VCSum: A Versatile Chinese Meeting Summarization Dataset

Han Wu[1,2], **Mingjie Zhan**[3,†], **Haochen Tan**[1,2,†], **Zhaohui Hou**[3,†], **Ding Liang**[3], **Linqi Song**[1,2]
[1] Department of Computer Science, City University of Hong Kong
[2] City University of Hong Kong Shenzhen Research Institute
[3] SenseTime Research
{hanwu32-c,haochetan-2}@my.cityu.edu.hk
{zhanmingjie,houzhaohui,liangding}@sensetime.com
linqi.song@cityu.edu.hk

Abstract

Compared to news and chat summarization, the development of meeting summarization is hugely decelerated by the limited data. To this end, we introduce a versatile Chinese meeting summarization dataset, dubbed VCSum, consisting of 239 real-life meetings, with a total duration of over 230 hours. We claim our dataset is *versatile* because we provide the annotations of topic segmentation, headlines, segmentation summaries, overall meeting summaries, and salient sentences for each meeting transcript. As such, the dataset can adapt to various summarization tasks or methods, including segmentation-based summarization, multi-granularity summarization and retrieval-then-generate summarization. Our analysis confirms the effectiveness and robustness of VCSum. We also provide a set of benchmark models regarding different downstream summarization tasks on VCSum to facilitate further research. The dataset and code will be released at https://github.com/hahahawu/VCSum.

Meeting Transcript about Fundamental Education
Speaker1: 我想请问一下，我们平时对于基础教育，包括现在因为就是只有哈工大有本科，所以哈工大是直接有对接这个本科高中生的这个需求的，在这方面有没有一些围绕着这个方面的分享。 *Speaker2*: 深圳这个教育和医疗是两个短板，尤其教育的话这个尤其是基础教育，现在高端教育慢慢这个步伐已经开始在加快了。但是在基础教育这块 ... *Speaker3*: 另一方面跟家长的这种这个观念也比较有关系。因为 ... [EOS]
Headline: 基础教育的短板 **Segmentation summary**: 目前基础教育是短板，处于未解决温饱问题阶段，学校重视不够。家长的观念也有影响，一直都在走应试教育的老路，最终目标都是在高考中去考高分 ...
Speaker1: 老师们讲得特别全面系统，而且把我们整个工作的衔接都联系起来了。接下来请 ... *Speaker4*: 我说一下我对人才培养的这个看法 ... [EOS]
Headline: 人才培养的方式 **Segmentation summary**: 要提供一个平台，为青少年的特长提供机会，同时还能培养特长，要从小发掘小朋友的特长, ... ，同时新技术新思想也要通过大城市蔓延到小城市。
...
Overall summary: 基础教育要多提供平台，重视基础教育的同时，也要多关注孩子的兴趣，学校要提供层次的选人标准，不要太局限在成绩上，也要有差异化 ...

Table 1: An example from our dataset. The green texts are the highlighted sentences. The token [EOS] is used to distinguish different segmentations. We provide multi-granularity summaries to a meeting transcript, including headline, segmentation summary and overall summary. See the English example in Appendix A.1.

1 Introduction

Meeting summarization (Janin et al., 2003; McCowan et al., 2005; Zhong et al., 2021) is the task of distilling the meeting transcript into a concise and readable summary that contains the most salient parts of a meeting. The summary can help the participants or absentees to quickly grape the highlight points. Therefore, a set of models have been proposed to comprehensively and succinctly summarize the content of a meeting (Zhu et al., 2020; Feng et al., 2021; Zhong et al., 2022b).

Compared to standard text summarization (Nallapati et al., 2016; Narayan et al., 2018), meeting summarization is a much more challenging task because it has more informal and oral expressions, topic shifts, multiple participants and longer context. For this reason, existing datasets for meeting summarization, i.e., AMI (McCowan et al., 2005) and ICSI (Janin et al., 2003), can hardly be used to train a robust summarization model owing to 1) their small size – AMI and ICSI only contain 137 and 59 pairs of meeting transcripts and summaries, respectively; 2) the specific domain – AMI just focuses on the product design and development while ICSI concentrates on the academic discussions; 3) coarse-grained summaries – the summaries in these two datasets are directly written for the whole meeting that might involve various topics. Although some larger dialogue summarization datasets have been created, e.g., SAMSum (Gliwa et al., 2019), DialogSUM (Chen et al., 2021) and

[†] Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2023, pages 6065–6079
July 9-14, 2023 ©2023 Association for Computational Linguistics

MediaSum (Zhu et al., 2021), their context and summaries are still much shorter than the meeting summarization data. On the other hand, some variants of the summarization task are recently studied in news summarization, such as segmentation-based summarization (Liu et al., 2022) aiming at jointly segmenting and summarizing the lengthy context, multi-granularity summarization (Zhong et al., 2022a) aiming at generating coarse-, middle- and fined-grained summaries at the same time, and *retrieval-then-generate* summarization (Mao et al., 2022) aiming at generating the summaries on the extracted salient sentences. While these variant tasks and methods have demonstrated their capacities on improving the news summarization performance, there is a huge demand for a large-scale *versatile* meeting summarization dataset to adapt these variants into the field.

To this end, we collect a **V**ersatile **C**hinese meeting **Sum**marization dataset based on the real-life meeting recordings, called VCSum. The dataset contains 239 meetings, with a total duration of over 230 hours, and each meeting transcript has over 14K tokens on average. To make the dataset versatile, we provide various annotations, including segmentation-based annotations, multi-granularity annotations and extractive annotations. Table 1 illustrates an example from our dataset. The meeting transcript is segmented into several sections according to the topics discussed. Then multi-granularity summaries are provided for each section, i.e., a coarse-grained headline summary with 5-20 words and a find-grained segmentation summary with 100-150 words. An overall summary with 200-250 words for the whole meeting is also annotated to formulate the challenging summarization tasks, such as lengthy meeting transcript summarization. Furthermore, we instruct annotators to highlight salient sentences in the meeting whose content is further verified to be highly consistent with the summaries. As such, our dataset is also a good testbed for the *extract-then-generate* summarization. In the experiment part, we evaluate several benchmark models on segmentation-based summarization, multi-granularity summarization, extract-then-generate summarization and highlight sentence extraction. We also provide a conversation solution to the dataset.

We summarize our contributions as follows: (1) we are the first one to collect a large and high-quality meeting summarization dataset from real-life videos in the last around 20 years; (2) we propose the first *versatile* summarization dataset that contains the annotations of extractive highlight sentences, topic segmentation, and multi-granularity summaries; (3) we conduct extensive experiments on the proposed dataset, constructing the benchmark to facilitate further research.

2 Related Work

2.1 Dialogue Summarization

Dialogue summarization aims to extract or summarize the most important information in the dialogue. Dialogue can take many forms, including chit-chat, meetings, and emails (Li et al., 2017; Zhang et al., 2021; Chen et al., 2021). A bunch of chat summarization datasets (Gliwa et al., 2019; Zhu et al., 2021; Chen et al., 2021) have been created yet. For example, the most widely used chat summarization dataset, i.e., SAMSum (Gliwa et al., 2019), is collected by linguists writing messenger-like conversations. This high-quality dataset greatly facilitates this research direction. A set of follow-up algorithms are proposed to solve the task by enhancing the dialogue context modeling(Chen and Yang, 2020; Zhong et al., 2022b) or dialogue participant modeling (Narayan et al., 2021). However, due to the much higher cost of collecting and annotating meetings, existing meeting summarization datasets are really limited, only AMI (McCowan et al., 2005) and ICSI (Janin et al., 2003), which were constructed around 20 years ago. The small size and inferior annotation quality of these two datasets are scarce to support the training of a robust meeting summarization model, while meeting summarization suffers more challenges than chat summarization, e.g., longer context and more topic shifts. Although some data augmentation techniques(Zhu et al., 2020; Zhong et al., 2022b) have been attempted, insufficient meeting summarization data remains a great barrier to this field moving forward. In this work, we create a high-quality and larger meeting summarization dataset to fill the gap of no new meeting summarization datasets proposed in the last 20 years.

2.2 Advanced Summarization Tasks

Thanks to the abundant data and simpler annotations, many variants of the standard summarization task have been attempted on news summarization, including segmentation-based summarization (Liu et al., 2022), multi-granularity summarization

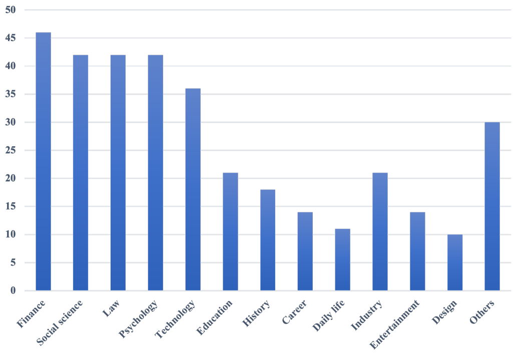


Figure 1: Topic distribution of VCSum.

(Zhong et al., 2022a). The task of segmentation-based summarization is proposed to address the problem of multiple topics discussed in the lengthy document. The multi-granularity summarization aims to provide summaries with different degrees of semantic coverage. Although these tasks could improve the summarization performance, they cannot be tested on meeting summarization due to a lack of data. In this work, we provide the meeting summarization data with segmentation-based and multi-granularity annotations and build the benchmarks of these tasks on meeting summarization. Besides, superior to SegNews dataset (Liu et al., 2022), VCSum focuses on real-life meetings and provides human-annotated summaries.

3 The VCSum Corpus

3.1 Data Selection

We collect the roundtable meetings from some Chinese video-sharing websites. We first obtain 1,419 videos from the websites by searching the keyword "圆桌会议(roundtable)". To select high-quality videos and alleviate potential ethical issues, we crowdsource the basic meeting information by asking the following questions: 1) are the audio and video of the meeting clear? 2) how many *valid* participants in the meeting? Participants are valid only when they have lots of expressions and clearly articulate their opinions. 3) is there any reference in the meeting? The reference can be the supporting materials of the meeting, such as slides, technical documentation, requirement document, reports, etc. 4) does the meeting involve any offensive or ethical content, including politics, religious issues, gender concerns, or violence?

Each meeting will be marked by two annotators, and the disagreements will be tackled by the third annotator. Then the candidate videos are selected by the conditions of 1) having clear audio and video; 2) the number of valid participants ranging from 1 to 10; 3) no references; 4) not involving any offensive or ethical concerns; and 5) the meeting duration ranging from 10 minutes to 200 minutes. We finally obtained 541 valid videos. The creation year of these videos ranges from 2017 to 2022. Owing to the limited budget for the annotation, we further selected 239 meetings, covering as many different domains as possible, including technology, finance, daily life and so on. We provide the topic distribution in Figure 1.

3.2 Data Annotation

The annotation is conducted on Feishu Minutes [1], an integrated platform for video parsing and automatic speech recognition (ASR). We upload candidate videos to the platform and parse them into meeting transcripts. Then annotators are asked to read the transcripts and provide five kinds of annotations.

Highlight sentences. Annotators should highlight the salient and informative sentences in the original meeting transcripts. The marked sentences must be fluent and complete. To avoid excessive labeling, the highlighted sentences should not exceed 10% of the entire transcript.

Topic segmentation. As a meeting generally involves multiple topics, we instruct annotators to identify different topic segments by inserting a special token [EOS] at the end of each segment. Note that the topic segmentation is annotated on the utterance level.

Segmentation headline. After obtaining topic segments, annotators should provide a headline to identify the topic of each segmentation. The word count for each headline should fall within the range of 5-20 words.

Segmentation summary. Annotators should write a summary for each segment. Different from headlines, segmentation summaries focus more on details of the content. Each segmentation summary should contain 100-150 words.

Overall summary. In the end, annotators should provide an overall summary that covers the most salient and informative content of the meeting. The word count for each overall summary should range from 200 to 250 words.

[1] https://meetings.feishu.cn/minutes

Dataset	Lan. style	Scenario	Domain	#Transcripts	#Tokens/trans.	#Turns/trans.	#Speakers/trans.	#Tokens/sum.
CNN	written	news	multiple	92465	654.0	-	-	42.1
DailyMail	written	news	multiple	219506	702.9	-	-	51.5
SAMSum	written	online	multiple	16369	93.8	11.2	2.4	20.3
DialogSum	spoken	-	multiple	13460	131.0	-	-	23.6
MediaSum	spoken	interview	multiple	463596	1553.7	30.0	6.5	14.4
CSDS	written (Chinese)	online	customer	10701	401.1	26.0	2.0	83.3
AMI	spoken	meeting	product	137	6007.7	535.6	4.0	296.6
ICSI	spoken	meeting	academia	59	13317.3	819.0	6.3	488.5
VCSUM	spoken (Chinese)	meeting	multiple	239	14106.9	73.1	5.6	231.9
VCSUM$_{seg}$*				1359	2480.9	12.9	3.0	139.1

Table 2: Comparison between VCSUM and other news, dialogue or meeting summarization datasets. * indicates the statistical results on segmentation transcripts. # stands for the average result.

VCSUM	#hl./trans.	#Token/hl.	R-1/2/L
Whole	71.7	32.5	88.65/50.46/62.56
Segmentation	12.7	32.5	81.17/49.22/67.54

Table 3: The highlight statistics of the whole meeting transcript and transcript segmentations. hl. stands for highlight sentences. R-1/2/L means the ROUGE-1/2/L recall score between the set of highlights and the corresponding summary.

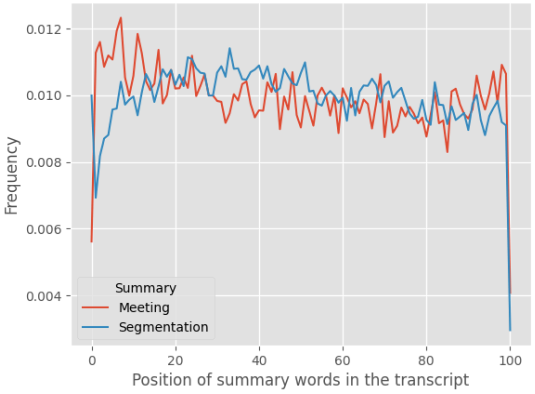


Figure 2: The frequency of non-stop summary words appearing at different positions of the transcript. The positions are normalized to [0, 100].

3.3 Quality Control

To obtain high-quality data, we release a set of data samples to select the most suitable and professional annotators. Finally, we recruit eight annotators majoring in law, finance or Chinese culture from the top universities in China (4 females and 4 males). Before the formal annotation, all annotators were asked to study the annotation protocols and practice on the training samples for a period of time. The annotation process began after all annotators passed our examination. We also have two annotation inspectors from our research group to monitor the whole process. During the annotation process, each sample is annotated by an annotator and checked by another annotator and an inspector. The annotation would be accepted only if both two checkers approved it. After the annotation, all results are further validated by ourselves. If any errors are found in an annotation batch, the corresponding annotator and checkers would be instructed to self-check and re-annotate the batch until the result meets our requirements. Additionally, to alleviate the error propagation from ASR, we manually compared several ASR technologies and finally selected Feishu Minutes. Our analysis of 100 randomly sampled meeting segmentations reveals a word error rate of 8%, which is significantly lower than the error rate of 30% reported in the AMI dataset. To further improve the data quality, we encourage annotators to revise any typos or grammatical errors they find during the annotation process.

3.4 The Characteristics of VCSUM

Table 2 illustrates the comparison between VCSUM and other news, dialogue or meeting summarization datasets. We see that VCSUM has much longer context and summaries compared to existing news and dialogue summarization datasets. The transcripts in VCSUM contain 72.8 turns and 5.7 speakers on average, suggesting that the dataset also has the characteristics of multi-turn and multi-party dialogues. While the traditional meeting datasets, like AMI and ICSI, only focus on a single domain, VCSUM involves multiple domains, ranging from daily life to academic discussions. Moreover, we also provide the extractive highlight sentences in VCSUM. Each meeting transcript averagely has 71.7 highlighted sentences, with each sentence containing 32.5 tokens. We calculate the ROUGE scores (Lin, 2004) between highlight sentences and the corresponding summary. The larger value of ROUGE-1 indicates that the written summaries are semantically consistent with the highlight annotations, while the smaller ROUGE-2 scores suggest the abstractiveness of the summary.

Previous study (Kedzie et al., 2018) reveals a critical problem of summarization task, called *positional bias*, which is that in existing datasets (McCowan et al., 2005; Chen et al., 2021; Zhu et al., 2021), most important information is often shown at the beginning of the context. This problem might bias the model to focus on the early words rather than the entire context. To this end, we also study the positional bias on our dataset. We evenly partition the transcript into 100 bins and count the frequency of the non-stop summary words appearing in each bin. As shown in Figure 2, the meeting transcripts contain more summary words near the beginning and end while segmentation transcripts hold more summary words in the middle part. However, the summary words of VCSUM are smoothly distributed in the transcripts. This observation indicates that our dataset does not suffer the positional bias, thus being a more challenging dataset.

4 Task Overview

Based on the annotation of our dataset, we propose three main challenging tasks. Given a meeting transcript $C = \{U_1, U_2, ..., U_N\}$ consisting of N utterances, wherein each utterance $U_i = \{w_1^i, w_2^i, ..., w_{|U_i|}^i\}$ is a sequence of words, we formulate the tasks as follows.

Highlight Sentence Extraction The task of highlight sentence extraction (HSE) aims to find a set of spans $H = \{(w_j^i, w_k^i)|0 < i \leq N, 0 < j < k \leq |U_i|\}$ that contain the most important information of the meeting. All highlight sentences are not overlapped and not across different utterances.

Segmentation-based Multi-granularity Summarization This task is essentially comprised of two sub-tasks, i.e., topic segmentation and multi-granularity summarization. Formally, we aim to generate three summaries at different granularities, i.e., a headline Y_h, a segmentation summary Y_s and a joint summary $Y_h\!:\!Y_s$, based on the transcript segments S, where $S \in \{(U_j, ..., U_k)|0 < j \leq k \leq N\}$. The segments are partitioned in the level of utterance, and there are no overlapping among different segments.

Abstractive Meeting Summarization The goal of this task is to generate the overall summary Y_{gold} covering the diverse topics in the meeting based on the transcript C.

Model	Chunk	F1	Gold R1
BERT	512	**39.47**	84.72
	1024	35.41	82.28
Longformer	512	39.24	**85.17**
	1024	38.31	84.35
	2048	37.31	84.16

Table 4: The F1 and gold ROUGE-1 recall scores on highlight sentence extraction task. BERT and Longformer are evaluated on the data with chunk size of 512, 1024 and 2048.

5 Experiments

To jointly accommodate the segmentation-based summarization and overall meeting summarization, we divide the train/dev/test datasets that contain around 80%/10%/10% segmentation summaries as well as overall summaries. There are totally 193/25/21 overall summaries and 1076/135/136 segmentation summaries in train/dev/test sets.

5.1 Highlight Sentence Extraction

Experiment Setup We solve the task as a sequence labeling problem. Due to the lengthy context of meeting transcripts, we try to split the transcript into small chunks. Here we evaluate the BERT (Devlin et al., 2019) and Longformer (Beltagy et al., 2020) models with chunk sizes of 512, 1024 and 2048. We report the F1 score and gold ROUGE-1 recall score as our evaluation metrics, wherein the gold ROUGE recall score is calculated between the generated highlight sentences and the gold overall meeting summary. Find the implementation details in Appendix B.

Results Table 4 illustrates the results of highlight sentence extraction on the test set. We can see that both BERT and Longformer perform better with the shorter input length. However, a huge performance drop is observed when BERT works on the longer inputs while Longformer could perform stably. This finding is consistent with the characteristics of these two models.

5.2 Segmentation-based Multi-granularity Summarization

Experiment Setup This task is comprised of two sub-tasks, i.e., topic segmentation and multi-

Model	Chunk Turns	*Pk*↓	*WinDiff*↓
RANDOM	-	0.421	0.555
EVEN	-	0.441	0.505
BERTSUMEXT	5	0.293	0.310
	10	**0.216**	**0.214**
	15	0.230	0.234
BART$_{enc}$	5	0.334	0.380
	10	0.222	0.216
	15	0.244	0.242

Table 5: The *Pk* and *WinDiff* scores on the topic segmentation task. BERT and BART$_{enc}$ are evaluated on the data with chunk turns of 5, 10 and 15, wherein BART$_{enc}$ stands for using only the encoder part of BART model. ↓ means lower is better.

granularity summarization, wherein the summarization method is applied on the segmented sections.

For the topic segmentation, we formulate the task as a sequence labeling problem. We solve it by BERTSUMEXT model Liu and Lapata (2019) and BART encoder model (Lewis et al., 2020). Specifically, we insert a special token [CLS] at the beginning of each utterance, which would be used to classify whether the utterance is the end of a segment. Then we use an interval segmentation indicator to distinguish different utterances. However, due to the context length exceeding 10K in most samples, we split the meeting transcript into chunks where each chunk contains 5/10/15 turns of utterance. In the stage of inference, we make the predictions on segmented chunks but calculate the scores on the entire meeting transcript. Following previous work (Zhong et al., 2022b), we use the standard metrics *Pk* (Beeferman et al., 1999) and *WinDiff* (Pevzner and Hearst, 2002) to evaluate the segmentation models.

For multi-granularity summarization, we evaluate two settings, i.e., *with predicted segments* and *with gold segments*. For the former setting, the model is trained to first segment the transcript and then generate summaries for the predicted segments. For the latter, we provide the gold segments and evaluate the capacities of generating multi-granularity summaries with correct segments. As the number of predicted segments is uncertain, we calculate the ROUGE-1/2/L scores for each meeting by joining all predicted segmentation summaries together with a special token [Y_SEP]. We employ two widely-used summarization models, i.e., BART and Pegasus (Zhang et al., 2020), as our backbone. We also report the results of RANDOM and ORACLE baselines. RANDOM is to randomly select sentences from the context as the summary, while ORACLE is to select the sentences with the highest ROUGE-1 scores against the ground truth.

In the setting of *with predicted segments*, we adopt the best-performing topic segmentation model, i.e., BERTSUMEXT trained with 10-turn chunks, to segment the transcripts. Then the summaries are generated by BART and Pegasus models on the segmented sections. Furthermore, we also try to initialize the encoder part of the generative BART with the weights from BART$_{enc}$. In this way, we expect the generative model can be aware of the segmentation features. Note that we do not follow the previous work (Liu et al., 2022) which jointly optimizes the segmentation and generation tasks with the pre-trained language models because our transcripts and summaries are much longer than theirs and the model cannot work well with the such lengthy context.

Results Table 5 illustrates the evaluation results of topic segmentation. As we can see, both BERTSUMEXT and BART models are much better than the baselines, i.e., RANDOM and EVEN. RANDOM is the baseline that randomly selects the utterances as the boundary of a segment, while EVEN is to partition the whole transcript evenly. BERTSUMEXT outperforms the BART$_{enc}$ across the board. This is reasonable since BERTSUMEXT captures more inter-segment features by the interval indicator and high-level interact layers. The results of these two models also show the same trend that segmenting the transcript into 10 turns is the best choice. This is consistent with the findings in Sankar et al. (2019) that around 8 turns of dialogue context are enough to capture the contextual features.

Table 6 shows the evaluation results on segmentation-based multi-granularity summarization. When given gold segments, Pegasus ($l = 2048$) achieves the best performance on most metrics. Both BART and Pegasus models outperform the extractive oracle methods across the board, suggesting that the written summaries are abstractive enough. For the summary generation at different granularities, we find that the headline generation is harder than the summary generation owing to its highly condensed information. Comparing the generation of segmentation summary and joint summary, slight improvements are spotted when prompting the segmentation summary generation with the headline.

Model		Headline			Segmentation Summary			Joint Summary		
Segmentor	Generator	R1	R2	RL	R1	R2	RL	R1	R2	RL
With gold segments										
-	Random	21.21	2.53	1.03	43.22	7.63	13.41	43.18	7.66	13.49
-	Oracle	39.90	20.01	34.72	55.10	25.19	32.66	55.21	24.82	31.77
-	BART($l = 1024$)	42.85	25.53	35.96	58.18	23.56	29.65	59.49	24.38	29.80
-	BART($l = 2048$)	41.92	23.73	34.53	59.06	24.35	29.25	59.14	24.35	29.25
-	Pegasus($l = 1024$)	**46.49**	**27.69**	38.92	59.31	25.10	33.19	59.78	25.66	34.11
-	Pegasus($l = 2048$)	45.68	27.22	**39.04**	**59.59**	**25.31**	**33.57**	**59.80**	**26.03**	**34.55**
With predicted segments										
BertSumExt	BART$_{ed}$($l = 1024$)	40.10	21.92	32.43	56.63	22.17	25.69	57.60	22.73	27.16
BertSumExt	Pegasus($l = 1024$)	41.41	22.43	34.93	57.90	23.45	30.98	58.14	23.39	30.80
BART$_{enc}$	Pegasus($l = 1024$)	40.14	20.13	32.76	54.16	21.15	26.89	54.01	20.97	27.51
BART$_{enc}$	BART$_{ed}$	37.44	19.74	30.40	53.30	20.12	23.46	54.20	20.33	23.90

Table 6: Evaluation results of ROUGE F1 on segmentation-based multi-granularity summarization. The scores here is calculated without sentence splitting. l stands for the truncation length. BART$_{ed}$ means the standard encoder-decoder-based BART model.

Method	R1	R2	RL
Vanilla BART($l = 1024$)	37.61	11.38	18.16
Vanilla BART($l = 2048$)	38.83	12.05	18.02
Vanilla Pegasus($l = 1024$)	29.43	17.19	19.42
Vanilla Pegasus($l = 2048$)	27.62	8.56	18.49
Pred. Joint Summary + BART	42.29	13.77	19.17
Pred. Highlights + BART	**45.76**	**16.33**	**22.26**
Gold Joint Summary + BART	55.85	31.33	34.19
Gold Highlights + BART	47.75	18.47	24.14

Table 7: Evaluation results on abstractive meeting summarization. Pred. joint summary/highlights means the pipeline method that first predicts the joint summaries/highlights and then generates the overall summary on the results of first step.

When evaluating with the predicted segments, we find that comparable performance could be reached with a strong topic segmentation model. After conducting more detailed analyses regarding the results of segmentation and generation, we surprisingly find that the most errors of our segmentation model are within three utterances, e.g., the label is the 10^{th} utterance but the prediction is 8^{th}, while the generation model can easily tolerate such deviations. An exception is tying the weights of the segmentation model with the generation model, i.e., BART$_{enc}$ + BART$_{ed}$, which performs much worse than others. We think this is because the generation process is not largely dependent on the segmentation features.

5.3 Abstractive Meeting Summarization

Experiment Setup It is a kind of *long text summarization* task (Liu et al., 2020; Gidiotis and Tsoumakas, 2020), which encourages the model to generate a concise summary (< 512 words) containing the main content of a long text (> 5,000 words). We evaluate the vanilla sequence-to-sequence models and *retrieval-then-generate* methods. Specifically, we fine-tune the BART and Pegasus models with the truncation length of 1024 and 2048 to generate the overall summary. For *retrieval-then-generate* methods, we first *retrieve* the joint summaries or highlight sentences of a meeting transcript, and then produce the overall summaries on the retrieval results. This method hugely decreases the input length to the summarization model. We use the best-performing highlight model and joint summary generation model to finish the retrieval process. We report the ROUGE-1/2/L F1 scores for performance comparison.

Results As shown in Table 7, the vanilla models perform weakly on meeting summarization owing to the lengthy input. As said in Section 3.4, the summary words are evenly distributed in the entire transcript. Therefore, directly truncating the input would cause information loss. The paradigm of *retrieval-then-generate* is a good solution to this problem since it retrieves the key information first and then generates the summary on the retrieval results. When the summarization is based on gold content, the joint summary is a better ground than the highlight information. This is reasonable since the joint summary contains more concentrated information with fewer words. However, when the summarization is based on the retrieval results, predicted highlights even achieve comparable summarization performance to the gold highlights. This is attributed to the effectiveness of our highlight ex-

Meeting ID: 71962448; Segment ID: 2	
Pegasus	中国当代艺术蓬勃发展的原因:中国有庞大的艺术教育的底盘，学校的空间很大，可以让新的年轻艺术家有很好的活动的条件，中国的材料很丰富，加工便宜，人力资源很便宜，在市场上还有快速发展的过程，至少有潜在的购买力吸引这个艺术家，在世界议坛上很流行的中国政治口号，在文化信息的交往中，中国艺术家比西方艺术家有更宽拓的空间。
Gold	艺术教育问题的分析:从艺术教育的角度来看，中国当代艺术蓬勃发展有六大原因，一是中国有庞大的艺术教育地盘，二是学校空间大，可以让新的年轻艺术家有很好的活动条件，三是中国的材料丰富，并且人力资源便宜，四是市场上有快速发展的过程，存在潜在的购买力，五是中国的元素，文化信息交流中，中国艺术家更具宽拓空间，六是中国正处于巨大变革中，批判性的建构是当代艺术的重要动力。
Meeting ID: 634267454	
Pred. Sum. + BART	区块链是一个学科化的研究而不仅仅是热点，它是一种从特殊走向另一种特殊的时代，它一直尝试对社会进行范式的迁移，从历史交汇到文化主导，在未来艺术形式的一个很关键的线索，可以发现不断不断地将权力分散到更多人的手中，能够参与到社会的运动中，这也是未来的艺术选择艺术家的标准。艺术作品的表达方式可以用画面表达和用音像的方式表达，或者是用任何一种行为方式，关键是能否用擅长的工具、或者愿意拓展你的想象力去表达你对加密世界的思考。
Gold	加密的爆发搅活了整个艺术品市场，带动了艺术领域改革开放，打破了原有的概念，带来了新的创作媒介，是对艺术行业的一个挑战，是一个转向，让创作者与世界更好的融合，但是艺术家还是要有标准的，要有人品、有创新的能力、要让作品更够更清晰的表达，加密是从底层起来的，有着蓬勃的生命力，每一个参与社会推进的人就已经是在参与艺术的创作了。

Table 8: Case studies on the joint summary (the first part) of the segment and the overall summary (the second part). Find the translations in Appendix A.2.

Method	Headline	Seg. Sum	Joint Sum
BART($l = 2048$)	41.92	59.06	59.14
Pegasus($l = 2048$)	**45.68**	59.59	59.80
CONVLM($l = 2048$)	45.25	**59.98**	**60.70**

Table 9: ROUGE-1 F1 scores of different models on multi-granularity summarization with gold segments.

traction model and the redundancy information of highlight sentences. The worse performance from the predicted joint summary is because the generation of joint summaries is essentially a challenging task, which might cause more error propagation.

6 Further Discussions

6.1 Conversational Solutions

DIALOGUELM (Zhong et al., 2022b) is a strong baseline for meeting summarization in English. Based on the fact that meeting is a kind of long conversation, DIALOGUELM adopts sparse architecture and dialogue-specific pre-training objectives, such as speaker masking, turn splitting, turn merging and turn permutation, to capture the conversational features. It finally demonstrates that the pre-trained dialogue model is also a good solution to the meeting summarization task. Inspired by this observation, we make a preliminary attempt at the conversational solution to VCSUM. Specifically, we pre-train an encoder-decoder-based dialogue model, dubbed CONVLM, using our in-house Chinese dialogue data. We provide details of CONVLM in Appendix C. We train the model with the objectives of speaker identification and response generation. After the pre-training, we fine-tune the model on VCSUM. To simplify the comparison, we evaluate on the task of multi-granularity summarization with gold segments.

As shown in Table 9, CONVLM achieves comparable or better performance against the summarization models. Especially for the generation of segmentation and joint summary, CONVLM consistently outperforms the baseline models, suggesting that modeling conversational features could benefit the summarization task on VCSUM. This finding demonstrates the conversational characteristics of our dataset and sheds light on using dialogue-specific pre-trained language models to solve the tasks of VCSUM.

6.2 Error Analysis

To study the difficulties of VCSUM, we take a detailed analysis of the error cases. We find that around 50% of errors is information missing which is essentially caused by lengthy input truncation. For example, in the first row of Table 8, the segmentation summary is much semantically close to the ground truth with only one key point (the red words) missing. This phenomenon is more severe when generating the overall summaries owing to the long and informative context. There are also some errors from irrelevant information or redundant information, accounting for around 20%. These kinds of errors are mostly found in the retrieval-then-generate methods that retrieve some insignificant content in the first stage, thus finally misleading the generation process, like the blue words in Table 8. The remaining 30% errors include factual errors and syntactic errors.

7 Conclusion and Future Work

In this work, we collect a large and high-quality Chinese meeting summarization dataset from real-life videos, namely VCSUM. The dataset is versatile to support the tasks of highlight sentence extraction, segmentation-based summarization, multi-granularity summarization and meeting summarization. Depth analyses demonstrate the superiority of our dataset. We then provide a strong benchmark for different downstream tasks on VCSUM. For future work, we believe the development of an end-to-end framework that could jointly solve all tasks of VCSUM is a promising direction. Furthermore, we will also consider to release the video and audio files of the annotated meetings to facilitate the research of multi-modality meeting summarization.

Acknowledgement

We sincerely appreciate the valuable and constructive comments from the reviewers. This work was supported in part by the Hong Kong ITF grant PRP/079/22FX, the Technological Breakthrough Project of Science, Technology and Innovation Commission of Shenzhen Municipality under Grants JSGG20201102162000001, InnoHK initiative, the Government of the HKSAR, Laboratory for AI-Powered Financial Technologies.

Limitations

A potential limitation of this work is that we just try some straightforward methods on the summarization tasks, such as vanilla generative pre-trained language models or pipeline retrieval-then-generate methods. We do not try the end-to-end two-stage approaches in this work since we are more focused on the contributions of the dataset construction and building the benchmark. We leave the development of advanced models as future work.

Ethics Statement

The construction of dataset. All videos in our newly-introduced dataset are available on the Chinese video sharing websites and are public to the download. To avoid the potential ethical issues, we carefully checked all videos in multiple aspects, as said in Section 3.1. We try to guarantee that all videos do not involve any offensive, gender-biased, political content and any other ethical issues. During the annotation, we instruct annotators to anonymize or remove the sensitive or private information.

We recruit eight annotators that passed our examinations from the crowdsourcing platform, and two quality inspectors from our research team. To fairly paid the annotations, we first take an in-house annotation to evaluate the speed and difficulty of the large-scale annotations. Finally, we pay each annotator \$25-\$30 per hour. Typically, it would take around 2 hours to annotate a one-hour meeting. So, the workers are compensated \$50-\$60 per sample.

Applications. We will release the code and dataset along with friendly instructions to support its correct use. However, we still need to emphasize that abstractive summarization is a kind of generation task, which is not as controllable as we think. It still would generate some novel or unexpected words occasionally. Therefore, further research on the summarization faithfulness is warmly needed.

References

Doug Beeferman, Adam Berger, and John Lafferty. 1999. Statistical models for text segmentation. *Machine learning*, 34(1):177–210.

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

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118.

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

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.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring dialogpt for dialogue summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1479–1491.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*, volume 1, pages I–I. IEEE.

Chris Kedzie, Kathleen Mckeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828.

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

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.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

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

Xiaojun Liu, Chuang Zhang, Xiaojun Chen, Yanan Cao, and Jinpeng Li. 2020. Clts: A new chinese long text summarization dataset. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I*, pages 531–542.

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.

Yang Liu, Chenguang Zhu, and Michael Zeng. 2022. End-to-end segmentation-based news summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 544–554.

Ziming Mao, Chen Henry Wu, Ansong Ni, Yusen Zhang, Rui Zhang, Tao Yu, Budhaditya Deb, Chenguang Zhu, Ahmed Awadallah, and Dragomir Radev. 2022. Dyle: Dynamic latent extraction for abstractive long-input summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1687–1698.

I McCowan, J Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al. 2005. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research.*, pages 88–100.

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

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Lev Pevzner and Marti A Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37.

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

Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. 2021. Emailsum: Abstractive email thread summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6895–6909.

Ming Zhong, Yang Liu, Suyu Ge, Yuning Mao, Yizhu Jiao, Xingxing Zhang, Yichong Xu, Chenguang Zhu, Michael Zeng, and Jiawei Han. 2022a. Unsupervised summarization with customized granularities. *arXiv preprint arXiv:2201.12502*.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022b. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

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.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A hierarchical network for abstractive meeting summarization with cross-domain pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203.

A Translations

A.1 Translated Examples.

See Table 10.

A.2 Translated Case Study

See Table 11.

B Implementation

We provide the details of implementations here. For all experiments, we take five runs with different random seeds and report the average score.

Highlight Sentence Extraction We solve the task at sentence-level. Specifically, we split the utterances into small sentences by *comma*. Then, we insert a special token [CLS] at the beginning of each sentence to represent the sentence. The binary classification is conducted on the special tokens to decide whether a sentence should be highlighted or not. Owing to the unbalanced ratio between positive and negative samples, we only active 40% negative samples in the training stage. The batch size is set to 16 or 32, depending on the chunk size. The models are optimized by Adam (Kingma and Ba, 2015) optimizer with an initial learning rate of 5e-5. The pre-trained BERT (`hfl/chinese-roberta-wwm-ext-large`) and Longformer (`IDEA-CCNL/Erlangshen-Longformer-330M`) models are loaded from Huggingface[2]. We train the models for 10 epochs and select the best model on the validation set to evaluate on the test set. All experiments are conducted on 8 NVIDIA A100 GPUs.

Meeting Transcript about Fundamental Education
Speaker1: I would like to ask about the basic education. Now only Harbin Institute of Technology has undergraduates in Shenzhen. So Harbin Institute of Technology needs to educate the fresh undergraduates who just finish the high school courses. Do you have any sharing regarding this aspect. *Speaker2*: Education and medical care are two shortcomings of Shenzen. Especially in terms of basic education, although the development of high-end education has been greatly improved, the basic education ... *Speaker3*: On the other hand, it is also related to the thoughts of parents. Because ... `[EOS]`
Headline: Shortcomings of Fundamental Education **Segmentation summary**: At present, basic education is weak, and it is still in the stage of unresolved food and clothing problems, and schools do not pay enough attention to it. The thought of parents also has an influence. They have been following the old path of exam-oriented education, and the ultimate goal is to get high scores in the college entrance examination.....
Speaker1: It is a very comprehensive sharing, covering the all aspects of our work. Next, we invite ... *Speaker4*: OK, I would like to share my view of talent training ... `[EOS]`
Headline: The methods of talent training **Segmentation summary**: It is necessary to provide a platform to provide opportunities for the young people, and at the same time to cultivate their strengths. It is necessary to discover the strengths of children from an early age. ... At the same time, new technologies and new ideas must be spread to small cities through big cities.
...
Overall summary: Basic education should provide more platforms. While paying attention to basic education, we should also pay more attention to children's interests. Schools should provide a level of selection criteria, not too limited in grades, but also differentiated ...

Table 10: An example from our dataset in English. The green texts are the highlighted sentences. The token `[EOS]` is used to distinguish different segmentations. We provide multi-granularity summaries to a meeting transcript, including headline, segmentation summary and overall summary.

Segmentation-based Multi-granularity Summarization For topic segmentation, the BERT-SUMEXT is based on the large BERT model while

[2] `https://huggingface.co/`

BART$_{enc}$ is based on large BART. We truncate each utterance with maximum utterance length of 128 words. The batch size is set to 16 or 32, depending on the chunk turn size. The other settings are mostly same to highlight sentence extraction.

For multi-granularity summarization, we employ the BART (fnlp/bart-large-chinese) and Pegasus (IDEA-CCNL/Randeng-Pegasus-523M-Summary-Chinese) models from Huggingface as our backbones. The input sequence is truncated to the maximum length of 1024 or 2048. To incorporate speaker information, we add a speaker indicator at the beginning of each utterance, e.g., *[Speaker_1: utterance_1; ...;Speaker_n: utterance_n;]*. The batch size is set to 64. All models are optimized by Adam, and the learning rate is initialized to 5e-5 and linearly updated. During the training process, the best model is selected on the validation loss. In the stage of inference, we generate the summary using beam search with beam size of 5 and length penalty of 1.0.

Abstractive Meeting Summarization For the vanilla generative models, we truncate the input sequence to the maximum length of 1024 or 2048. The batch size is set to 64 or 32, depending on the truncation length. For the retrieval-then-generate methods, all predictions are obtained from the corresponding best-performing models, i.e., BERTSUMEXT+Pegasus($l = 1024$) and BERT trained with chunk size of 512. During the decoding, we generate the summary using beam search with beam size of 5 and length penalty of 1.2.

C Details of CONVLM

CONVLM employs the architecture of large BART model. The training data is collected from Chinese social media, including Zhihu[3], Weibo[4] and Douban[5]. We totally crawl 10 billion pieces of dialogue data, and select 10 million pieces of high-quality data to train CONVLM. The pre-training objectives are speaker identification and response generation, wherein speaker identification is to predict the masked speaker indicators and response generation is to generate the response based on the context. We truncate the input sequence into 512. The batch size is set to 128. The learning rate is set to 5e-5 and linear updated along with the training process. We train the CONVLM for 300K steps, which takes around 50 hours on 16 NVIDIA A100 GPUs. We validate the model each 1K steps. We save the model with the best validation loss and the last 10 checkpoints.

D Screenshot of Annotation Platform

See Figure 3.

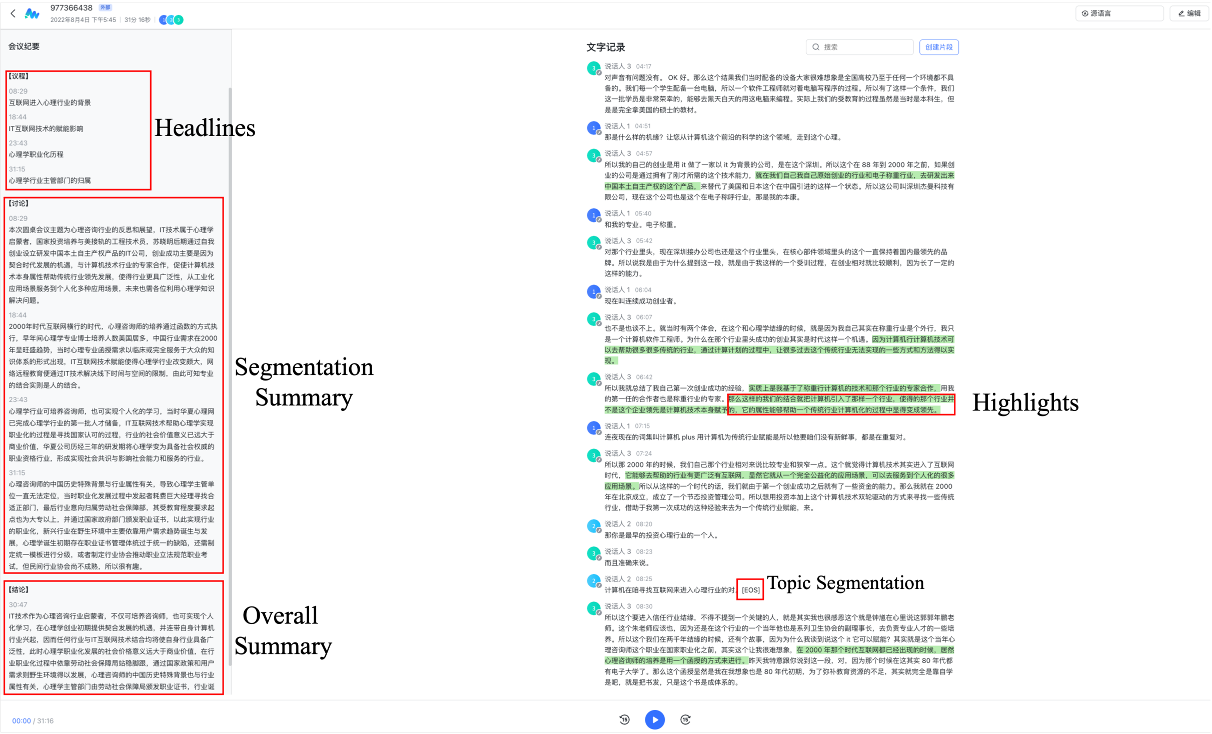
[3] https://www.zhihu.com/

[4] https://www.weibo.com/

[5] https://www.douban.com/

	Meeting ID: 71962448; Segment ID: 2
Pegasus	Reasons for the vigorous development of Chinese contemporary art: China has a huge infrastructure for art education, and the school has a lot of space, which allows new young artists to have good conditions for activities. China's materials are abundant, processing is cheap, human resources are cheap, and there is still rapid development in the market. At least there is the potential purchasing power to attract the artist. Chinese political slogans are very popular in the world forum. In the exchange of cultural information, Chinese artists have a wider space than Western artists.
Gold	Analysis of Art Education Problems:From the perspective of art education, there are six reasons for the vigorous development of Chinese contemporary art. One is that China has a huge territory for art education. Second, the school has a large space, which allows new young artists to have good conditions for activities. Third, China is rich in materials and cheap in human resources. Fourth, there is a process of rapid development in the market and potential purchasing power. The fifth is Chinese elements. In the exchange of cultural information, Chinese artists have more room to expand. Sixth, China is undergoing tremendous changes, and critical construction is an important driving force for contemporary art.
	Meeting ID: 634267454
Pred. Sum. + BART	Blockchain is a disciplined research rather than just a hotspot. It is an era from a special to another special, and it has been trying to transfer the paradigm of society, from the intersection of history to the dominance of culture. A key clue in the future of art forms can be found in the continued decentralization of power into more hands. Being able to participate in social movements is also the standard for future artists to choose artists.The expression of works of art can be expressed in pictures, audio and video, or in any kind of behavior. The key is whether you can use the tools you are good at, or be willing to expand your imagination to express your thinking about the encrypted world.
Gold	The outbreak of encryption has stirred up the entire art market, driven the reform and opening up of the art field, broken the original concept, and brought new creative media. This is a challenge to the art industry and a turning point, allowing creators to better integrate with the world. But artists still need standards, they must have character, they must have the ability to innovate, and they must make their works clearer and more expressive. Encryption starts from the bottom and has vigorous vitality. Everyone who participates in social advancement is already participating in the creation of art.

Table 11: Case studies on the joint summary (the first part) of the segment and the overall summary (the second part).



977366438

2022年8月4日 下午5:45 | 31分 16秒

会议纪要

【议程】

08:29
互联网进入心理行业的背景

18:44
IT互联网技术的赋能影响

23:43
心理学职业化历程

31:15
心理学行业主管部门的归属

【讨论】

08:29
本次圆桌会议主题为心理咨询行业的反思和展望，IT技术属于心理学启蒙者，国家投资培养与美接轨的工程技术员，苏晓明后期通过自我创业设立研发中国本土自主产权产品的IT公司，创业成功主要是因为契合时代发展的机遇，与计算机技术行业的专家合作，促使计算机技术本身属性帮助传统行业领先发展，使得行业更具广泛性，从工业化应用场景服务到个人化多种应用场景，未来也需各位利用心理学知识解决问题。

18:44
2000年时代互联网横行的时代，心理咨询师的培养通过函数的方式执行，早年间心理学专业博士培养人数美国居多，中国行业需求在2000年呈旺盛趋势，当时心理专业函授需求以临床或完全服务于大众的知识体系的形式出现，IT互联网技术赋能使得心理学行业改变颇大，网络远程教育便通过IT技术解决线下时间与空间的限制，由此可知专业的结合实则是人的结合。

23:43
心理学行业可培养咨询师，也可实现个人化的学习，当时华夏心理网已完成心理学行业的第一批人才储备，IT互联网技术帮助心理学实现职业化的过程是寻找国家认可的过程，行业的社会价值意义已远大于商业价值，华夏公司历经三年的研发期将心理学变为具备社会权威的职业资格行业，形成实现社会共识与影响社会能力和服务的行业。

31:15
心理咨询师的中国历史特殊背景与行业属性有关，导致心理学主管单位一直无法定位，当时职业化发展过程中发起者耗费巨大经理寻找合适正部门，最后行业意向归属劳动社会保障部，其受教育程度要求起点也为大专以上，并通过国家政府部门颁发职业证书，以此实现行业的职业化，新兴行业在野生环境中主要依靠用户需求趋势诞生与发展，心理学诞生初期存在职业证书管理体统过于统一的缺陷，还需制定统一模板进行分级，或者制定行业协会推动职业立法规范职业考试，但民间行业协会尚不成熟，所以很有趣。

【结论】

30:47
IT技术作为心理咨询行业启蒙者，不仅可培养咨询师，也可实现个人化学习，在心理学创业初期提供契合发展的机遇，并连带自身计算机行业兴起，因而任何行业与IT互联网技术结合均将使自身行业具备广泛性，此时心理学职业化发展的社会价格意义远大于商业价值，在行业职业化过程中依靠劳动社会保障局站稳脚跟，通过国家政策和用户需求则野生环境得以发展，心理咨询师的中国历史特殊背景也与行业属性有关，心理学主管部门由劳动社会保障局颁发职业证书，行业诞

00:00 / 31:16

文字记录

说话人 3 04:17
对声音有问题没有。 OK 好。那么这个结果我们当时配备的设备大家很难想象是全国高校乃至于任何一个环境都不具备的。我们每一个学生配备一台电脑，所以一个软件工程师就对着电脑写程序的过程。所以有了这样一个条件，我们这一批学员是非常荣幸的，能够去黑天白天的用这电脑来编程。实际上我们的受教育的过程虽然是当时是本科生，但是是完全拿美国的硕士的教材。

说话人 1 04:51
那是什么样的机缘？让您从计算机这个前沿的科学的这个领域，走到这个心理。

说话人 3 04:57
所以我的自己的创业是用 it 做了一家以 it 为背景的公司，是在这个深圳。所以这个在 88 年到 2000 年之前，如果创业的公司是通过拥有了刚才所需的这个技术能力，就在我们自己我自己原始创业的行业和电子称重行业，去研发出来中国本土自主产权的这个产品，来替代了美国和日本这个在中国引进的这样一个状态。所以这公司叫深圳杰曼科技有限公司，现在这个公司也是这个在电子称呼行业，那是我的本康。

说话人 1 05:40
和我的专业。电子称重。

说话人 3 05:42
对那个行业里头，现在深圳接办公司也还是这个行业里头，在核心部件领域里头的这个一直保持着国内最领先的品牌。所以说我是由于为什么提到这一段，就是由于我这样的一个受训过程，在创业相对就比较顺利，因为长了一定的这样的能力。

说话人 1 06:04
现在叫连续成功创业者。

说话人 3 06:07
也不是也谈不上。就当时有两个体会，在这个和心理学结缘的时候，就是因为我自己其实在称重行业是个外行，我只是一个计算机软件工程师。为什么在那个行业里头成功的创业其实是时代这样一个机遇。因为计算机行计算机技术可以去帮助很多很多传统的行业，通过计算计划的过程中，让很多过去这个传统行业无法实现的一些方式和方法得以实现。

说话人 3 06:42
所以我就总结了我自己第一次创业成功的经验，实质上是我基于了称重行计算机的技术和那个行业的专家合作，用我的第一任的合作者也是称重行业的专家。那么这样的我们的结合就把计算机引入了那样一个行业，使得的那个行业并不是这个企业领先是计算机技术本身赋予的，它的属性能够帮助一个传统行业计算机化的过程中显得变成领先。

说话人 1 07:15
连夜现在的词集叫计算机 plus 用计算机为传统行业赋能是所以他要咱们没有新鲜事，都是在重复对。

说话人 3 07:24
所以那 2000 年的时候，我们自己那个行业相对来说比较专业和狭窄一点。这个就觉得计算机技术其实进入了互联网时代，它能够去帮助的行业有更广泛有互联网，显然它就从一个完全公益化的应用场景，可以去服务到个人化的很多应用场景。所以从这样的一个时代的话，我们就由于第一个创业成功之后就有了一些资金的能力。那么我就在 2000 年在北京成立，成立了一个节态投资管理公司。所以想用投资本加上这个计算机技术双轮驱动的方式来寻找一些传统行业，借助于我第一次成功的这种经验来去为一个传统行业赋能，来。

说话人 2 08:20
那你是最早的投资心理行业的一个人。

说话人 3 08:23
而且准确来说。

说话人 2 08:25
计算机在咱寻找互联网来进入心理行业的对。[EOS]

说话人 3 08:30
所以这个要进入信任行业结缘，不得不提到一个关键的人，就是其实我也很感恩这个就是钟馗在心里说这郭郭年鹏老师。这个朱老师应该也，因为还是在这个行业的一个当年他也是系列卫生协会的副理事长，去负责专业人才的一些培养。所以这个我们在两千年结缘的时候，还有个故事，因为为什么我谈到说这个 it 它可以赋能？其实就是这个当年心理咨询师这个职业在国家职业化之前，其实这个让我很难想象，在 2000 年那个时代互联网都已经出现的时候，居然心理咨询师的培养是用一个函授的方式来进行。昨天我特意跟你说到这一段，对，因为那个时候在这其实 80 年代都有电子大学了。那么这个函授显然是我在我想象也是 80 年代初期，为了弥补教育资源的不足，其实就完全是靠自学是吧，就是把书发，只是这个书是成体系的。

Figure 3: The screenshot of the annotation platform.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
See the Limitation part.

☑ A2. Did you discuss any potential risks of your work?
See the Ethics Statement part.

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

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

See Section 3.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Upon acceptance, we will release our dataset and provide the guidelines for its use and distribution.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Upon acceptance, we will release our dataset and provide the guidelines for its use and distribution.

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

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

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

C ☑ Did you run computational experiments?

See Section 5 and Appendix B.

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

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

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

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

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
See Section 5 and Appendix B.

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

See Section 3 and Ethics Statement.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
See Section 3 and Appendix D.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
See Section 3 and Ethics Statement.

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

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

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
See Section 3 and Ethics Statement.