MUG: A GENERAL MEETING UNDERSTANDING AND GENERATION BENCHMARK

Qinglin Zhang[1], *Chong Deng*[1], *Jiaqing Liu*[1], *Hai Yu*[1], *Qian Chen*[1]

Wen Wang[1], Zhijie Yan[1], Jinglin Liu[2], Yi Ren[2], Zhou Zhao[2]

[1] Speech Lab of DAMO Academy, Alibaba Group [2] Zhejiang University

[1]{qinglin.zql, dengchong.d, mingzhai.ljq, yuhai.yu}@alibaba-inc.com
[1]{tanqing.cq, w.wang, zhijie.yzj}@alibaba-inc.com
[2]{jinglinliu, rayeren, zhaozhou}@zju.edu.cn

ABSTRACT

Listening to long video/audio recordings from video conferencing and online courses for acquiring information is extremely inefficient. Even after ASR systems transcribe recordings into long-form spoken language documents, reading ASR transcripts only partly speeds up seeking information. It has been observed that a range of NLP applications, such as keyphrase extraction, topic segmentation, and summarization, significantly improve users' efficiency in grasping important information. The meeting scenario is among the most valuable scenarios for deploying these *spoken language processing* (SLP) capabilities. However, the lack of large-scale public meeting datasets annotated for these SLP tasks severely hinders their advancement. To prompt SLP advancement, we establish a large-scale **general Meeting Understanding and Generation Benchmark (MUG)** to benchmark the performance of a wide range of SLP tasks, including topic segmentation, topic-level and session-level extractive summarization and topic title generation, keyphrase extraction, and action item detection. To facilitate the MUG benchmark, we construct and release a large-scale meeting dataset for comprehensive long-form SLP development, *the AliMeeting4MUG Corpus*, which consists of 654 recorded Mandarin meeting sessions with diverse topic coverage, with manual annotations for SLP tasks on manual transcripts of meeting recordings. To the best of our knowledge, the AliMeeting4MUG Corpus is so far the largest meeting corpus in scale and facilitates most SLP tasks. In this paper, we provide a detailed introduction of this corpus, SLP tasks and evaluation methods, baseline systems and their performance[1].

Index Terms— Long-form Spoken Language Processing, Keyphrase Extraction, Topic Segmentation, Title Generation, Summarization, Action Item Detection

1. INTRODUCTION

Millions of meetings occur daily around the world, vital for seeking information, sharing knowledge, and improving productivity. They produce an enormous number of video/audio recordings. Listening to the long recordings for acquiring information is extremely inefficient. Even after ASR systems convert recordings into spoken language documents, reading these documents only partly ameliorates the efficiency issue. It has been observed that NLP technologies on spoken language documents, such as keyphrase extraction and

---

[1]The Alimeeting4MUG corpus is released at `https://modelscope.cn/datasets/modelscope/Alimeeting4MUG/summary` and the baseline system is released at `https://github.com/alibaba-damo-academy/SpokenNLP`.

summarization, are crucial for distilling, organizing, and prioritizing information and significantly improves users' efficiency in grasping important information, as shown in prior studies on topic segmentation [1], keyphrase extraction [2] and summarization [3]. Topic segmentation and keyphrase extraction are also useful for downstream information retrieval [4, 5, 6] and summarization [7].

Compared to written&formal text, meeting transcripts pose great challenges to spoken language processing (SLP) tasks. First, meeting transcripts exhibit *a wide variety of spoken language phenomena*, such as disfluencies and redundancy, grammar errors, coreference and information drop, colloquial expressions, and incomplete/fragmented sentences due to speaker interactions. These spoken language phenomena cause a significant discrepancy to the majority of training data of NLP models, i.e., written text, hence leading to dramatic performance degradation. Second, meeting transcripts are usually *long-form documents* with several thousand words or more, challenging to mainstay Transformer-based models with quadratic complexity to the input sequence length. Although ASR errors also pose a great challenge to SLP on meetings [8], we focus on the two key challenges above in this paper since many prior works demonstrate that even on manual transcripts of meetings, i.e., transcripts **without ASR errors**, performances of strong NLP models degrade dramatically compared to their performances on written text, due to these two key challenges. For example, abstractive summarization performance on Chinese written text dataset CLES [9] is about 20 absolute higher than that on our meeting corpus.

However, publicly available meeting corpora supporting SLP tasks are limited and on small scale, severely hindering the advancement of SLP in meetings. The ICSI meeting corpus [10] comprises 75 recorded academic meetings of research discussions (72 hours in total) at ICSI in Berkeley. The AMI meeting corpus [11] consists of 137 meetings (100 hours in total) about industrial product design. The ELITR Minuting Corpus [12] consists of 179 project meetings, with 120 meetings in English and 59 in Czech. There are also publicly available committee meetings of formal discussions, which are quite different in domains and style from academic or project meeting corpora. The QMSum dataset [13] includes AMI and ICSI meetings as well as 25 committee meetings of the Welsh Parliament and 11 from the Parliament of Canada. Most other public meeting corpora are quite small and have no SLP annotations, such as the ISL meeting corpus [14] and the NIST Meeting Room Corpus [15].

To address this issue, we establish a **general Meeting Understanding and Generation benchmark (MUG)** for the community to benchmark the performance of a wide range of SLP tasks, including topic segmentation, topic-level and session-level extractive summarization, topic title generation, keyphrase extraction, and ac-

Datasets	#Sessions	#Avg. Turns	#Avg. Speakers	Avg. Session Len.	Supported Tasks	Language
AMI	137	535.6	4.0	5,570.4	Action, SUM, TS	English
ICSI	59	819.0	6.3	8,567.7	Action, SUM, TS	English
ELITR Minuting Corpus (English)	120	727	5.9	7,066	SUM	English
ELITR Minuting Corpus (Czech)	59	1,205	7.6	8,534	SUM	Czech
QMSum	232	556.8	9.2	12,026.3	QA, SUM, TS	English
AMC (ours)	**654**	376.3	2.5	10,772.5	**Action, KPE, SUM, Title, TS**	Mandarin

Table 1. Statistics of existing meeting corpora and our **AliMeeting4MUG Corpus (AMC)**. For tasks supported by meeting corpora, **TS**:Topic Segmentation, **SUM**:Summarization, **QA**:Question Answering, **KPE**:Keyphrase Extraction, **Title**:Topic Title Generation, **Action**:Action Item Detection. Avg. Turns and Avg. Speakers are per session. Avg. Session length is the number of tokens per session after tokenization.

tion item detection. To facilitate the MUG benchmark, we construct and release a large-scale meeting dataset for long-form SLP development, **the AliMeeting4MUG Corpus (AMC)**, which consists of 654 recorded Mandarin meeting sessions with diverse topic coverage, with manual annotations for SLP tasks on manual transcripts of meeting recordings. As compared to the existing meeting corpora in Table 1, to the best of our knowledge, AMC is so far the **largest meeting corpus in scale and facilitates the most SLP tasks**. We build baseline systems and report evaluation results on the MUG tasks. Next, we provide a detailed introduction to AMC, SLP tasks and evaluation methods, baseline systems and performance.

2. THE ALIMEETING4MUG CORPUS (AMC)

2.1. Data Collection

The AliMeeting4MUG Corpus extends the **224** meetings from our previously released *Alimeeting* corpus for the ICASSP2022 M2MeT (multi-channel multi-party meeting transcription) challenge [16] with **430** additionally collected meetings. Each meeting session consists of a 15-minute to 30-minute discussion by 2-4 participants covering certain topics. *We make sure all 654 meetings have no personally identifiable information, nor sensitive information about projects or organizations.* Topics span medical treatment, education, business, organization management, industrial production, and other daily routines. The extended meeting sessions generally adopt the same data collection procedure as the Alimeeting corpus [16], except that the extended meeting sessions have a smaller number of participants (from avg. 3.2 speakers per session in Alimeeting to 2.1 in the extended sessions) and a much lower speech overlap ratio (from 42.27% and 34.76% on Alimeeting Train and Eval sets to around 10% in extended sessions). For creating AMC, we adopt the same manual transcripts with manually inserted punctuation and manual speaker labels for the Alimeeting corpus, and follow the same procedure for manual transcription, manual punctuation insertion and speaker labeling as conducted for the Alimeeting corpus [16] on the extended sessions. We create manual annotations for all 5 SLP tasks (Table 1) for 524 meetings and manually annotate the rest 130 meetings with only topic segmentation. For Track2-5 described in Section 3, we partition the 524 meetings with all 5 SLP annotations into 295 sessions for training (**Train**), 65 sessions for system development (**Dev**), 82 sessions for Stage1 test set (**exceptTS-Test1**) and 82 sessions for Stage2(Final) test set(**exceptTS-Test2**). For Track1, the same Train and Dev sets are used and we partition the 130 meetings with only TS labels into 65 sessions for Stage1 test set(**TSonly-Test1**) and 65 sessions for Stage2(Final) test set(**TSonly-Test2**). Statistics of AMC Train and Dev sets are summarized in Table 2.

	Train	Dev
#Sessions	295	65
Avg. #Speakers	2.7	2.6
Avg. #Turns	440.1	439.4
Avg. #Paragraphs	469.7	463.4
Avg. #Sentences	722.8	705.7
Avg. Session Length	11253.5	11516.1
Avg. Turn Length	25.6	26.2
Avg. Paragraph Length	24.0	24.8
Avg. Sentence Length	15.6	16.3

Table 2. Statistics of Train and Dev sets of the AliMeeting4MUG Corpus. All average numbers are per-session average. All lengths are measured by the number of tokens after tokenization.

2.2. Spoken Language Processing Annotations

All sessions in AMC are recorded with both far-field and near-field recordings. Before manual annotations, we align the signal from the near-field recordings with far-field recordings and select the signal with higher quality for manual transcription. We create manual transcripts for audio together with manually inserted punctuation and manual speaker labels, with careful quality control [16]. Semantic units ended with a manually labeled period, question mark, and exclamation are treated as **sentences** for the following SLP annotations and baseline systems (Section 4). We use character tokenization for Chinese and whitespace tokenization for other languages. Next, we segment the resulting spoken documents into paragraphs using our sequence model that sets new SOTA for paragraph segmentation on both written text benchmarks and spoken documents [17]. The model improves performance on spoken document segmentation by exploiting phonetic information and improves efficiency by processing multiple sentences simultaneously and employing a self-adaptive sliding window. To simplify display, we break paragraphs on speaker changes based on the manual speaker labels. Then, we apply a series of manual annotations on the paragraph-segmented spoken documents, including Topic Segmentation, Topic-level and Session-level Extractive Summarization, Keyphrase Extraction, Topic Title Generation, and Action Item Detection. Our annotation guidelines fully consider synergy between tasks with details described as follows.

Topic Segmentation (TS) Following TS annotations on ICSI Meeting Corpus [18], we consider purposes of TS as high-level understanding and presentation of discourse and enabling users to browse the content efficiently. Our topic segments denote a fine-grained, focused discussion of a subject matter. The annotators first go through the entire document, acquiring a basic understanding of its content and discourse, then label key sentence candidates in the document for Topic-level Extractive Summarization. A valid topic segment is

required to contain 1-3 key sentences conveying the central ideas of the topic. Topic segments are non-overlapping, both semantically self-contained and discriminative from other topics. When some key sentence candidates among neighboring topics are semantically similar, these topics are scrutinized to determine whether they should be merged. Similar to [18], we allow topics to cross speaker changes; whereas, in order to maintain the paragraph segmentation structures, a topic boundary is annotated only at the last sentence of paragraphs. TS-annotated documents are then annotated for Topic-level Extractive Summarization and Topic Title Generation **simultaneously**.

Topic-level and Session-level Extractive Summarization (ES) The core annotation schema for *Topic-level Extractive Summarization* is described above. We restrict reference key sentences to sentence types other than interrogative sentences, and also conduct deduplication. To address incomplete meanings of key sentence candidates, their context sentences may be concatenated to create semantically complete key sentences. For *Session-level Extractive Summarization*, the reference key sentences are based on deduplicating and filtering reference key sentences for Topics. Key sentence candidates that are not semantically complete are filtered out. We limit the number of session-level reference key sentences to be about *twice* the number of annotated topics.

Topic Title Generation (TTG) Based on the annotations of TS and ES, annotators create a title for each topic by summarizing its central idea concisely with a phrase or short sentence, in formal language with an **(optional subject)-predicate-object** expression focusing on entities and events. TTG can be considered as *Topic-level Abstractive Summarization into a very short summary*. TTG also serves quality assessment for ES: annotation errors on ES observed in this stage are fixed before annotating topic titles. A title may use phrases from annotated topic-level key sentences as well as contain information not covered by the key sentences, but may not create new information not conveyed in the topic. Also, titles for different topics should be significantly discriminative.

Keyphrase Extraction (KPE) Exploiting the annotations of TS, ES, and TTG, annotators extract top-K ($K <= 20$) keyphrases (KPs) for a session. KPs need to be in formal language and semantically complete. The majority of KPs are entities and events.

Action Item Detection (AID) Following [18], action items refer to a task discussed in the meeting and assigned to the participant(s) and expected to complete *within a short time window* after the meeting. The input to AID annotations is manual transcripts with manually inserted punctuation. We label sentences containing information about actionable tasks (task description, time frame, owner) as positive samples and otherwise negative samples.

Exploring Multi-Annotator Annotations For TS annotation, one annotator conducts initial annotations and an expert reviews and fixes the annotations to create the final manual topic labels. For each of ES, TTG, KPE, and AID annotations, each document is annotated by **three annotators**. We compute the standard ROUGE-1,2,L F-score [19](R-1,R-2,R-L)[2] to evaluate the Inter-Annotator Agreement (**IAA**) for ES and TTG between each pair of annotations. We compute exact F_1 (based on exact match) to evaluate IAA for KPE and compute the Kappa coefficient [20] to evaluate IAA for AID. For KPE, the union of the labels from the three annotators is used as the final manual labels for training and evaluation. For topic- and session-level ES, the union of labeled key sentences from three annotators for each topic or session is used as a reference for training. For TTG, we create three copies of training data with reference from

---

[2]We use https://pypi.org/project/rouge/ for $ROUGE-1,2,L$ computation.

each annotator as a target and pool the data for training. For ES and TTG, we report average and best ROUGE scores based on the three versions of references. For AID, for inconsistent labels from annotators, an expert decides the final manual label for training and evaluation. Table 3 shows statistics and IAA of SLP annotations on AMC. The moderate IAA values indicate great challenges for SLP annotations on meetings, which demand more studies.

3. TRACK SETTING AND EVALUATION

In this section, we introduce the five tracks in the MUG benchmark for benchmarking the performance of five SLP tasks on AMC, including task definitions and evaluation methods.

Track 1: Topic Segmentation (TS) The Topic Segmentation (TS) track requires segmenting the manual transcripts of a session into a sequence of non-overlapping, topically coherent segments. For evaluation, we use three standard evaluation metrics, including positive F_1, P_k [21], and *WinDiff* (*WD*) [22][3]. Both P_k and *WD* use a fixed sliding window over the document, with the window size usually set to half of the average true segment length. P_k is proposed to resolve the drawbacks of positive F_1, including the inherent trade-off between precision and recall and insensitivity to near-misses. While sliding the window, P_k determines whether the two ends of the window are in the same or different reference segments and increases a counter for a mismatch. The final score is computed by scaling the counter by the total number of measurements taken. However, P_k still has problems, including sensitivity to variation in segment size distribution, penalizing false negatives more than false positives, and over-penalizing near-misses. To remedy these issues, *WD* is proposed in [22] as an update to P_k. We use $1 - P_k$ and *1-WD* so that larger values indicate better performance, same as positive F_1.

Track 2: Topic-level and Session-level Extractive Summarization (ES) The Topic-level and Session-level Extractive Summarization track requires extracting key sentences for each reference topic segment and the entire session, without modifying original sentences. We use the standard ROUGE-1,2,L F-scores [19](R-1,R-2,R-L) as the evaluation metrics for the two subtasks of topic-level and session-level extractive summarization, respectively. Since the references are from three annotators, we report both average and best ROUGE-1,2,L scores based on the three references. We then average the ROUGE-1,2,L scores of the two subtasks as the final score for this track. Larger values indicate better performance.

Track 3: Topic Title Generation (TTG) The Topic Title Generation track requires generating an informative and concise title for each reference topic segment. Similar to the ES track, we use the standard ROUGE-1,2,L F-score evaluation metrics and report both average and best ROUGE-1,2,L scores based on the three references.

Track 4: Keyphrase Extraction (KPE) The Keyphrase Extraction track requires extracting top-K keyphrases from a session that can reflect its main content. We compute *exact* F_1 and *partial* F_1 at top-K between the predicted KPs and the labeled KPs as evaluation metrics. Exact F_1 is computed when predicted KPs exactly match the labeled KPs. For computing partial F_1, we define two KPs i, j as partially matched if $|LCS(i, j)| >= L$, where LCS denotes their longest common sequence and L is set to 2 characters on this dataset. Partial F_1 is computed correspondingly based on the partial match.

Track 5: Action Item Detection (AID) The Action Item Detection track requires detecting sentences containing information about actionable tasks as positive samples. We report positive precision, recall, and F_1 for this task.

---

[3]We use https://segeval.readthedocs.io/en/latest/ for P_k and *WD* computation.

	TS		Topic-level ES	Session-level ES	TTG	KPE	AID
IAA	N/A		49.53/30.50/41.13	55.65/28.40/34.97	30.79/16.63/28.17	55.62	0.50
	#Topics	Len.	Count/Topic	Count/Session	Len.	Count/Session	Count/Session
mean	9.81	996.1	2.41	10.81	11.26	17.37	3.22
std	2.22	353.9	0.66	2.93	1.85	3.53	3.86
25%	8	714	2	9	10	15	0
50%	9.5	950	3	10	11	17	2
75%	11	1230	3	12	13	20	5

Table 3. Statistics and Inter-Annotator Agreement (**IAA**) for all SLP annotations in the entire AliMeeting4MUG Corpus. Counts for KPE are computed on **union of three versions of annotations** while all other counts are based on individual annotations. For evaluating IAA, we report ROUGE F-scores R-1/R-2/R-L for Topic-/Session-level Extractive Summarization (ES) and Topic Title Generation (TTG) tasks, exact F_1 for Keyphrase Extraction (KPE), and Kappa coefficient for Action Item Detection (AID).

4. BASELINE SYSTEMS

One key challenge of the MUG benchmark is handling long-form documents (Section 1). The avg. session length of AMC is 10,772.5 tokens (Table 2). Hence, our baseline systems for all tracks carefully consider handling long-input text. Although Transformer [23] has become the SOTA architecture for sequence modeling on a wide variety of NLP tasks and transformer-based pre-trained language models (PLMs) become dominant in NLP, such as BERT [24], GPT-2 [25], and RoBERTa [26], the core self-attention mechanism has quadratic time and memory complexity to the input sequence length [23], limiting the max sequence length during pre-training (e.g., 512 for BERT) for a balance between performance and memory usage. The standard approaches for BERT-series PLMs to handle long input are to truncate the input into the max sequence length, or to use a sliding window and aggregate the model results on each window as the final results. However, these approaches may degrade performance due to losing contextual information. Consequently, for **baseline systems for the three understanding tasks of Track 1/2/5**, to better handle long-form documents than BERT-series PLMs, we build a unified framework of fine-tuning a pre-trained efficient Transformer, Longformer-base [27][4]. Longformer provides a drop-in replacement of self-attention in Transformer, achieving linear complexity by using a window-based self-attention to capture local context and a global attention to encode inductive bias about the task. We model TS/ES/AID as sequence labeling tasks. We use a fixed sliding window with size 4096 and allow one sentence overlap. We set batch size as 2 and learning rate as 5e-5. We also evaluate BERT-base[5] and Longformer-base with 512 sliding window size and find that Longformer-base with 4096 window size outperforms both of them on these tasks. **The baseline system for the generative TTG track** is based on fine-tuning the pre-trained BART-base model [28][6]. We truncate input to the max sequence length 512 of this model. We set learning rate as 3e-5 with warmup_steps 300 and linear lr_scheduler, batch size 16, and use label smoothing [29] with $\beta = 0.1$. For baseline systems for TS/ES/TTG/AID, we use Adam optimizer and set dropout prob as 0.1. We train 10 epochs for TS/ES/AID and 5 epochs for TTG and select the best-performing model on the Dev set, respectively. **The baseline system for the KPE track** uses an unsupervised approach YAKE [30], which employs document-based statistical features for ranking candidates. Our earlier work [31] shows that YAKE performs relatively stable on documents with varying lengths, es-

pecially on long documents, compared to many other unsupervised approaches. We evaluate YAKE[7] with the default parameter settings (ngram=3, window_size=1) except Jieba [8] for Chinese word segmentation on input documents (scoring exact/partial F_1 does not need to segment reference KPs) and an extended stopword list for both Chinese and English. Table 4 reports baseline system results on the *TSonly-Test1* set for Track1 and *exceptTS-Test1* set for Track2-5 (Section 2) with evaluation metrics described in Section 3. We find that TS performance on AMC **manual** transcripts is worse than that from applying the same baseline system on QMSUM **ASR** transcripts, and our baseline TTG results on AMC **manual** transcripts are worse than abstractive summarization SOTA on **ASR** transcripts of AMI,ICSI and QMSUM [32, 33, 34], all suggesting that the SLP tasks on AMC are quite challenging.

Track 1 Topic Segmentation (TS)			
Model	positive F_1	$1 - p_k$	1-WD
Longformer	$22.7_{\pm 0.98}$	$0.583_{\pm 0.008}$	$0.56_{\pm 0.008}$
Track 2 Extractive Summarization (ES) (AVG)			
Model	R-1 Avg./Best	R-2 Avg./Best	R-L Avg./Best
Longformer	$53.83_{\pm 0.39}/61.64_{\pm 0.68}$	$32.33_{\pm 0.60}/42.73_{\pm 0.84}$	$42.94_{\pm 0.61}/53.87_{\pm 0.68}$
Topic-level ES			
Model	R-1 Avg./Best	R-2 Avg./Best	R-L Avg./Best
Longformer	$51.16_{\pm 0.68}/63.0_{\pm 1.03}$	$34.4_{\pm 0.78}/49.61_{\pm 1.19}$	$45.03_{\pm 1.02}/59.61_{\pm 1.2}$
Session-level ES			
Model	R-1 Avg./Best	R-2 Avg./Best	R-L Avg./Best
Longformer	$56.5_{\pm 0.94}/60.28_{\pm 1.2}$	$30.26_{\pm 0.77}/35.85_{\pm 1.07}$	$40.84_{\pm 0.53}/48.13_{\pm 0.43}$
Track 3 Topic Title Generation (TTG)			
Model	R-1 Avg./Best	R-2 Avg./Best	R-L Avg./Best
BART	$32.16_{\pm 0.21}/45.11_{\pm 0.22}$	$17.87_{\pm 0.22}/28.26_{\pm 0.32}$	$30.1_{\pm 0.26}/43.16_{\pm 0.22}$
Track 4 Keyphrase Extraction (KPE)			
Model	Exact/Partial F_1@10	Exact/Partial F_1@15	Exact/Partial F_1@20
YAKE	15.2/24.9	17.5/27.8	19.1/29.5
Track 5 Action Item Detection (AID)			
Model	positive P	positive R	positive F_1
Longformer	$60.18_{\pm 5.06}$	$66.89_{\pm 3.29}$	$63.14_{\pm 1.41}$

Table 4. Baseline system performance on the {*TSonly,exceptTS*}-*Test1* sets for the five tracks. Each mean and std are computed on results from 5 runs with different random seeds.

5. CONCLUSION

We establish a general and comprehensive Meeting Understanding and Generation benchmark (**MUG**) to drive the spoken language processing (SLP) research on meetings. To facilitate MUG, we construct the AliMeeting4MUG Corpus. We define SLP tasks, build and evaluate baseline systems. Next, we plan to add tasks such as QA and Session-level Abstractive Summarization, cover more languages such as English, and facilitate multi-modality MUG research.

---

[4]https://huggingface.co/IDEA-CCNL/Erlangshen-Longformer-110M
[5]https://huggingface.co/bert-base-chinese
[6]https://huggingface.co/fnlp/bart-base-chinese

[7]https://github.com/LIAAD/yake
[8]https://github.com/fxsjy/jieba

6. REFERENCES

[1] Parviz Ajideh and Bio Data, "Schema-theory based considerations on pre-reading activities in esp textbooks," *The Asian EFL Journal Teaching Articles*, pp. 1–19, 2006.

[2] Peter D. Turney, "Learning algorithms for keyphrase extraction," *Inf. Retr.*, vol. 2, no. 4, pp. 303–336, 2000.

[3] Iryna Gurevych and Michael Strube, "Semantic similarity applied to spoken dialogue summarization," in *COLING*, 2004.

[4] Gerard Salton, Amit Singhal, Chris Buckley, and Mandar Mitra, "Automatic text decomposition using text segments and text themes," in *The Seventh ACM Conference on Hypertext*, David Stotts, Ed., 1996, pp. 53–65.

[5] Marti A. Hearst and Christian Plaunt, "Subtopic structuring for full-length document access," in *ACM-SIGIR*, Robert Korfhage, Edie M. Rasmussen, and Peter Willett, Eds., 1993, pp. 59–68.

[6] Carl Gutwin, Gordon W. Paynter, Ian H. Witten, Craig G. Nevill-Manning, and Eibe Frank, "Improving browsing in digital libraries with keyphrase indexes," *Decis. Support Syst.*, vol. 27, no. 1-2, pp. 81–104, 1999.

[7] Gaël Dias, Elsa Alves, and José Gabriel Pereira Lopes, "Topic segmentation algorithms for text summarization and passage retrieval: An exhaustive evaluation," in *AAAI*. 2007, pp. 1334–1339, AAAI Press.

[8] J. Koay et al., "How domain terminology affects meeting summarization performance," in *COLING*, 2020.

[9] K. Chen et al., "A large-scale chinese long-text extractive summarization corpus," in *ICASSP*, 2021.

[10] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, and Chuck Wooters, "The ICSI meeting corpus," in *ICASSP*. 2003, pp. 364–367, IEEE.

[11] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner, "The AMI meeting corpus: A pre-announcement," in *MLMI*. 2005, vol. 3869 of *Lecture Notes in Computer Science*, pp. 28–39, Springer.

[12] Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar, "ELITR minuting corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech," in *LREC*, 2022, pp. 3174–3182.

[13] Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev, "QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization," in *NAACL*, 2021.

[14] Susanne Burger, Victoria MacLaren, and Hua Yu, "The ISL meeting corpus: the impact of meeting type on speech style," in *INTERSPEECH*, John H. L. Hansen and Bryan L. Pellom, Eds., 2002.

[15] Martial Michel, Jerome Ajot, and Jonathan G. Fiscus, "The NIST meeting room corpus 2 phase 1," in *MLMI*, Steve Renals, Samy Bengio, and Jonathan G. Fiscus, Eds., 2006, vol. 4299 of *Lecture Notes in Computer Science*, pp. 13–23.

[16] Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu, "M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge," in *ICASSP*. 2022, pp. 6167–6171, IEEE.

[17] Qinglin Zhang, Qian Chen, Yali Li, Jiaqing Liu, and Wen Wang, "Sequence model with self-adaptive sliding window for efficient spoken document segmentation," in *ASRU*. 2021, pp. 411–418, IEEE.

[18] Alexander Gruenstein, John Niekrasz, and Matthew Purver, "Meeting structure annotation: Data and tools," in *SIGDIAL*, Laila Dybkjær and Wolfgang Minker, Eds. 2005, pp. 117–127, Special Interest Group on Discourse and Dialogue (SIGdial).

[19] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, Barcelona, Spain, July 2004, pp. 74–81, Association for Computational Linguistics.

[20] Jean Carletta, "Assessing agreement on classification tasks: The kappa statistic," *Computational Linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[21] Doug Beeferman, Adam L. Berger, and John D. Lafferty, "Statistical models for text segmentation," *Mach. Learn.*, vol. 34, no. 1-3, pp. 177–210, 1999.

[22] Lev Pevzner and Marti A. Hearst, "A critique and improvement of an evaluation metric for text segmentation," *Computational Linguistics*, vol. 28, no. 1, pp. 19–36, 2002.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, Jill Burstein, Christy Doran, and Thamar Solorio, Eds. 2019, pp. 4171–4186, Association for Computational Linguistics.

[25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9, 2019.

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[27] Iz Beltagy, Matthew E. Peters, and Arman Cohan, "Longformer: The long-document transformer," *CoRR*, vol. abs/2004.05150, 2020.

[28] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL*, 2020, pp. 7871–7880.

[29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[30] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt, "Yake! collection-independent automatic keyword extractor," in *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR*. 2018, vol. 10772 of *Lecture Notes in Computer Science*, pp. 806–810, Springer.

[31] Linhan Zhang, Qian Chen, Wen Wang, Chong Deng, Shiliang Zhang, Bing Li, Wei Wang, and Xin Cao, "Mderank: A masked document embedding rank approach for unsupervised keyphrase extraction," in *ACL*, 2022, pp. 396–409.

[32] C. An et al., "Retrievalsum: A retrieval enhanced framework for abstractive summarization," *CoRR*, vol. abs/2109.07943, 2021.

[33] M. Zhong et al., "Dialoglm: Pre-trained model for long dialogue understanding and summarization," in *AAAI*, 2022.

[34] W. Xiong et al., "Adapting pretrained text-to-text models for long text sequences," *CoRR*, vol. abs/2209.10052, 2022.