy to target resources at the poorest countries.
4. International Development Secretary Justine Greening today insisted aid money [...]
5. Miss Greening held talks with ministers from [...]
7. Miss Greening said: 'I don't think it's right that the EU still gives money to those countries higher up the [...]
9. Her intervention comes amid mounting concern about the EU aid budget, which [...] total aid budget. [...]

Keywords: countries, budget, development, 10-billion, Turkey

Table 5: An example article from CNN/DM and keywords extracted. These keywords indicate both highly specific concepts and broad topic, but a model trained on data with appropriate reference summaries can learn to leverage either specific *or generic* keywords in the summarization process.

4.1 Keyword-controlled Data

We present a scheme to generate keywords for each document from the original dataset. CNN/DM consists of pairs (D, S) of a document D and associated summary S. We aim to augment these to form (D, K, S') triples with keywords K and a possibly modified summary S'. Our mixed augmentation technique requires training the model on **both** (D, S) and (D, K, S') for a given document. We now describe the steps to create this data.

Keyword Extraction For each document in CNN/DM, we calculate the most important tokens in that document according to their TF-IDF ranking with respect to the entire corpus. Of these tokens, we select the ones that are present in the reference summary. This process selects tokens that are more likely to be consequential in affecting the output summary.

Reference Summary Computation Since CNN/DM reference summaries are abstractive, we need to derive extractive oracle summaries for training; these consist of sentence-level binary decisions $\mathbf{E} = E_1, \ldots, E_m$ for each sentence. Traditionally, this is done by finding a set of sentences that maximize ROUGE-2 (R2) with respect to the reference: $\text{argmax}_{\mathbf{E}} R2(\mathbf{E}, S)$ (Gillick and Favre, 2009; Nallapati et al., 2017). However, training the model to predict $P(S_1, \ldots, S_m \mid D, k)$, an extractive analogue of He et al. (2020), was insufficient for our extractive model to learn to be sensitive to keywords; it merely learned to return a good generic summary regardless of what keywords were given.

To instill stronger dependence on the keywords, we made two modifications to this process. First, we modified the reference summary by concatenating the keywords with the reference summary *before* computing the extractive oracle summary. This concatenation makes the oracle extraction more likely to select sentences containing the keywords, though modifying the reference summary requires maintaining a balance between the influence of keywords and of the original gold summary.

Second, we use BERTScore (Zhang et al., 2020b, BS) rather than ROUGE-2 to identify sentences that closely match the reference summary. BERTScore turns out to boost the evaluation performance by a large margin, as shown in Table 12, so we use BERTScore for oracle extraction for all our experiments. One reason for this is that the ROUGE-2 summaries favor exact keyword matches in selecting sentences, so the trained model simply learned to keyword matching in extreme cases. Our final reference summary is therefore $\text{argmax}_{\mathbf{E}} BS(\mathbf{E}, S + nK)$, where n is a hyperparameter we discuss next.

Keyword Intensity To compute n, we introduce another parameter r that controls the ratio of keyword tokens to original reference summary tokens. Higher values of r lead to extracting sentences in a manner more closely approximating keyword matching, but yielding poor standalone summaries. On the other hand, lower values of r may lead to generic summaries insensitive to the keywords. In practice, the number of times a keyword w is concatenated to the original summary S is defined as $n = r \times \frac{\text{len}(S)}{\#(\text{keywords})}$ where $\text{len}(S)$ is the number of tokens in the original summaries and $\#(\text{keywords})$ is the total number of keywords available. When $r = 1$, the concatenated keywords have the same length of the original summary.

Mixed Training We explore a variant of training where we include training data with multiple variants of each original document from the dataset. Each document in the original dataset is mapped to two training samples, (1) a document without keywords and an unmodified oracle extractive summary, (2) a document with keywords and an oracle extractive summary using our modification procedure.

4.2 Aspect-Oriented Model

Our model is trained to predict a summary S from a document-keywords pair (D, K). Following BERTSUM (Liu and Lapata, 2019), we fine-tune BERT (Devlin et al., 2019) for extractive summarization using our modified CNN/Daily Mail dataset with keywords. During training, we prepend a special token followed by the keywords to the original document, and use the modified oracle extractive summary as the gold outputs. During inference, the keywords are user-defined. This scheme is similar to He et al. (2020), but differs in that it is extractive.

We refer to this model, trained on our BERTScore references with the mixed training scheme, as AOSUMM.

5 Experiments

We evaluate our model on the ASPECTNEWS dataset, comparing performance on aspect-oriented summarization to several baselines. We additionally experiment on the SPACE multi-document dataset (Angelidis et al., 2021) to provide a point of comparison on a prior dataset and show that our aspect-oriented method is competitive with other systems.

5.1 Metrics

On ASPECTNEWS, we evaluate our model against the annotations using using F_1 score and ROUGE scores. It is impossible to achieve 100 F_1 on this task due to inherent disagreement between annotators. One downside of F_1 is that the model may be penalized even when the predicted sentence is very similar to the annotation, for this reason we also calculate ROUGE-1, -2, and -L scores (Lin, 2004). On the SPACE dataset, the gold summaries are abstractive, so we only calculate ROUGE scores.

5.2 Baselines & Competitor Models

On the SPACE corpus, we primarily focus on comparisons to quantized transformer (QT) (Angelidis et al., 2021) and CTRLSUM (He et al., 2020). For the ASPECTNEWS dataset, we benchmark our system against several other models and baselines which we now describe.

Heuristic and QA Baselines KEYWORD takes the keywords described in Table 1 and greedily finds the first occurrence of each keyword in the input document. STDREF stands for the extractive oracle given the original reference summaries from CNN/DM. QA uses an ELMo-BiDAF question answering model (Seo et al., 2017; Peters et al., 2018) to find answers to synthetic questions "*What is {keyword}?*" for each keyword in the article. We select the sentence where the selected span is located as a sentence to extract. Each of these three technique is an extractive baseline where top sentences are selected.

Summarization Baselines We also compare our AOSUMM model against text summarization models, and query-focused models from previous work (retrained or off-the-shelf). (i) BERTSUM is a `bert-base-cased` extractive summarization model fine-tuned on CNN/DM (Liu and Lapata, 2019). (ii) BERT-FK shares the similar model architecture as BERTSUM but the training data comes from Frermann and Klementiev (2019). This data is constructed by interleaving several articles from the CNN/DM dataset together, extracting a coarse aspect from the original URL of one of the article, and setting the new gold summary to match that article. (iii) CTRLSUMis an off-the-shelf **abstractive** summarization model with the capability of conditioning on certain queries or prompts (He et al., 2020). (iv) Our model AOSUMM is based on BERTSUM and trained with techniques described in Section 4.

5.3 Results

ASPECTNEWS The experimental results on ASPECTNEWS are shown in Table 6. We find that our model outperforms our baselines across F_1, ROUGE-1, ROUGE-2, and ROUGE-L scores. Significantly, our model generally outperforms keyword matching, demonstrating that semantic match information from training with the BERTScore oracle may be more useful than training with a ROUGE oracle in terms of reproducing annotators' judgments; recall that our model has not been trained on any ASPECTNEWS data and only on our synthetic data.

Model	PENANNOT				NATUREANNOT				GEOANNOT				RECVANNOT			
	F_1	R-1	R-2	R-L	F_1	R-1	R-2	R-L	F_1	R-1	R-2	R-L	F_1	R-1	R-2	R-L
STDREF	32.9	51.7	39.5	40.7	33.5	53.0	41.3	42.0	34.9	51.9	41.3	42.1	28.2	45.7	33.0	37.4
KEYWORD	39.2	62.0	50.6	47.1	38.3	58.7	46.6	45.0	50.9	67.9	59.9	53.7	32.8	53.3	41.6	43.9
QA	30.7	46.9	36.8	37.7	26.5	39.1	28.8	32.2	**52.4**	63.0	58.9	**56.8**	32.9	46.6	36.5	38.5
BERTSUM	40.1	60.1	47.8	46.5	41.6	63.5	51.7	**49.4**	46.4	65.4	56.4	51.4	37.3	55.8	44.8	44.6
BERT-FK	24.5	43.9	28.9	33.2	21.0	40.8	23.4	28.3	23.9	42.4	30.3	32.9	21.4	35.4	21.3	26.9
CTRLSUM	N/A	47.8	30.2	33.0	N/A	51.7	35.3	35.4	N/A	21.6	8.0	19.6	N/A	32.3	11.6	19.2
AOSUMM	**44.8**	**64.2**	**54.1**	**51.6**	**45.2**	**64.4**	**53.9**	48.0	49.9	**69.1**	**61.2**	54.2	**39.6**	**59.5**	**49.1**	**46.7**
Max	60.3				61.5				70.2				61.4			

Table 6: Performance comparison of our model (AOSUMM) versus baselines on the ASPECTNEWS dataset in both the earthquakes and fraud domains, using our geography (GEOANNOT) and recovery (RECVANNOT) aspects for the former and penalty (PENANNOT), and nature (NATUREANNOT) aspects for the latter. The last row displays the maximum possible F_1 score due to the disagreement of annotation.

	Service	Location	Food	Building	Cleanliness	Rooms
BERTSUM	12.4	16.7	13.0	15.6	13.8	12.5
CTRLSUM	20.1	18.6	17.4	18.9	23.3	19.7
QT	26.0	23.6	17.7	16.0	25.1	21.6
AOSUMM	26.9	20.3	17.4	16.4	22.8	21.6

Table 7: ROUGE-L scores on the SPACE dataset of our model, AOSUMM, versus BERTSUM, CTRLSUM, and quantized transformer (QT). Despite being an extractive model, our approach is competitive with strong query-focused or aspect-based models.

KW	F_1	R-1	R-2	R-L	F_1	R-1	R-2	R-L
	PENANNOT				NATUREANNOT			
PEN	**44.8**	64.2	54.1	**51.6**	41.8	60.8	49.5	46.5
NATURE	44.3	**65.5**	**56.0**	51.3	**45.2**	**64.4**	**53.9**	**48.0**
	GEOANNOT				RECVANNOT			
GEO	**49.9**	**69.1**	**61.2**	**54.2**	38.0	56.2	45.3	46.2
RECV	42.8	60.4	49.7	47.8	**39.6**	**59.5**	**49.1**	**46.7**

Table 8: Keyword sensitivity analysis broken down by domain of ASPECTNEWS.

	Jaccard Sim.	EM (%)
PENKW vs. NATUREKW	0.657	21.0
GEOKW vs. RECVKW	0.559	22.0

Table 9: Difference in AOSUMM outputs with different keywords. We compute Jaccard similarity between the sets and the and percentage of Exact Matches (EM).

We note that our model's performance falls behind keyword matching some baselines in the geography aspect; this may be because the aspect is relatively homogeneous and can be easily approximated by keyword matching.

SPACE The results on all the aspects of the SPACE dataset are shown in Table 7. All of the aspect-oriented models exceed the performance of the generic summaries produced by BERTSUM. We also find that our model performs competitively with the quantized transformer (QT) (Angelidis et al., 2021) and CTRLSUM (He et al., 2020) methods in this dataset. This is a surprising result: the AOSUMM model is trained *only* with out-of-domain synthetic data, without access to the aspects prior to keywords specified at test time. Additionally, this is an *abstractive* task that we are applying an *extractive* model to.

5.4 Ablations and Analysis

Keyword Sensitivity We evaluate the sensitivity of the model to different keywords. There is some overlap between the summaries returned by different keyword sets, as shown by the Jaccard similarity: some sentences may fit under both GEO and RECV, or both PEN and NATURE. Table 9 shows statistics of this, with the Fraud keyword sets yielding more similar summaries than those in Earthquake. We also confirm that using the keywords "matched" to our setting outperforms using other sets of keywords in that domain (Table 8) suggesting that our model is picking summaries in a keyword-driven fashion.

Keyword Intensity We can vary the parameter k controlling the number of times we append the keywords to the reference summary in order to generate the oracle extractive summary. We experiment with different level of intensity and show the result in Table 10. For most cases, $r = 1$ works well among all the datasets.

	GEO	RECV	PEN	NATURE
$r = 0.5$	48.4	**40.0**	41.9	42.7
$r = 1.0$	**49.9**	39.6	**44.8**	**45.2**
$r = 2.0$	49.0	39.4	41.9	42.0

Table 10: Comparison of various levels of keyword intensity. We experiment with different level of keyword intensity for different oracle and train our AOSUMM model on these setting. We show the F_1 of model's prediction and human annotation. The larger the r, the more keywords will be concatenated.

6 Qualitative Evaluation & Comparison

Extractive vs. Abstractive Comparison It is difficult to directly compare the quality of summaries produced by an extractive model to those produced by an abstractive model. Abstractive models do not extract individual sentences from a summary so direct F_1 evaluations cannot be compared in the manner of Table 6. ROUGE scores are a misleading comparison given that an extractive model will be better matched to our extractive ground truths. Therefore, we perform a qualitative analysis to determine the models' relative responsiveness to keywords and relative advantages and disadvantages.[5]

Keyword Sensitivity Comparison Although both CTRLSUM and AOSUMM are sensitive to the choice of keywords and alter their summary in response to different keywords, CTRLSUM often either hallucinates false information (Maynez et al., 2020) or simply rewords the prompt in the generated summary. We found that just under the GEO keywords in the earthquakes domain, out of 100 sample articles the bigram "not known" appears 27 times in relation to describing the location of the earthquake and "not immediately known" appears another 24 times. The CTRLSUM model frequently rephrases the prompt rather than synthesizing information in the document related to the keywords into a cogent summary.

Comparison of Factuality of Output Table 11 shows one example of CTRLSUM hallucination in the GEO case. Here, the model also rewords the prompt and inserts it into the summary without adding new information. Although such behavior may possibly perform well on automated metrics, it does not serve the purpose of query-focused summarization.

[5]Note that for the abstractive SPACE dataset we considered here, we found that the performance difference between our model and abstractive models is small. Our investigation found that, at least on this dataset, abstractive models are engaging in heavy copying of the source text, suggesting that extractive models may be almost as well suited for this task as abstractive models.

Sel	Article
	(CNN) – A 7.2-magnitude earthquake has struck south of the Mariana Islands, according to the U.S. Geological Survey.
G	The Mariana Islands – an archipelago in the western Pacific Ocean – are made up of two U.S. territories, Guam and the Commonwealth of the Northern Mariana Islands.
	The islands sit about three-quarters of the way from Hawaii to the Philippines.
R	The Pacific Tsunami Warning Center did not issue a tsunami warning after the quake, which struck at 7:19 a.m. Saturday (5:19 p.m. ET Friday).
R	"We wouldn't expect any kind of significant tsunami for this event," said the center's director, Charles McCreery, noting that the quake's magnitude was relatively low to provoke one.
R	There were no immediate reports of casualties or damage, emergency management officials said.
G	The quake struck about 375 kilometers (233 miles) west-southwest of Hagatna, Guam, and 445 kilometers (276 miles) west-southwest of Rota, Northern Mariana Islands.
	CTRLSUM GEO
	NEW: The location of the quake is in the province of Yucatán. NEW: There are no immediate reports of casualties or damage. The quake is centered about 375 kilometers (233 miles) west-southwest of Hagatna, Guam. The U.S. Geological Survey says it was a 7.2-magnitude quake. The earthquake is centered in the Yucatan province of Mexico. The country's geography is similar to that of the U.N. region.
	CTRLSUM RECV
	NEW: The death toll from the quake is not immediately known. The U.S. Geological Survey reports a 7.2-magnitude quake. The Mariana Islands sit about three-quarters of the way from Hawaii to the Philippines. "There is a survivor. There is an injury. There will be an aid.recovery. process," the U.N. secretary-general says. The quake is centered about 375 kilometers (233 miles) west-southwest of Hagatna, Guam.

Table 11: An example article from the earthquakes domain, along with summaries selected by AOSUMM (denoted as *G* and *R*) and CTRLSUM with GEO and RECV keyword.

Extractive summaries Table 11 shows that our model is able to successfully extract relevant parts of the document for our aspects under consideration. There are some features which may make these summaries hard to process in isolation, such as the quake in the first *R* sentence; our method could be extended with prior techniques to account for anaphora resolution (Durrett et al., 2016).

7 Conclusion

In this paper, we present a new dataset for aspect-oriented summarization of news articles called ASPECTNEWS. Unlike query-focused summarization datasets which are often driven by document specific facts or knowledge, this aspect-oriented task is designed to mimic common user intents in domain-specific settings. We present a keyword-controllable system trained on synthetic data and show that it can perform well on ASPECTNEWS without training on the target domains, performing

better than a range of strong baseline methods.

Acknowledgments

This work was chiefly supported by funding from Walmart Labs and partially supported by NSF Grant IIS-1814522, a gift from Amazon, and a gift from Salesforce Inc. Opinions expressed in this paper do not necessarily reflect the views of these sponsors. Thanks to Ido Dagan for helpful discussion and suggestions about this paper, as well to the anonymous reviewers for their thoughtful comments.

References

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics (TACL)*, 9:277–293.

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.

Tal Baumel, Raphael Cohen, and Michael Elhadad. 2014. Query-chain focused summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 913–922, Baltimore, Maryland. Association for Computational Linguistics.

Daniel Matthew Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, C. Tar, Yun-Hsuan Sung, B. Strope, and R. Kurzweil. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Janara Christensen, Stephen Soderland, Gagan Bansal, and Mausam. 2014. Hierarchical summarization: Scaling up multi-document summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 902–912, Baltimore, Maryland. Association for Computational Linguistics.

Hoa Trang Dang. 2005. Overview of duc 2005.

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.

Li Dong, Shaohan Huang, Furu Wei, Mirella Lapata, Ming Zhou, and Ke Xu. 2017. Learning to generate product reviews from attributes. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 623–632, Valencia, Spain. Association for Computational Linguistics.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008, Berlin, Germany. Association for Computational Linguistics.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Lea Frermann 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.

Dan Gillick and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18, Boulder, Colorado. Association for Computational Linguistics.

Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *arXiv preprint arXiv:2012.04281*.

Karl Moritz Hermann, Tomás Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Karen Sparck Jones. 1998. Automatic summarising: Factors and directions. In *Advances in Automatic Text Summarization*, pages 1–12. MIT Press.

Kundan Krishna and Balaji Vasan Srinivasan. 2018. Generating topic-oriented summaries using neural attention. 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 1697–1705, New Orleans, Louisiana. 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.

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.

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.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan Thomas Mcdonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

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*, Brussels, Belgium.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Ori Shapira, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. 2021. Extending multi-document summarization evaluation to the interactive setting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 657–677, Online. 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.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of ACL-08: HLT*, pages 308–316, Columbus, Ohio. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Furu Wei, Wenjie Li, Q. Lu, and Y. He. 2008. Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization. In *Proceedings of the Special Interest Group on Information Retrieval (SIGIR)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

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.

Yumo Xu and Mirella Lapata. 2020a. Abstractive query focused summarization with query-free resources. *arXiv preprint arXiv:2012.14774*.

Yumo Xu and Mirella Lapata. 2020b. 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.

Ouyang You, Wenjie Li, Sujian Li, and Qin Lu. 2011. Applying regression models to query-focused multi-document summarization. *Information Processing & Management*, 47:227–237.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

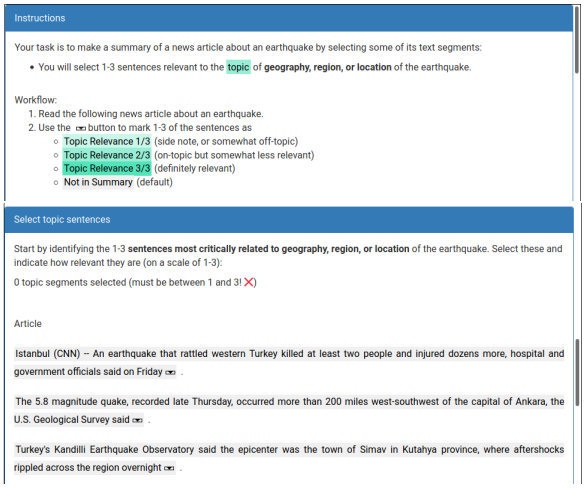

Figure 2: User interface for Turkers' annotation.

A Appendices

A.1 Training Details

For all models, we split CNN/Daily Mail set into the standard 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs following See et al. (2017).

We follow the training procedure for BERTSUM (Liu and Lapata, 2019) with modifications. We use the cased variant of `bert-base-cased` available through HuggingFace (Wolf et al., 2019) instead of uncased and do not lowercase the dataset during preparation. Our learning rate schedule follows Vaswani et al. (2017) with

$$lr = 2e^{-3} \cdot \min(\text{step}^{-0.5}, \text{step} \cdot \text{warmup}^{-1.5})$$

where warmup $= 10000$.

For fine-tuning AOSUMM on the modified CNN/DM dataset, the training completes in 8 hours on a single NVIDIA Quadro RTX 8000.

A.2 Exemplar Sentences

In order to generate earthquake and fraud domain data we filter the CNN/DM dataset using similarity between latent representations of Universal Sentence Encoder (USE) (Cer et al., 2018). To find domain-related articles, we need to generate a sentence that is vague enough to match most in-domain articles but specific enough to exclude articles outside the domain. For earthquakes we found the sentence "*An earthquake occurred.*" to work well. We embedded this sentence with USE, and calculated distance in latent space to articles in CNN/DM. For the fraud dataset we use the simlar sentence "*A fraud occured.*" After inspecting the matches, we manually exclude articles that are outside the domain.

	F_1	R-1	R-2	R-L	F_1	R-1	R-2	R-L
	PEN				NATURE			
RS	36.3	55.8	42.1	43.0	38.0	57.6	44.8	43.3
BS	**44.8**	**64.2**	**54.1**	**51.6**	**45.2**	**64.4**	**53.9**	**48.0**
	GEO				RECV			
RS	39.5	59.2	49.1	47.2	34.9	54.9	44.3	45.2
BS	**49.9**	**69.1**	**61.2**	**54.2**	**39.6**	**59.5**	**49.1**	**46.7**

Table 12: Comparison of our AOSUMM model trained on data using ROUGE (RS) or BERTScore (BS) as the scoring metric for oracle extraction. Training with BERTScore oracle summaries gives much stronger performance.

	GEO	RECV	PEN	NATURE	Avg.
Non-Mixed	48.0	**41.8**	43.9	43.7	44.3
Mixed	**49.9**	39.6	**44.8**	**45.2**	**44.9**

Table 13: Comparison of AOSUMM with or without mixed training data. We show the F_1 of the system output and human annotation on four domains.

A.3 Crowdsourcing

To improve the quality of the data collected, we educate annotators with detailed instruction and user-friendly interface shown in Figure 2. We also manually sample and check the collected data.

A.4 Oracle Derivation: BERTScore vs. ROUGE

In Table 12 we show the performance improvement from replacing ROUGE-derived oracle labels with their BERTScore-derived counterparts. Using BERTScore (Zhang et al., 2020b) to obtain oracle extractive summaries for training data produces models that are significantly stronger than models trained on sentences selected by maximizing ROUGE score. We hypothesize this is because ROUGE score maximization essentially limits what the model learns to lexical matching, while BERTScore can score based on more abstract, semantic criteria.

A.5 Mixed vs. Non-Mixed

We compare models trained using the mixed technique against models trained without any augmentation, and find that the mixed technique generally provides some benefit, but inconsistently. In Table 13, the Mixed technique is effective on GEO, PEN, and NATURE, but not RECV. The small per-

formance improvement from Mixed training may result from the model more easily learning the relationship between the keywords and the aspect-oriented summaries due to mixed examples. Another benefit of this technique is that a single model is capable of producing both generic and aspect-oriented summaries.

A.6 SPACE Evaluation Details

Several adjustments were made in order to run our model on the SPACE dataset. Since there are multiple input documents per summary, we first concatenated all documents together and treated the result as a single article. In order to process this large "article" with our model, we processed it in 512-token chunks using BERT in order to obtain representations from the `[CLS]` token, and then concatenated those representations together before passing them through the classification layer. This allowed selection of any sentence from any part of the input. The following keywords were used for each of the aspects in the dataset: (i) service, customer, staff, employee, assistance; (ii) location, room, region, hotel, place; (iii) food, dining, restaurant, dinner, meal; (iv) building, establishment, room, property, site; (v) cleanliness, sanitary, polished, clean, washed; (vi) rooms, chair, table, bed, wall.