Unsupervised Multi-Granularity Summarization

Ming Zhong[§] **Yang Liu**[†] **Suyu Ge**[§] **Yuning Mao**[§] **Yizhu Jiao**[§]
Xingxing Zhang[‡] **Yichong Xu**[†] **Chenguang Zhu**[†] **Michael Zeng**[†] **Jiawei Han**[§]
[§]University of Illinois at Urbana-Champaign
[†]Microsoft Cognitive Services Research [‡]Microsoft Research Asia
{mingz5, suyuge2, yuningm2, yizhuj2, hanj}@illinois.edu
{yaliu10, xizhang, yichong.xu, chezhu, nzeng}@microsoft.com

Abstract

Text summarization is a user-preference based task, i.e., for one document, users often have different priorities for summary. As a key aspect of customization in summarization, granularity is used to measure the semantic coverage between summary and source document. However, developing systems that can generate summaries with customizable semantic coverage is still an under-explored topic. In this paper, we propose the first unsupervised multi-granularity summarization framework, GRANUSUM. We take events as the basic semantic units of the source documents and propose to rank these events by their salience. We also develop a model to summarize input documents with given events as anchors and hints. By inputting different numbers of events, GRANUSUM is capable of producing multi-granular summaries in an unsupervised manner. Meanwhile, we annotate a new benchmark *GranuDUC* that contains multiple summaries at different granularities for each document cluster. Experimental results confirm the substantial superiority of GRANUSUM on multi-granularity summarization over strong baselines. Furthermore, by exploiting the event information, GRANUSUM also exhibits state-of-the-art performance under conventional unsupervised abstractive setting.[1]

Multiple News Articles about Hurricane Mitch
Honduras braced for potential catastrophe Tuesday as Hurricane Mitch roared through the northwest Caribbean, churning up high waves and intense rain ... (**Total 3,358 words**)
Summary of Coarse Granularity Level
Hurricane Mitch, category 5 hurricane, brought widespread death and destruction to Central American, and Honduras was especially hard hit. (**Total 19 words**)
Summary of Medium Granularity Level
Hurricane Mitch approached Honduras on Oct. 27, 1998 with winds up to 180mph a Category 5 storm ... The European Union, international relief agencies, Mexico, the U.S., Japan, Taiwan, the U.K. and U.N. sent financial aid, relief workers and supplies. (**Total 53 words**)
Summary of Fine Granularity Level
A category 5 storm, Hurricane Mitch roared across the northwest Caribbean with 180 mph winds across a 350-mile front ... The greatest losses were in Honduras where 6,076 people perished ... At least 569,000 people were homeless across Central America. Aid was sent from many sources (European Union, the UN, US and Mexico). The U.S. and European Union were joined by Pope John Paul II in a call for money and workers to help the stricken area. However, Relief efforts are hampered by extensive damage ... (**Total 133 words**)

Table 1: An example from our multi-granularity summarization benchmark GranuDUC. Texts of the same color (blue, red) denote similar points described in different ways. Finer-grained summaries have higher semantic coverage with the original text.

1 Introduction

Text summarization aims to condense and summarize long documents into a concise paragraph containing the essential points of the original texts (See et al., 2017; Liu and Lapata, 2019; Wang et al., 2020; Zhong et al., 2020; Liu et al., 2022; An et al., 2022). Notably, the requirements for summarization are highly customized and personalized for different users (Díaz and Gervás, 2007; Lerman et al., 2009; Yan et al., 2011; Chen et al., 2021b). Therefore, generating quality summaries to meet different preferences should be a natural capability of summarization systems.

Granularity, a key aspect of customization in summarization, is used to measure the degree of semantic coverage between summary and source documents (Mulkar-Mehta et al., 2011). To cater to the diverse needs of readers, the granularity level of summaries often varies in a wide range. As shown in Table 1, given multiple news about Hurricane Mitch, the most compact summary (Coarse Granularity Level) accommodates only the most important event to help people grasp the overall picture of the input documents. Interested readers, on the other hand, may prefer more fine-grained

[1]Dataset for this paper can be found at: https://github.com/maszhongming/GranuDUC.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4980–4995

summaries (Medium and Fine Granularity Level) to acquire additional details, such as how many casualties were caused and how different countries aided Honduras. Thus, multi-granularity summaries can meet the intent of different users and are more versatile in real-world applications.

Most existing summarization models and benchmarks focus solely on single-granularity summarization. It limits the ability of these systems to adapt to different user preferences and generalize to a wider range of granularity scenarios. To alleviate this issue, some recent studies are dedicated to controlling the length of summary (Kikuchi et al., 2016; Fan et al., 2018; Liu et al., 2018). However, as a surface-level feature of the summary, longer length does not equate to a higher degree of semantic coverage. In other words, the length limit can be easily satisfied by talking less/more details about the same event, but this is in contrast with the concept of summarization. Another research direction is query/aspect-based (Zhong et al., 2021; Hayashi et al., 2021; Ge et al., 2021) and interactive summarization (Shapira et al., 2017, 2021). Based on different queries, models can focus on different parts of the document and create summaries of various granularities. In practice, it requires a user to provide a query, implying that the user must have prior knowledge of the topic of the source text. Therefore, automatic granularity-aware summarization model is still an under-explored topic.

In this paper, we propose an unsupervised multi-granularity summarization framework called GRANUSUM. Unlike previous work based on supervised learning to provide guidance signals, such as salient sentences (Dou et al., 2021), keywords (He et al., 2020), and retrieved summaries (An et al., 2021), our approach does not rely on any manually labeled data. To measure the granularity, we first regard events as the basic semantic units of the input texts because events carry rich semantic information and are considered as informative representations in many NLP tasks (Zhang et al., 2020a; Li et al., 2020; Chen et al., 2021a). Overall, our system consists of two event-related components: Event-aware Summarizer and Event Selector. Specifically, given the document and randomly selected events in it as hints, we pre-train an abstractive Summarizer that can recover event-related passages. Furthermore, in an unsupervised manner, our Event Selector selects the events with high salience from the original text by candidate events pruning and ranking. Finally, through selecting different numbers of anchor events based on Event Selector, we can control the Summarizer to generate summaries containing different events, thus covering different numbers of semantic units of the original text. With our proposed approach, GRANUSUM becomes an unsupervised framework for multi-granularity summary generation.

To evaluate the multi-granularity summarization systems, we re-annotate DUC2004 (Dang, 2005) as the first benchmark in this direction (denoted as *GranuDUC*). Given multiple documents on the same topic, we annotate summaries at three levels of granularity with different semantic coverage. Also, to utilize the existing datasets for a supplement evaluation, we propose to divide several large-scale summarization datasets into buckets with summaries at different granularity levels to further evaluate the model performance. Experimentally, GRANUSUM surpasses strong summarization systems on all the multi-granularity evaluations. Additionally, we conduct conventional unsupervised abstractive summarization experiments on three typical benchmarks in different domains. Results demonstrate that GRANUSUM also substantially improves the previous state-of-the-art model under the traditional setting.

2 Related Work

2.1 Customized Summarization

In order to meet the needs of different users, existing neural summarization systems attempt to control customization of the summary, such as the aspects of content (Zhong et al., 2021; Hayashi et al., 2021), summary length (Christensen et al., 2014; Kikuchi et al., 2016; Liu et al., 2018) and writing style (An et al., 2021). Also, several studies seek to accommodate multiple types of preferences simultaneously to achieve customized summarization. Fan et al. (2018) additionally introduces different special marker tokens to the model to generate user-controllable summaries. He et al. (2020) allows for entity-centric, length-controllable, and question-guided summarization by adjusting the prompts, i.e., changing the textual input in the form of a set of keywords or descriptive prompt words. However, the unavailability of large-scale data containing customized summaries limits the development of these systems that rely on supervised learning. Thus, we focus on unsupervised approaches and are committed to solving the granularity aspect, which

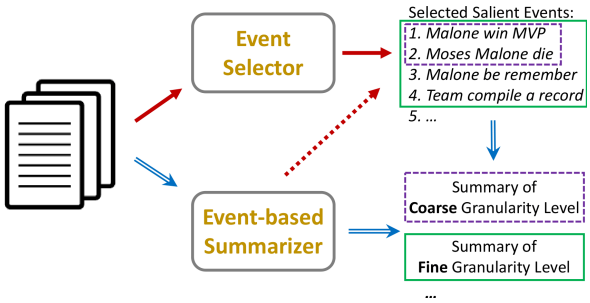

Figure 1: Overview of GRANUSUM. It consists of two components: Event Selector and Event-aware Summarizer. The red line (→) indicates that Selector extracts the salient events from the original text, and the dotted line means that Summarizer assists in this process. The blue line (⇒) denotes the multi-granularity summary generation process. By inputting different numbers of events as anchors (purple and green boxes), GRANUSUM can generate multi-granularity summaries.

remains an under-explored direction in customized summarization.

2.2 Unsupervised Summarization

In contrast to supervised learning, unsupervised models do not require any human-annotated summaries during training. Unsupervised summarization can also be divided into two branches: extractive methods and abstractive approaches. Most extractive methods rank the sentences and select the highest-ranked ones to form the summary. Specifically, they score sentences based on graph (Erkan and Radev, 2004; Hirao et al., 2013; Parveen et al., 2015), centrality (Zheng and Lapata, 2019; Liang et al., 2021), point-wise mutual information (Padmakumar and He, 2021), or sentence-level self-attention in pre-trained models (Xu et al., 2020). Another direction is unsupervised abstractive approaches, and these studies typically employ sequence-to-sequence auto-encoding method (Chu and Liu, 2019) with adversarial training and reinforcement learning (Wang and Lee, 2018). In addition, Yang et al. (2020) pre-train a Transformer model for unsupervised abstractive summarization by exploiting the lead bias phenomenon (See et al., 2017; Zhong et al., 2019a) in the news domain. In this work, our framework is an unsupervised abstractive framework, and can be further enhanced on top of the extractive method.

3 Multi-Granularity Framework

In this section, we first describe in detail our framework GRANUSUM, which has two major components: Event-aware Summarizer and Event Selector. Combining them enables multi-granularity generation. The overall framework can be seen in Figure 1. Then, we introduce the new human-annotated benchmark, GranuDUC, which can be used for multi-granularity evaluation.

3.1 Event-Aware Summarizer

In this work, we focus on abstractive summarization approaches. The way we make the model perceive the granularity is by inputting hints with different degrees of specificity, and here we format the hints as a sequence of events.

Event Extraction We follow previous work to define an event as a verb-centric phrase (Zhang et al., 2020a). A lightweight method[2] is utilized to extract events from open-domain unstructured data: we extract frequently-occurring syntactic patterns that contain verbs as events. On the basis of Zhang et al. (2020a), we extend a total of 76 syntactic patterns for matching events. For instance, the most common patterns contain n_1-nsubj-v_1 (e.g., *Hurricane hits*) and n_1-nsubj-v_1-dobj-n_2 (e.g., *Earthquake damages buildings*).[3] More details and concrete examples can be found in Appendix A.1.

Event-based Summarizer Pre-training Previous studies reveal that event information can be an effective building block for models to perform text generation (Daniel et al., 2003; Glavaš and Šnajder, 2014), so we attempt to obtain a Summarizer with the ability to generate event-related text in an unsupervised way. In the pre-training phase, it is trained to regenerate sentences based on a list of events and the remaining source text. Then we use it to generate a summary at inference time. Concretely, we pre-train a sequence-to-sequence model in the following steps:

1) randomly select a few sentences from the text,
2) extract events in these selected sentences,
3) mask these sentences in the source document,
4) take extracted events and unmasked text as input.

Then we use these selected sentences as the target for the model. For example, for a dialogue text as *"Do you have any plans tomorrow? How about playing basketball? Sure, I just finished my homework, it's time to exercise."*, we can select *How about playing basketball?* and extract the event

[2]Code for this part is available at: https://github.com/yzjiao/Open-vocabulary-event-extraction.

[3]nsubj and dobj indicate nominal subject and direct object. They are different relations between verbs and nouns.

play basketball. In this case, the specific format given to the model is:

- Input: play basketball ⟨seg⟩ Do you have any plans tomorrow? ⟨mask⟩ Sure, I just finished my homework, it's time to exercise.
- Target: How about playing basketball?

where ⟨seg⟩ is the segmentation token and ⟨mask⟩ indicates that a sentence at this position is masked. We use 'I' token to split the different events, and another example in news domain to further explain the four steps can be found in Appendix A.2.

3.2 Event Selector

The salience of the selected events determines whether the Summarizer can generate a quality summary or an irrelevant and uninformative paragraph. A long document can contain hundreds of events, and finding the best event subset involves an exponential search space. Therefore, it is crucial to have an Event Selector that selects the most important events in the text to feed to the Summarizer. Our event selector first reduces the search space by pruning out less salient events and sentences, and then ranks the remaining events using the pre-trained Summarizer.

Event Ranking The salience of the different events extracted from the documents varies. Some of the events are informative and relevant to the original text, but others are too general or specific. For instance, two events *club say* and *Malone be remember* can be extracted from the sentence *"The club said Malone will forever be remembered as a genuine icon and pillar in the Philadelphia 76ers team"*. The former is not important for this news, while the latter is indispensable. And in a sentence *"Malone won MVP awards by averaging 24.5 points and 15.3 rebounds"*, *"average 24.5 points and 15.3 rebounds"* is too detailed to be included in a high-level summary. Thus, ranking candidate events is a key function of Event Selector.

Inspired by Yuan et al. (2021), where a pre-trained generative model is capable of evaluating the correlation between the input and the target, we also use our pre-trained Event-based Summarizer to calculate the salience score for each event. Given the candidate event set E and the source document D, our Summarizer can generate a candidate summary c_E. Whenever an event e in the input is removed, if the generated candidate summary $c_{E\setminus\{e\}}$ differs greatly from c_E, this indicates that the removed event e is salient. As in the example above, removing *"club say"* does not cause an obstacle for the model to recover the sentence whose main meaning is that Malone is remembered by people, while removing *"Malone be remember"* makes the model unable to output the correct sentence. Thus, the latter should be the more important event. Formally, the **Salience Score** of event e can be defined as:

$$\mathrm{Sal}(e) \stackrel{\text{def}}{=} -\mathrm{Sim}(c_{E\setminus\{e\}}; c_E), \quad (1)$$

$$\mathrm{Sim}(x_1, x_2) \stackrel{\text{def}}{=} \mathrm{R1}(\mathrm{x}_1, \mathrm{x}_2) + \mathrm{R2}(\mathrm{x}_1, \mathrm{x}_2), \quad (2)$$

where $\mathrm{Sim}(x_1, x_2)$ is a function based on ROUGE score (Lin, 2004) to measure the similarity between any two text sequences x_1 and x_2. R1 and R2 are ROUGE-1 and ROUGE-2 scores, respectively. Based on the salience score, Event Selector can rank all the events in the candidate set. However, a single sentence may contain multiple events, so a long document can encompass hundreds of events. Using all events as a candidate set leads to unaffordable computational consumption. Therefore, we prune the candidate events before ranking them.

Candidate Pruning We expect to capture a small set of events that are relevant to the main topic while pruning redundant parts. Events with high relevance provide an efficient summary of the central points in the original text, while low redundancy ensures that the final summary is concise. To this end, we first select several salient sentences and extract the events in them as a candidate set. For relevance, if a sentence has a high semantic overlap with other input sentences, it should have a higher centrality and a higher probability to be included in the summary (Padmakumar and He, 2021). Thus, we define the **Relevance Score** of each sentence as:

$$\mathrm{Rel}(s, D) \stackrel{\text{def}}{=} \mathrm{Sim}(s; D \setminus \{s\}), \quad (3)$$

where s means the sentence and D represents the given document. $D\setminus\{s\}$ indicates that the sentence s is removed from the original text D.

For redundancy, the sentences in the summary should contain low redundant information when compared with each other. So when extracting the k-th sentence, we define its **Redundancy Score** as follows:

$$\mathrm{Red}(s, S) \stackrel{\text{def}}{=} \sum_{i=1}^{k-1} \mathrm{Sim}(s_i; s), \quad (4)$$

where S is a set of the k-1 sentences in the summary so far. We follow the idea of Maximal Marginal Relevance (Carbonell and Goldstein, 1998) to maximize relevance and minimize redundancy to calculate the **Importance Score** of each sentence as:

$$\mathrm{Imp(s, S, D)} = \lambda_1 \mathrm{Rel}(s, D) - \lambda_2 \mathrm{Red}(s, S). \quad (5)$$

Through iteratively calculating the score of each sentence, we can eventually obtain a fixed number of sentences and extract the events from them as a candidate set.

3.3 Multi-Granularity Summary Generation

With Event-aware Summarizer and Event Selector, it is feasible to generate multi-granularity summaries. By taking different numbers of ranked events as hints, the Summarizer can perceive the specific level of semantic coverage required to enable the generation of different summaries. For example, the Summarizer can generate a concise coarse-grained summary when only the two events with the highest salience scores (see Equation 1) are input. A case study to illustrate the overall flow of the multi-granularity summary generation can be found in Appendix A.4. During inference, instead of using the same setting as Zhang et al. (2020c), i.e., placing the ⟨mask⟩ token at the beginning of the article, we simply omit it. Because we already provide enough event information to guide the model to generate a summary in our framework.

3.4 New Benchmark: GranuDUC

Considering that there is no dataset for evaluating multi-granularity summarization models, we re-annotate a new benchmark called GranuDUC on the basis of DUC2004 (Dang, 2005). Our annotation team consists of 5 graduate students in NLP or people with equivalent expertise. For each document cluster, annotators are required to read multiple source documents and write summaries at three different granularities. The annotators are informed to be aware that granularity is not distinguished by the number of sentences, but is defined by different semantic coverage of the original text. Specifically, we inform the annotators that "coarse granularity level" should include only the main event of the entire documents, "medium granularity level" should include several important conditions, results and processes surrounding the main topic, and "fine granularity level" should further include the details such as time and location for each sub-event. Summaries at different granularities require significantly different levels of semantic coverage. Newly annotated sentences are allowed to be copied or rewritten from DUC2004's original reference summaries. In addition, we require annotators not to use the same sentences in different summaries of a sample, even when describing the same event. Each annotated summary is required to be reviewed by another annotator, then these two people discuss and revise until an agreement is reached. In the end, GranuDUC contains a total of 50 clusters, each cluster contains an average of 10 related documents and 3 summaries of different granularity, ranging from 10 words to more than 200 words in length. To demonstrate the quality of GranuDUC, we include the annotations of two samples in Appendix 8.

4 Experiments

We design three settings of experiments:
1) experiments on GranuDUC,
2) bucket-based evaluation,
3) unsupervised abstractive summarization.
The first two settings constitute a new testbed for multi-granularity summarization, where bucket means that we divide the existing dataset into different buckets according to semantic coverage to make the evaluation more comprehensive. In addition to this scenario, the last experiment auxiliarily evaluates the quality of summaries generated by our framework under the conventional unsupervised abstractive summarization setting.

4.1 Experimental Setup

Datasets Because the conclusions obtained on the summarization dataset of a single domain are not generalizable (Wang et al., 2019; Zhong et al., 2019b; Chen et al., 2020), we select two widely varying domains: news and scientific papers for our experiments Notably, we focus on two types of datasets, multi-document and long-document summarization, which are two main scenarios where users call for a multi-granularity system. For multi-document summarization, we concatenate the multiple articles into a single sequence as the source text. In addition to our benchmark GranuDUC, we use the following three datasets. Detailed statistics are listed in Table 2.

Multi-News (Fabbri et al., 2019) is a large-scale multi-document summarization dataset in the news domain. We use it in bucket-based evaluation (Sec-

Datasets	# Samples	Len. of Doc.	Len. of Sum.
Multi-News	56K	1793	217
arXiV	214K	6021	272
DUC2004	50	5882	115
GranuDUC	50	5882	24/68/135

Table 2: Statistics of all datasets we used in this paper. DUC2004 and GranuDUC are for testing only.

tion 4.2.2) and unsupervised summarization experiments (Section 4.3).

DUC2004 (Dang, 2005) contains 50 clusters, each with 10 relevant news articles and 4 reference summaries written by humans. Due to its small size, it is usually used directly as a test set. We utilize it in the unsupervised summarization experiment (Section 4.3).

arXiv (Cohan et al., 2018) is a collection of long documents derived from scientific papers. It takes the full text of the paper as input, and the corresponding abstract as the reference summary. We use it in the unsupervised summarization experiment (Section 4.3).

Implementation Details To process long input text in Table 2, we choose the Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) as our backbone model, and train it with typical cross entropy loss. For Multi-News and arXiv, we further pre-train LED with our event-related generation task on their training corpora (without using reference summaries) for a total of 10,000 and 30,000 steps, respectively. We set batch size to 32 and the maximum learning rate to 2e-5. λ_1 in the importance score is 1.0 and λ_2 is 0.4. By tuning the hyper-parameters on the validation set, we empirically extract 9 sentences for Multi-News and 4 sentences for arXiv to form a candidate set, and input 90% events according to salience score to the Summarizer under unsupervised summarization setting. For DUC2004 and GranuDUC, we test directly with the Summarizer pre-trained on Multi-News, since these datasets are both in the news domain. In all experiments, we use standard pyrouge[4] to calculate ROUGE scores. Due to the limitation of computational resources, we truncate an input text to 3,072 tokens for LED models.

Baselines We use the following baselines:

BART (Lewis et al., 2020) is the state-of-the-art sequence-to-sequence pre-trained model for various generation tasks, including abstractive dialogue generation, question answering, and text summarization. We use BART-large in all the experiments.

PEGASUS (Zhang et al., 2020b) is a powerful generation model with gap-sentences generation as a pretraining objective tailored for abstractive summarization. We use the large version of PEGASUS for comparison.

PEGASUS-event indicates that on top of PEGASUS, additional event information is prepended to the input before the ⟨mask⟩ token. We compare it to see if additional event information can be captured without our event-aware pre-training stage.

LED (Beltagy et al., 2020) has the same architecture as BART, except that the attention in the encoder introduces additional local attention and extends the position embedding to 16K tokens by copying the original embedding. The parameters in the LED are initialized by the weights in BART.

LED-Length-Control (LED-LC) is a baseline that we obtained by further pre-training LED. Inspired by Fan et al. (2018), given a document and the desired number of sentences k, we randomly place k sentences in the document with the ⟨mask⟩ token, and let the model recover these sentences. During inference, we input the text and the desired number of sentences as a hint to the model so that it can control the length of the output summary.[5]

PRIMERA (Xiao et al., 2022) is a pre-trained model for multi-document summarization that reduces the need for dataset-specific architectures and extensive labeled data. It achieves state-of-the-art results on multi-document summarization datasets under multiple settings.

4.2 Multi-granularity Evaluation

The first testbed we built for multi-granularity summarization includes two evaluation methods:

1) To test the ability of the model to generate summaries with different granularity levels when given the same input, we evaluate different models on our benchmark GranuDUC.

2) To supplement the limited size of GranuDUC, we design a bucket-based evaluation approach, where we divide a large-scale test set into different buckets based on their granularity levels, and test the ability of models to generate quality summaries in different granularity buckets.

[4] pypi.python.org/pypi/pyrouge/0.1.3

[5] If we need a two-sentence summary, the input format is: "⟨2⟩ ⟨seg⟩ ⟨mask⟩ source documents". It is exactly the same as GRANUSUM in terms of the training details and data.

	Coarse Granularity Level			Medium Granularity Level			Fine Granularity Level		
Model	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
PEGASUS	20.74	4.20	15.11	24.86	4.39	14.34	29.79	5.70	14.83
PEGASUS-event	20.68	4.18	15.12	24.72	4.28	14.25	29.58	5.52	14.61
LED-LC	21.83	4.80	15.29	26.73	5.59	15.76	30.18	5.57	15.24
GRANUSUM	**23.61**	**6.60**	**17.12**	**29.69**	**6.84**	**16.23**	**34.71**	**7.49**	**17.42**
Model	**Flu.**	**Rel.**	**Faith.**	**Flu.**	**Rel.**	**Faith.**	**Flu.**	**Rel.**	**Faith.**
PEGASUS	3.25	3.36	3.15	3.46	3.49	2.72	3.73	3.44	2.58
LED-LC	3.97	3.39	3.08	3.93	3.57	3.14	3.67	3.62	2.73
GRANUSUM	**4.13**	**3.82**	**3.59**	**4.09**	**3.78**	**3.46**	**3.82**	**4.05**	**3.17**

Table 3: Results on GranuDUC. The top half of the Table shows the result of the automatic metric ROUGE, and the bottom half presents the result of human evaluation, including fluency, relevance and faithfulness.

Model	Low			Medium			High		
	R-1	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**	**R-1**	**R-2**	**R-L**
PRIMERA	37.21	9.92	17.68	42.50	13.19	**20.24**	46.95	18.10	23.99
LED-LC	37.28	9.56	16.64	42.37	12.65	19.15	47.57	17.88	22.40
GRANUSUM	**38.19**	**10.27**	**18.07**	**44.73**	**14.12**	20.10	**50.23**	**19.62**	**24.11**
- Ranking	37.34	9.36	16.69	43.41	13.28	19.12	49.66	19.35	23.37

Table 4: Result of bucket-based evaluation on Multi-news. We design Granularity Score to divide the test set into three buckets. Low means that the summary has low semantic coverage with the source documents.

4.2.1 Results on GranuDUC

The summaries of each sample in GranuDUC can be divided into three granularity levels, where coarse granularity level represents the most compact summary, and fine granularity level is the most fine-grained summary. We use automatic metrics ROUGE and perform the human evaluation to evaluate the performance of different models in GranuDUC. Notably, both LED-LC and GRANUSUM have the ability to adjust the output according to specific granularity scenarios. At three different granularity levels on GranuDUC, we let LED-LC output 1, 3 and 8 sentences which correspond to the average length of reference summaries at different granularities. For our model, we take the top 90% events with the highest salience score in the selected 1, 3, 8 sentences as the input hint. For all baselines, we control the length of the model output to be similar to the reference summary to get the best performance.

Automatic Evaluation As illustrated in Table 3, compared to PEGASUS, LED-LC can bring a certain degree of improvement due to the ability to control the length of the output summary. This improvement is not remarkable at fine granularity level. For coarse and medium granularity levels, LED-LC can control the number of output sentences, while PEGASUS does not have a similar capability and it can only generate shorter summaries by truncating the output (to 32 and 64 words), which leads to performance degradation. On the other hand, GRANUSUM exceeds LED-LC and PEGASUS by a large margin in all the granularity levels. Although GRANUSUM and LED-LC are trained on the same data, GRANUSUM increases the R-1 score by 1.78 at coarse granularity level (21.83→23.61), and the improvement reaches to 4.53 at fine granularity level (30.18→34.71). With the benefit of event information, our model can generate more relevant and quality summaries, and the advantage is more pronounced in fine-grained summaries. Therefore, GRANUDUC is a more suitable system for multi-granularity scenarios than existing controllable summarization models.

Human Evaluation We also conduct human evaluation to have a more comprehensive understanding of the model output. Six graduate students are involved in this process to score the generated summaries from three different perspectives: fluency, relevance and faithfulness to the source documents. The score range is 1-5, with 1 being the worst and 5 the best. Each sample requires two people to discuss and agree on the scoring. According to the fluency scores in Table 3, both LED-LC and GRANUDUC can generate coherent sentences, while PEGASUS performs poorly in coarse and

Model	Multi-News			arXiv			DUC2004		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
LEAD	42.9	14.3	19.2	32.7	8.1	17.5	32.3	6.5	16.3
LED	17.3	3.7	10.4	15.0	3.1	10.8	16.6	3.0	12.0
BART	27.3	6.2	15.1	29.2	7.5	16.9	24.1	4.0	15.3
PEGASUS	32.0	10.1	16.7	29.5	7.9	17.1	32.7	7.4	17.6
PEGASUS-event	31.5	10.2	15.8	29.2	7.7	17.0	31.8	7.1	16.9
PRIMERA	42.2	13.7	**20.6**	34.6	9.4	18.3	34.7	6.9	17.6
Selector	43.3	14.1	19.1	35.3	10.8	17.8	34.3	7.1	17.1
LED-LC	42.0	13.3	19.2	34.9	9.9	18.1	33.9	6.6	16.8
GRANUSUM	**43.7**	**14.2**	20.1	**36.0**	**11.3**	**18.6**	**34.8**	**7.3**	**17.9**
- Ranking	43.5	14.0	19.7	35.4	10.8	18.5	34.3	7.0	17.2

Table 5: Results of unsupervised abstractive summarization on three datasets.

medium granularity levels due to truncating the output to a fixed length. From the perspective of relevance and faithfulness, a clear trend is that the more fine-grained the summary, the more relevant it is to the original text and the more likely it is to contain factual errors. Specific to the models, GRANUSUM generates more relevant and faithful summaries in all granularity scenarios compared to other baselines by exploiting event information.

4.2.2 Bucket-based Evaluation

In addition to GranuDUC, we seek to utilize existing large-scale datasets for multi-granularity evaluation. Unlike the previous approach of using a single reference summary to evaluate multiple lengths of summaries (Shapira et al., 2018), we divide the reference summaries into different buckets based on semantic coverage and then compare the performance of each model in each bucket. We first design a metric to calculate the granularity score between the source document and the reference summary to categorize the different samples. Because the same events in original text and human-written summary may have different descriptions, we design a granularity score on the basis of BERTScore (Zhang et al., 2019) to perform soft matching due to its ability to measure semantic coverage between two sequences. Specifically, we extract all the events in the source document and the reference summary as two event sequences, and calculate **Granularity Score** as:

$$\text{Granu}(D, r) = f(Event_D, Event_r), \quad (6)$$

where D is the source documents and r represents the reference summary. $Event_D$ denotes that we extract all events from D by using the approach in Section 3.1, and concatenate them into an event sequence. f means that BERTScore is used to calculate the recall score between two event sequences. Intuitively, a high recall score of the reference summary to the original text indicates that it has high semantic coverage and thus it is a summary at a high granularity level. We sort all samples in the test set of Multi-News dataset according to Granularity Score and divide them into three buckets with the same number of samples. The average length of summaries in the three buckets are 198, 214, and 236 words, respectively.

Although PRIMERA is the state-of-the-art model, it does not have the flexibility to change the output in response to different buckets. For LED-LC, we let the model generate 7, 8, and 9 sentences in low, medium, and high buckets, respectively. For our model, we take the top 70%, 80%, and 90% of the events with the higher salience score (see Section 3.2) in 9 selected sentences as the input for three different buckets. As shown in Table 4, LED-LC has no significant benefits over PRIMERA, indicating that controlling the output length and ignoring its connection to the original text is not a good solution for the multi-granularity system. In contrast, GRANUSUM achieves substantial improvements in all buckets compared to powerful baselines. In particular, in buckets with high semantic coverage, our model improves R-1 score by 3.28 compared to PRIMERA. Also, "-Ranking" means that we no longer filter out events based on the salience score, which causes a performance drop. It confirms that our selector can indeed exclude irrelevant and redundant events and thus improve the quality of the generated summary.

4.3 Unsupervised Abstractive Summarization

The quality of the summary is a key factor for all summarization systems. So in addition to the

multi-granularity scenario, we likewise compare GRANUSUM with conventional unsupervised abstractive summarization models. Table 5 provides results on three datasets. The first section includes a simple yet effective approach LEAD, which refers to extracting the first few sentences at the beginning of the text as a summary. It is a strong baseline in the news domain due to the lead bias problem (See et al., 2017; Zhong et al., 2019a). The second section lists the strong baselines and the last section contains the results of our models. Selector indicates that we extract several sentences from the source document based on our importance score described in Section 3.2 as the summary.

Surprisingly, although GRANUSUM is not specially designed for the conventional unsupervised summarization task, it still beats all the competitors and achieves new state-of-the-art results on most metrics across datasets. Despite inputting the same hints, PEGASUS-event does not show the ability to exploit event information and even performs worse than PEGASUS. In contrast, our pre-trained Event-aware Summarizer incorporates event information well into the generated summaries and thus boosts performance. Furthermore, GRANUSUM outperforms Selector, which is a strong extractive baseline, and extractive approaches usually dominate unsupervised summarization tasks. We think the improvement comes from two reasons:

1) In the pre-training stage, important content in the masked sentences is easier to reconstruct due to the redundancy of input texts. Thus, GRANUSUM learn to filter those unimportant content in inference, generating more concise summaries.

2) Event Selector screens out less critical events which should not appear in the summary.

Overall, GRANUSUM improves R-1 score by 1.0 on average compared to the previous best results, indicating that it is sufficient to generate quality summaries besides the multi-granularity ability.

5 Conclusion

In this paper, we highlight the importance of multi-granularity summarization systems in catering to user preferences and applying them to real-world scenarios. To facilitate research in this direction, we propose the first unsupervised multi-granularity summarization framework GRANUSUM and build a well-established testbed. Experiments demonstrate the effectiveness of our framework.

Limitations

We state the limitations of this paper from the following four aspects:

1) Unlike previous work that uses summary length to approximate granularity, we adopt an event-based definition, which can be extended to be more flexible. For example, introducing phrases, entities, relationships, etc. as part of the granularity may be a feasible way to further enhance the granularity-aware summarization system.

2) Despite being the first multi-granularity summarization benchmark, GranuDUC can only be used as a test set due to its small size. Thus, we call for the emergence of customized summarization datasets, which can greatly facilitate the development of customizable summarization models.

3) Specific to the method, we extract events from the source text as hints, which may reduce the abstractness of the generated summaries to some extent. In pursuit of a more abstractive summary, rephrasing events into different forms may be a viable option, and we leave it as future work.

4) In this paper we focus on three different levels of granularity and take document clusters containing thousands of words as input. A promising extension could be to input longer text and to add finer levels of granularity, for example, to generate summaries for an entire book (e.g., a novel) at multiple granularities.

Acknowledgements

We thank Wen Xiao for providing the output of PRIMERA. We would also like to thank anonymous reviewers for valuable comments and suggestions. Research was supported in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and INCAS Program No. HR001121C0165, National Science Foundation IIS-19-56151, IIS-17-41317, and IIS 17-04532, and the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, and the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Chenxin An, Ming Zhong, Zhichao Geng, Jianqiang Yang, and Xipeng Qiu. 2021. Retrievalsum: A retrieval enhanced framework for abstractive summarization. *arXiv preprint arXiv:2109.07943*.

Chenxin An, Ming Zhong, Zhiyong Wu, Qin Zhu, Xuan-Jing Huang, and Xipeng Qiu. 2022. Colo: A contrastive learning based re-ranking framework for one-stage summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5783–5793.

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

Jaime G. Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 335–336. ACM.

Muhao Chen, Hongming Zhang, Qiang Ning, Manling Li, Heng Ji, Kathleen McKeown, and Dan Roth. 2021a. Event-centric natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 6–14.

Yiran Chen, Pengfei Liu, Ming Zhong, Zi-Yi Dou, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2020. Cdevalsumm: An empirical study of cross-dataset evaluation for neural summarization systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3679–3691.

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

Janara Christensen, Stephen Soderland, Gagan Bansal, et al. 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.

Eric Chu and Peter Liu. 2019. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12.

Naomi Daniel, Dragomir Radev, and Timothy Allison. 2003. Sub-event based multi-document summarization. In *Proceedings of the HLT-NAACL 03 Text Summarization Workshop*, pages 9–16.

Alberto Díaz and Pablo Gervás. 2007. User-model based personalized summarization. *Information Processing & Management*, 43(6):1715–1734.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. Gsum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

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

Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54.

Suyu Ge, Jiaxin Huang, Yu Meng, Sharon Wang, and Jiawei Han. 2021. Fine-grained opinion summarization with minimal supervision. *arXiv preprint arXiv:2110.08845*.

Goran Glavaš and Jan Šnajder. 2014. Event graphs for information retrieval and multi-document summarization. *Expert systems with applications*, 41(15):6904–6916.

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.

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

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1515–1520.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338.

Kevin Lerman, Sasha Blair-Goldensohn, and Ryan McDonald. 2009. Sentiment summarization: evaluating and learning user preferences. In *Proceedings of the 12th conference of the European chapter of the ACL (EACL 2009)*, pages 514–522.

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.

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020. Connecting the dots: Event graph schema induction with path language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 684–695.

Xinnian Liang, Shuangzhi Wu, Mu Li, and Zhoujun Li. 2021. Improving unsupervised extractive summarization with facet-aware modeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1685–1697.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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 3721–3731.

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.

Yizhu Liu, Zhiyi Luo, and Kenny Zhu. 2018. Controlling length in abstractive summarization using a convolutional neural network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4110–4119.

Rutu Mulkar-Mehta, Jerry R Hobbs, and Eduard Hovy. 2011. Granularity in natural language discourse. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Vishakh Padmakumar and He He. 2021. Unsupervised extractive summarization using pointwise mutual information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2505–2512.

Daraksha Parveen, Hans-Martin Ramsl, and Michael Strube. 2015. Topical coherence for graph-based extractive summarization. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1949–1954.

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)*, volume 1, pages 1073–1083.

Ori Shapira, David Gabay, Hadar Ronen, Judit Bar-Ilan, Yael Amsterdamer, Ani Nenkova, and Ido Dagan. 2018. Evaluating multiple system summary lengths: A case study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 774–778.

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.

Ori Shapira, Hadar Ronen, Meni Adler, Yael Amsterdamer, Judit Bar-Ilan, and Ido Dagan. 2017. Interactive abstractive summarization for event news tweets. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 109–114.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuan-Jing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6209–6219.

Danqing Wang, Pengfei Liu, Ming Zhong, Jie Fu, Xipeng Qiu, and Xuanjing Huang. 2019. Exploring domain shift in extractive text summarization. *arXiv preprint arXiv:1908.11664*.

Yaushian Wang and Hung-Yi Lee. 2018. Learning to encode text as human-readable summaries using generative adversarial networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4187–4195.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. Primera: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263.

Shusheng Xu, Xingxing Zhang, Yi Wu, Furu Wei, and Ming Zhou. 2020. Unsupervised extractive summarization by pre-training hierarchical transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1784–1795.

Rui Yan, Jian-Yun Nie, and Xiaoming Li. 2011. Summarize what you are interested in: An optimization framework for interactive personalized summarization. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1342–1351.

Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. 2020. Ted: A pretrained unsupervised summarization model with theme modeling and denoising. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1865–1874.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *arXiv preprint arXiv:2106.11520*.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Wingki Leung. 2020a. Aser: A large-scale eventuality knowledge graph. In *The Web Conference 2020-Proceedings of the World Wide Web Conference, WWW 2020*, page 201.

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020c. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Hao Zheng and Mirella Lapata. 2019. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247.

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, ACL 2020, Online, July 5-10, 2020*, pages 6197–6208. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2019a. Searching for effective neural extractive summarization: What works and what's next. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1058.

Ming Zhong, Danqing Wang, Pengfei Liu, Xipeng Qiu, and Xuan-Jing Huang. 2019b. A closer look at data bias in neural extractive summarization models. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 80–89.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 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.

A Method

Here we provide more details about our method part. The workflow of GRANUSUM and case study are listed in Table 7.

A.1 Event Extraction

Specifically, given a sentence s, we use a dependency parser to obtain its dependency parse tree and select all non-auxiliary verbs as centric tokens. Then, along the syntactic relationships between the selected verbs and other tokens, we extract the longest phrase that matches the designed patterns as events. As illustrated in Table 6, the most frequent pattern is n_1-nsubj-v_1, such as *Hurricane hit*. Another common pattern is n_1-nsubj-v_1-dobj-n_2, like *Hurricane damage buildings*. Here "nsubj" denotes an active relationship between nouns and verbs, while "nsubjpass" in another example represents a passive relationship between them. More detailed examples can be found in Table 7, we extract events from four selected sentences, and the colored text shows the locations of the events in the original document.

A.2 Event-based Summarizer Pre-training

We further explain the four steps of Event-based Summarizer pre-training with the help of the following example. For a paragraph of news as *"Honduras braced for potential catastrophe Tuesday. Hurricane Mitch roared through the Caribbean, churning up high waves and intense rain that sent coastal residents scurrying for safer ground. President declared a state of maximum alert and the Honduran military sent planes to pluck residents from their homes on islands near the coast"*, we

1) first randomly select a sentence: *"Hurricane Mitch roared through the Caribbean, churning up high waves and intense rain that sent coastal residents scurrying for safer ground"*,

2) extract events in it such as *Mitch roar*, *Mitch churn up wave and rain*, *send* and *resident scurry*,

3) then mask this sentence in the original paragraph, and finally

4) use extracted events and masked text as the input and regard the selected sentence as the target as follows:

- Input: Mitch roar | Mitch churn up wave and rain | send | resident scurry ⟨seg⟩ Honduras braced for potential catastrophe Tuesday. ⟨mask⟩ President declared a state of maximum alert and the Honduran military sent planes to pluck residents from their homes on islands near the coast.
- Target: Hurricane Mitch roared through the Caribbean, churning up high waves and intense rain that sent coastal residents scurrying for safer ground.

In our experiments, we randomly mask 1 to n sentences from a document, which leads to n samples to pre-train our Summarizer. Here we set n to the smaller of a constant number 10 and one-third of the number of sentences in the document.

A.3 Event Selector

We use the example in Table 7 to further explain the flow of the Event Selector. When we obtain candidate events from selected sentences, there are still different types of issues in the candidate set. Some generic and uninformative events, such as "*club say*" and "*let him know*", should have a lower priority for a summary. Although we introduce sentence-level redundancy score in the pruning step, as a finer-grained unit, events still suffer from redundancy problem (see events in Table 7 with the same color), e.g., both "win MVP", "Malone win MVP" and "average 31.1 points and 14.7 rebounds", "average 24.5 points and 15.3 rebounds" appear in the candidate set. However, after the events ranking and filter using our Event Selector, all of these issues are alleviated. In this case, our Selector regards "Malone win MVP", "Moses Malone die" and "Malone be remember" as the three most salient events, which is consistent with the original news. In addition, uninformative events ("*club say*" and "*let him know*") are ranked at the end of the candidate sets, and duplicate events ("win MVP" and "average 24.5 points and 15.3 rebounds") are filtered out due to the lowest salience score. In general, the reasonable ranking of candidate events by the Selector plays a crucial role in improving the quality of subsequent multi-granularity summaries.

A.4 Multi-Granularity Summary Generation

We can see from Table 7, to obtain the most condensed summary, the two most important events ("Malone win MVP" and "Moses Malone die") and the original news are fed to the model. Then, the pre-trained Summarizer can be aware of event-based cues and generate the corresponding sentence: "Moses Malone, a three-time NBA MVP and one of basketball's most ferocious rebounders,

Patterns	Examples
n_1-$nsubj$-v_1	Hurricane hit
n_1-$nsubj$-v_1-$dobj$-n_2	Hurricane damage buildings
n_1-$nsubj$-v_1-$xcomp$-a	People feel scared
n_1-$nsubj$-v_1-$xcomp$-v_2-$dobj$-n_2	Police want to save people
n_1-$nsubjpass$-v_1	Residents are injured

Table 6: Five typical patterns and corresponding examples when we extract events (76 patterns in total). Here 'v' is a verb, 'n' stands for a noun, and 'a' denotes an adjective. All verbs remain in their original form. 'nsubj', 'dobj', 'xcomp', and 'nsubjpass' are syntactic relations.

died on Sunday". As more events are input, our Summarizer also has the ability to adjust the order of the narrative to make the content more logical. In the summary of granularity level 2, the order in the prompt is "Malone be remember" then "team compile a 65-17 record", but the model first output "He helped the team compile a 65-17 record in the first season" and then "These achievements make him be remembered as a genuine icon and pillar in the history of 76ers basketball" to make the whole summary more coherent and intuitive. Compared to sentences selected from the source documents (see Step 1 in Table 7), the summary generated by GranuSum omits unimportant details and paraphrases to make it more concise. Abstractive models without guidance signals, such as PEGASUS, tend to generate some repetitive sentences (the first two sentences), and generate several less relevant sentences without capturing important events. In contrast, GRANUSUM can output summaries that are more relevant and faithful to the original text.

B Examples for GranuDUC

We provide two annotation examples for our proposed GranuDUC benchmark in Table 8.

Step 1: Select Important Sentences based on Relevance and Redundancy Score, and Extract Events

- Malone was part of the 76ers' 1983 NBA championship team, and the club said he will forever be remembered as a genuine icon and pillar of the most storied era in the history of Philadelphia 76ers basketball. ⟶ club say | Malone be remember
- In the initial meeting in New York, Cunningham pulled Malone aside and let him know his expectations of the player who had won MVP honors in Houston the previous season by averaging 31.1 points and 14.7 rebounds. ⟶ Cunningham pull Malone | let hime know | win MVP | average 31.1 points and 14.7 rebounds
- In his first season with the Sixers, Malone won MVP awards by averaging 24.5 points and 15.3 rebounds during the regular season in which the team compiled a 65-17 record. ⟶ Malone win MVP | average 24.5 points and 15.3 rebounds | team compile a 65-17 record
- Moses Malone, a three-time NBA MVP and one of basketball's most ferocious rebounders, died Sunday, the Philadelphia 76ers said. ⟶ Moses Malone die | 76ers say

Step 2: Obtain a Candidate Set by Combining the Above Events

- Original Candidate Events: club say | Malone be remember | Cunningham pull Malone | let him know | win MVP | average 31.1 points and 14.7 rebounds | Malone win MVP | average 24.5 points and 15.3 rebounds | team compile a 65-17 record | Moses Malone die | 76ers say

Step 3: Event Ranking and Filtering (Event Selector)

- Ranked Candidate Events: Malone win MVP | Moses Malone die | Malone be remember | team compile a 65-17 record | Cunningham pull Malone | average 31.1 points and 14.7 rebounds | 76ers say | let him know

Step 4: Multi-Granularity Summary Generation (Event-based Summarizer)

- Coarse Granularity Level
  - Input: Malone win MVP | Moses Malone die ⟨seg⟩ ⟨mask⟩ Source News
  - Generated Summary: Moses Malone, a three-time NBA MVP and one of basketball's most ferocious rebounders, died on Sunday.
- Fine Granularity Level
  - Input: Malone win MVP | Moses Malone die | Malone be remember | team compile a 65-17 record ⟨seg⟩ ⟨mask⟩ Source News
  - Generated Summary: Moses Malone, a three-time NBA MVP and one of basketball's most ferocious rebounders, died on Sunday. He helped the team compile a 65-17 record in the first season. These achievements make him be remembered as a genuine icon and pillar in the history of 76ers basketball.

Summary Generated by PEGASUS

- Moses Malone, a three-time NBA MVP and one of basketball's most ferocious rebounders, died Sunday, the Philadelphia 76ers said. The 76ers issued a statement that said Malone had died. Malone was inducted into the Naismith Memorial Basketball Hall of Fame in 2001 and attended the induction ceremonies for the year's class in Springfield, Massachusetts this weekend.

Reference Summary

- Three-time NBA MVP and Philadelphia 76ers legend Moses Malone, who with Julius Erving in 1983 brought the City of Brotherly Love its first championship since 1967, has died at the age of 60, reports the Inquirer. Moses holds a special place in our hearts and will forever be remembered as a genuine icon and pillar of the most storied era in the history of Philadelphia 76ers basketball.

Table 7: Workflow of GRANUSUM and case study. The colored text in Step 1 indicates the location of the extracted event in the original sentence. Events of the same color in Step 2 are redundant. Underlined text in Step 4 represents the overlap with the reference summary. Notably, we pre-train an Event-based Summarizer before Step 1.

Sample 1: News about the Civil Suit against Microsoft

• **Summary of Coarse Granularity Level**: The Justice Department filed a civil suit against Microsoft to change its pattern of anti-competitive conduct on browser software.

• **Summary of Medium Granularity Level**: Business rivals have filed an anti-trust suit against Microsoft to break Microsoft Corp.'s monopoly on computer operating systems. The suit began with a Microsoft vs Netscape battle. The Government is examining Microsoft's financial records and painting a dark image of its Chairman Bill Gates. An unpublished book may be crucial to the trial.

• **Summary of Fine Granularity Level**: The Justice Department filed a suit against Microsoft for violation of the Sherman Act to change its anti-competitive conduct. The heart of the suit is the Internet browser battle between Microsoft and Netscape. Microsoft, it is argued, has told computer manufacturers that if they want Windows, they must forgo Netscape. Netscape complaint over browsers was central to the case, which grew to include Intel, IBM, Sun, Apple, AOL, and Intuit. The battle now extends far beyond that aiming at Microsoft's overall aggressive anti-competitive conduct. Microsoft's chairman, Bill Gates, usually seen as a visionary is portrayed in much darker tones in the trial. Microsoft was ordered to let Justice examine its records and sought a trial delay. An unpublished book provided evidence, which can be crucial to the trial.

Sample 2: News about the Health Condition of the Russian President

• **Summary of Coarse Granularity Level**: Russia President Boris Yeltsin's worsening heath condition caused great concern to the Russian leadership.

• **Summary of Medium Granularity Level**: During Russia President Boris Yeltsin's seven years in power, illness has often sidelined him. He recently cut short a trip to Central Asia because of a respiratory infection and he later canceled two out-of-country summits. Russia's leaders are calling for his resignation and question his legal right to seek reelection.

• **Summary of Fine Granularity Level**: Russia President Boris Yeltsin had a heart attack in 1996, followed by multiple bypass surgery. The cause of minor burns on his hand were not disclosed. On a trip to Uzbekistan he walked stiffly, stumbled, rambled and seemed confused. Ceremonies were canceled and the trip ended a day early. Yeltsin refuses to admit he is seriously ill and his condition is kept secret. He was treated with antibiotics and ordered to bed but went to the office anyway. Many Russians suspect he is sicker, question his ability to do his job, and want him to resign. The court was to judge on whether he could serve a third term, but he already has said he will not run.

Table 8: Annotation of two samples in GranuDUC.