发表于 *Interspeech 2025* 论文集，第 276–280 页。DOI：
https://doi.org/10.21437/Interspeech.2025-2792 论文集：
https://www.isca-archive.org/interspeech_2025/

arXiv:2601.02128v1 [cs.CL] 5 Jan 2026

迈向多层级转录文本分段：用于目录生成的 LoRA 微调

Steffen Freisinger、*Philipp Seeberger*、*Thomas Ranzenberger*、*Tobias Bocklet*、*Korbinian Riedhammer*

纽伦堡应用技术大学，德国 y

firstname.lastname@th-nuernberg.de

摘要

将语音转录文本划分为主题小节，既有利于下游处理，也能帮助依赖书面文本获取信息的用户。我们提出一种全新的转录文本层级主题分割方法，可生成多层级目录，标示主题和子主题的边界。我们比较了大语言模型上的零样本提示与 LoRA 微调，并探索了高层级语音停顿特征的整合方式。在英语会议录音和多语言讲座转录文本（葡萄牙语、德语）上的评估结果表明，该方法显著优于已有的主题分段基线方法。此外，我们还针对多层级分割调整了一种常用评估指标，使其能够通过单一指标综合考量所有层级。

索引术语：主题分割、口语内容分段、目录生成、层次分段

1. 引言

口语内容的主题分割对于许多语音技术应用至关重要，尤其是在处理讲座、会议和研讨会等时长较长或涉及多位说话人的录音时 [1, 2, 3]。与通常按照章节或小节组织的书面文本（如书籍、论文等）不同，自动生成的口语转录文本缺乏这种半结构化信息。尽管自动语音识别（ASR）技术不断进步，但缺乏结构仍使用户难以查找特定信息，而信息检索（IR）和检索增强生成（RAG）等下游任务也依赖有意义的分段。

以往大多数关于口语文档分割的研究都采用传统的线性分段方法，即根据时间维度将语音转录文本划分为多个片段。然而，主题往往存在多个粒度层级，而单纯从时间维度出发的视角完全忽视了这些组合结构。例如，像“项目需求”这样的宽泛主题可能包含“用户界面”和“测试策略”等更具体的子主题。将这些子主题视为与更广泛的主题同等的内容，会破坏其固有的层级关系，导致结构清晰度下降。层次化方法不仅能改善所有用户的内容导航体验，还能通过提供清晰的多层级书面结构，帮助有听力障碍的人士。这有助于实现对口语内容的包容性访问，而转录文本则是获取信息的主要途径。

诸如 LLaMA[1], Qwen[2],和 GPT-4[3] 这样的大语言模型（LLM）展现出强大的语言理解能力，并且在摘要生成[4, 5] 和报告生成 [6], 等需要理解连贯内容的任务中取得了显著进展。此外，它们能够进行零样本学习和上下文学习，结合任务描述、大规模预训练和指令遵循数据，解锁了新的能力。然而，将大语言模型应用于口语内容的目录（ToC）生成仍未得到充分探索。

本研究受到近期将大语言模型用于相关任务所取得成果的启发 [4, 5, 6]。我们提出一种方法，以弥合大语言模型在目录生成方面的潜力与层次化主题分段之间的差距。我们的目标是为口语内容自动生成层次化主题结构[4]。

我们的主要贡献总结如下：

- 我们基于大语言模型摘要技术的进展，提出了一种用于转录文本层次化主题分段的新方法。
- 我们证明了该方法优于现有的分段基线方法，并在三个数据集上评估了线性分割任务。
- 在多层级分割方面，我们调整了一种常见的线性分段指标，使其能够考量所有层次。

2. 相关工作

2.1. 口语文档分割

口语文档分割适用于多个领域，包括讲座视频 [3, 7], 会议 [8, 9], 广播节目 [2, 10], 对话 [11], 播客 [12, 13], 和 YouTube 视频 [14]。大多数研究聚焦于文本转录稿的分割，也有一系列研究将声学特征纳入其中。TextTiling [15] 通过分析句子间的词汇偏差来确定主题转变，还可以结合预训练的句子嵌入 [16] 或随机生成的超维向量 [1] 加以扩展。典型的监督方法将分割视为序列标注任务，把句子标记为主题变化或无变化，并使用基于长短期记忆网络和 Transformer 的模型 [11, 17]。其他研究则将其视为分类问题，根据句子的左、右上下文 [18, 19], 判断给定的句子边界是否标志着主题变化。除了文本之外，其他方法还会纳入停顿、音高、音量和说话人变化等声学线索 [7, 20] ，以及预训练音频模型生成的嵌入 [2, 10]。尽管口语文档分割已取得进展，研究重点仍主要放在扁平层级上，而忽略了层级结构。

[1]https://arxiv.org/abs/2307.09288[2] https://arxiv.org/abs/2412.15115[3] https://arxiv.org/abs/2303.08774

[4]代码和提示词：https://github.com/steffrs/toc-llm

2.2. 层次主题分割

与传统的主题分割不同，层次主题分割会识别多个层级上的主题转变。早期研究将词汇衔接度量与对话特征相结合，以检测多方对话中的顶层边界和子主题边界 [21]。后续方法利用贝叶斯多尺度模型和亲和传播，在文本中构建嵌套主题结构 [22, 23]。层次方法也被用于口语对话摘要生成。在 [24], 中，作者提出了一种两阶段系统，将自动语音识别与基于文本的摘要生成相结合，并采用专门的分割和合并算法处理较长的对话数据。尽管该方法并非专门针对分割任务，但它凸显了多层级组织对于提升可读性的价值。

2.3. 目录生成

主题分割的一种特殊形式是目录生成，它将分割与标题生成结合起来。 [25] 使用层次判别模型为教科书生成目录。他们的方法分别结合局部和全局模型，用于候选标题生成和候选标题选择。这里，全局模型旨在确保目录结构的连贯性。同样， [26] 研究了自动识别层级的方法，利用维基百科和文学作品等多样化语料库中的语言特征和监督分类。为了解决数据稀疏问题， [27] 采用半监督学习，将词簇融入结构化模型，以提升长文本的分割效果。针对语音转录文本， [28] 在概念词向量上应用层次聚类，从会议转录文本中提取分层主题结构。相比之下，PODTILE [29]引入了一种经过微调的 Transformer，用于生成带标题的播客章节，但未纳入层级结构。尽管早期研究探索了这一方向，但在口语文档分割中，使用大语言模型端到端生成目录仍鲜有研究。

3. 方法

3.1. 问题定义

设 D 是由 n 个句子 $\{s_1, \ldots, s_n\}$ 组成的文档。线性划分将 D 划分为一组连续片段 $\{S_1, \ldots, S_m\}$，使得 $\bigcup_{i=1}^{m} S_i = D$ 且任意两个片段互不重叠。我们将其扩展为层次化划分

划分 跨越 L 个层级，其中每个层级 l 也定义了对 D的完整划分。第 l 层级的每个边界都会保留在第 $l+1$层级。形式上，如果 S_i^l 是第 l层级的一个片段，它可以在第 $l+1$层级保持完整，也可以拆分为更小的连续子片段 $\{S_i^l\} \rightarrow \{S_{i,1}^{l+1}, S_{i,2}^{l+1}, \ldots\}$。由此形成嵌套结构，从 $l = 1$ 处较粗略的顶层主题，到 $l = L$处更细致的子主题。对于线性分段实验，我们仅单独考虑最低层级 L 。

3.2. ToC-LLM：目录生成

我们提出一种方法，利用大语言模型直接从语音转录文本生成多级目录（参见图 1）。与仅预测边界位置不同，我们的方法生成结构化大纲，使每个片段都能标注其层级。具体做法是通过系统提示指示模型为整篇转录文本生成目录。输入的每一行恰好包含

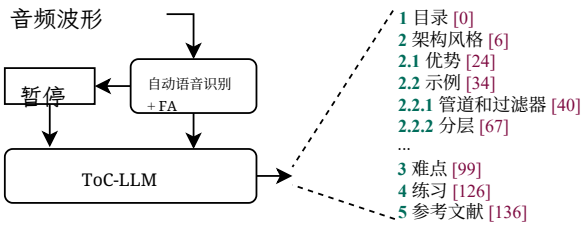


图 1： 我们方法的概览，以VIDEOAULA （已翻译）中的示例为例：首先，我们使用自动语音识别系统生成转录文本，并通过强制对齐（*FA*）使转录文本与音频同步。然后，我们提取句间停顿时长，并将其与文本一同用于提示大语言模型。输出为目录，展示了主题边界。

一个句子，并标注其对应的索引 [30, 31]。最终输出采用点分编号（例如 2.2.1）表示片段的层级，同时包含主题名称及对应的句子索引。图 1 展示了生成结构的示例。

3.3. 融入声学线索

为了探究声学线索对主题分段的影响，我们选择性地将句间停顿时长纳入转录文本。既有研究表明，停顿是主题边界的重要指标 [2, 3, 32]。在我们的设置中，每行句子都可以包含停顿标注（例如，42（停顿=0.62s）：现在让我们继续……），使模型能够识别这些时间间隔。由于预处理过程中进行了精确对齐（参见第 4.1 节），我们可以直接根据句子边界获取停顿区间，无需额外的音频处理。此可选功能旨在利用高层次的声学信息来改进边界检测。

3.4. 建模与实现

我们的主要模型是*MistralNemo*[5]，，这是一个在多语言数据上训练的 120 亿参数大语言模型。它支持 128k 令牌的上下文窗口，因此适用于不同语言的长篇转录文本。作为比较，我们还采用 *Qwen2.5- 7B*[6]，，这是一个拥有 76.1 亿参数、上下文长度相同且支持多语言的模型。这两个模型都经过指令微调，能够更好地处理自然语言查询。我们将这两个模型分别称为 TOC-NEMO 和 TOC-QWEN。

我们使用来自*unsloth*库的4位量化检查点[7][8]，以减少内存用量并加快推理。为探索领域适配，我们比较零样本场景（不进行微调）与基于 LoRA 的微调 [33]。LoRA 会冻结主模型权重，并更新一组低秩适配器。结合4位量化后，这种方法已显示出能显著降低内存占用和训练时间，同时保持与全精度微调相当的性能 [34]。在零样本和微调配置中，我们都采用相同的提示格式，以确保比较公平。

[5]https://mistral.ai/news/mistral-nemo [6] https://arxiv.org/abs/241215115 7 . *unsloth/Qwen2.5-7B-Instruct-unsloth-bnb-4bit*[8] *unsloth/Mistral-Nemo-Instruct-2407-bnb-4bit*[9] https://github.com/unslothai/unsloth

4. 实验

4.1. 数据集

在实验中，我们使用了三个主题分割数据集，并在表 2 中进行了概述。AMI 语料库 [9] 包含 171 场英语会议，其中 139 场会议场次带有主题分割标注。我们将 AMI 的主题区间标注调整为我们的分割方案（第 4.2 节）：每当出现间隔时，就插入一个填充主题。这样可以保留仅预测主题变化的做法 [14, 16, 17, 18]。为了构建层次结构，我们将任何在较大片段内开始的主题视为子主题。不过，与另外两个数据集相比，AMI 的层次深度相对较浅。VIDEOAULA 数据集 [3] 包含 34 场关于计算机科学主题的葡萄牙语视频讲座。每场讲座均由不同的说话人主讲，并包含主题片段标注。LECTUREDE [32] 是一个内部收集的德语视频讲座集，内容涵盖工程学、社会科学和教学法。我们为这些录音手动标注了层次主题标签，以便开展多层级分割实验。

对于所有数据集，我们都采用相同的预处理流程。首先，我们使用 Whisper large-v2[35]，（一种基于 68 万小时语音训练的端到端多语言自动语音识别模型）生成转录文本。接下来，我们使用 Montreal Forced Aligner[36] 将转录文本与相应的音轨同步。最后，我们使用 SpaCy[10]将对齐后的文本切分为句子。该流程旨在模拟内容自动转录和分段的真实场景。

4.2. 层次化评估

在评估生成层次分段的方法时，标准的单层级指标（例如，F_1、B或 P_k）会丢失更精细的细节，或迫使人们在评估时只能选择一个层级。为解决这一问题，我们提出了一种层级指标包装器，它能为多层级分段生成单一数值指标，同时保留所有粗粒度和细粒度层级。我们采用一种简单的动态规划方法，将每个参考层级与一个非递减的假设层级对齐，这与单调序列对齐方法（例如动态时间规整）相似，从而确保参考层级由粗到细的进展与假设层级以全局最优的方式匹配。

设 D为一个包含 n个句子的文档，并假设我们在每个层级 $l \in \{1, \ldots, L_{\text{ref}}\}$都有一个参考 层次分段 $\{S_1^l, \ldots, S_{m_l}^l\}$。同样，设 $\{H_1^k, \ldots, H_{p_k}^k\}$为每个层级 $k \in \{1, \ldots, L_{\text{hyp}}\}$的假设 分段。对于每个单独的层级，我们可以计算一个线性度量 $M(R^l,\ H^k)$，例如边界相似度 B [37]。

为了定义一个单一的层次 分数，我们要求：如果参考分段从较粗的层级 l 推进到较细的层级，$l+1$，那么假设分段不能“向上回退”到较不精细的层级。形式上，我们选择一个满足以下条件的函数 ϕ：

$$\phi : \{1, \ldots, L_{\text{ref}}\} \to \{1, \ldots, L_{\text{hyp}}\}, \quad \text{where} \quad \phi(l+1) \geq \phi(l) \quad \forall l. \tag{1}$$

在此约束下，我们寻求能够最大化 总边界相似度的 ϕ^*：

$$\phi^* = \underset{\phi(\cdot)}{\operatorname{argmax}} \sum_{l=1}^{L_{\text{ref}}} B\big(R^l, H^{\phi(l)}\big), \text{subject to } \phi(l+1) \geq \phi(l). \tag{2}$$

最后，我们通过对所有参考层级取平均值来定义层次边界相似度：

$$B_{\text{hier}}(R, H) = \frac{1}{L_{\text{ref}}} \sum_{l=1}^{L_{\text{ref}}} B\big(R^l, H^{\phi^*(l)}\big). \tag{3}$$

因此，每个参考层级 l 都与使其局部边界指标达到最大的假设层级相匹配，同时遵守公式（1）中的非递减约束。公式（2）和（3）由此得出一个标量，用于概括整个嵌套 分段从粗到细的对齐程度。

4.3. 基线方法

我们将提出的方法与几种线性基线方法进行比较，以探究它们在解决主题片段划分任务方面的效果：

TT-BERT：作为无监督基线方法，我们采用 TextTiling [15] 的变体，并结合句子嵌入 [16]。为计算嵌入，我们使用来自 *SentenceTransformers* 库 [38]的预训练文本编码器[11] 。

MINISEG：作为有监督基线方法，我们采用 MINISEG [14], 的序列标注方法，该方法基于 MiniLM [39] 和 RoFormer [40]。对于每个句子，我们判断是否存在主题边界。

CS-BERT：另一种有监督基线方法是 cross-segmentBERT [18],，其中包含二元片段边界分类。我们将左侧和右侧上下文拼接起来，并使用 ［CLS］标记进行分类。

SEGMENTLLM：最后，我们采用 [30]提出的基于提示的方法。该零样本方法指示大语言模型生成一个句子索引列表，其中每个列表代表一个主题片段。

4.4. 线性分段结果

表1总结了线性分段结果，我们采用最细粒度的分段层级（参见第3.1节）。表1中的所有监督方法均在 WIKI-727语料库的前10,000篇文章上进行预训练， [17], 与 [14, 19]类似。对于 AMI，我们按照 [18]的做法，报告100个自助法测试集上的均值 ± 标准差。对于 VIDEOAULA 和 LECTUREDE，我们采用留一说话人交叉验证策略，并报告所有验证集上的均值 ± 标准差。每个验证集的指标按该集合中所有样本的平均值计算。

在微调后的模型中，加入暂停时长的 TOC-NEMO 在所有数据集上均取得最佳表现。这表明，在应用监督微调时，模型能够利用额外的声学线索。相比之下，暂停时长会降低零样本版本的得分，这说明这些线索需要经过特定训练才能带来提升。

使用 ChatGPT 的 SEGMENTLLM [30] 取得了最高的零样本得分。然而，当 SEGMENTLLM 和我们的目录提示都使用 Nemo 模型时，目录方法在所有数据集上的表现始终优于 SEGMENTLLM。这一发现凸显了目录格式的优势，即使评估的仅是线性分段。我们没有在 LECTUREDE 上运行 SEGMENTLLM CHATGPT 方法，因为隐私限制禁止对该数据集进行外部云端处理。

[10] https://github.com/explosion/spaCy

[11] *paraphrase-multilingual-mpnet-base-v2*

表1： 不同数据集上的方法比较。我们报告*100*个自助法测试集（AMI）上的均值 ± 标准差，以及留一说话人交叉验证集（VIDEOAULA、LECTUREDE）上的均值 ± 标准差。✓ 表示在相应的目标训练集上进行微调，✗ 表示零样本设置。每组中的最佳分数以粗体突出显示。

		AMI		VIDEOAULA		LECTUREDE	
方法	**调优**	$F_1 \uparrow$	$B \uparrow$	$F_1 \uparrow$	$B \uparrow$	$F_1 \uparrow$	$B \uparrow$
TT-BERT [16]	✗	7.25 ± 1.47	6.71 ± 0.97	9.42 ± 2.45	8.73 ± 2.87	6.94 ± 4.43	8.31 ± 2.55
SEGMENTLLM CHATGPT [30]	✗	**10.35** ± 2.14	**7.97** ± 1.31	**27.57** ± 9.27	**22.29** ± 8.32	-	-
SEGMENTLLM NEMO 参见 [30]	✗	5.42 ± 1.44	5.79 ± 1.02	4.92 ± 2.64	4.68 ± 2.18	9.87 ± 9.17	11.37 ± 10.22
TOC-NEMO	✗	7.36 ± 1.51	7.68 ± 1.07	12.27 ± 6.45	14.06 ± 5.80	20.55 ± 5.97	17.04 ± 4.98
TOC-NEMO + PAUSE	✗	6.50 ± 2.23	7.05 ± 1.72	9.90 ± 3.14	12.4 ± 3.79	17.5 ± 7.64	15.79 ± 6.26
MINISEG [14]	✓	18.61 ± 2.54	14.90 ± 1.68	38.72 ± 7.80	28.28 ± 4.42	23.86 ± 11.18	19.21 ± 8.11
CS-BERT [18]	✓	20.50 ± 2.71	15.77 ± 1.78	47.77 ± 8.47	34.22 ± 6.09	31.08 ± 5.99	22.88 ± 6.37
TOC-NEMO	✓	26.55 ± 2.92	22.79 ± 2.40	64.26 ± 7.86	51.94 ± 8.10	58.47 ± 15.86	49.50 ± 17.04
TOC-NEMO + PAUSE	✓	**30.34** ± 3.28	**24.81** ± 2.30	**67.34** ± 6.57	**55.18** ± 6.42	**58.89** ± 13.65	**49.77** ± 14.72
TOC-QWEN +PAUSE	✓	24.71 ± 3.91	18.68 ± 2.95	45.69 ± 14.72	35.95 ± 13.59	51.05 ± 17.89	42.20 ± 18.40

表 2： 三个数据集的统计信息。

数据集	**记录**	**时长**	**说话人**	**层级**
AMI	139	73h	-	3
VIDEOAULA	34	23h	5	4
LECTUREDE	96	27h	4	5

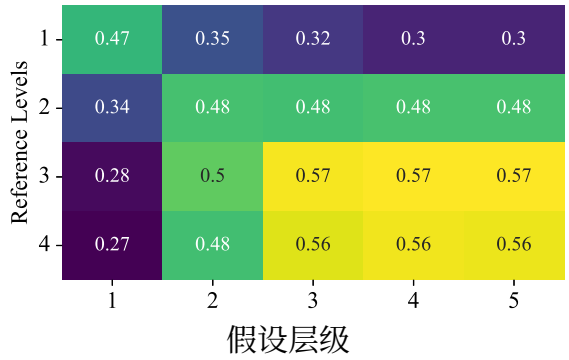


图 2：特定层级的 B 参考层级（行）与假设层级（列）各组合的得分。对 TOC-NEMO + PAUSE 在 VIDEOAULA 上的所有样本取平均

4.5. 层次分段结果

表1中的线性得分无法说明我们的方法是否真正生成了嵌套结构，即包含粗粒度顶层片段和更细粒度子层级片段。为此，我们采用了第 4.2 节介绍的扩展边界相似度指标（B_{hier}）。该指标评估假设层级与参考层级的匹配情况，从而反映主题的嵌套组织。

图 2 展示了 VIDEOAULA 数据集在不同参考层级和假设层级上的 B分数，这些分数是对所有讲座取平均后得出的。所选方法为 TOC-NEMO + PAUSE。如图所示，比较粒度相近的层级时，得分最高。这表明模型成功捕捉了顶层的宽泛主题和下层的细粒度主题，生成了有意义的目录结构。T

abl报告 B_{hier} for th零样 h本方法hods from Ta-

表3： 不同零样本方法和数据集的层级 B 分数（与线性 B 分数的差异）。

方法	**AMI**	**VIDEOAULA**	**LECTUREDE**
TT-BERT	6.70（-0.01）	8.64（-0.09）	7.92（-0.39）
S.LLM CHATGPT	8.49 (+0.52)	20.2 (-2.09)	-
S.LLM NEMO	5.77 (-0.02)	4.17 (-0.51)	8.72 (-2.65)
TOC-NEMO	8.50 (+0.82)	14.01 (-0.05)	21.56 (+4.52)
TOC-NEMO + P.	8.10 (+1.05)	12.00 (-0.40)	21.73 (+5.94)

表 1 还列出了与线性 B 分数的差异。在层级结构更深的 LECTUREDE 数据集上（见表 2），仅线性分段器 SEGMENTLLM NEMO 的层级指标低于其标准线性 B分数。相反，专为多层级输出设计的方法，其 B_{hier} 值高于线性 B，凸显了显式建模嵌套主题的优势。对于标注相对浅层的 AMI 数据集，所有方法的B与 B_{hier}之间差距都不太明显。这些发现凸显了层级分段在结构上的附加价值，也表明评估指标需要超越简单的线性边界。

5. 结论

我们提出了一种层次化主题分段方法，可从冗长的语音转录文本中生成多层级目录。实验表明，经过微调后，大语言模型能够超越传统基线方法。纳入停顿时长线索可进一步提升分段效果。此外，我们还提出了一种层级评估方法，相较于线性指标，它能提供更深入的见解。

尽管如此，零样本模型的表现仍存在显著差距，这表明我们需要更有效的解决方案，避免依赖来自目标领域或语言的监督数据。我们希望本研究的发现能推动层次分段领域的进一步研究，为自动生成的转录文本增添有价值的结构。

6. 参考文献

[1] S. Park、J. Seo 和 J. Lee，《超高维空间中的无监督对话主题分段》，载于*INTERSPEECH*，2023，

第 730–734 页。[2] I. Ghinassi、L. Wang、C. Newell 和 M. Purver，《利用预训练神经编码器对播客节目进行多模态主题分段》，载于 ACM ICMR 会议论文集，2023，第 602–606 页。[3] E. R. Soares 和 E. Barr´ere，《利用低层和高层音频特征对视频讲座进行自动主题分段》，载于 WebMedia 会议论文集，2018，第 189–196 页。[4] H. Shang、Z. Li、J. Guo、S. Li、Z. Rao、Y. Luo、D. Wei 和 H. Yang，《利用大语言模型进行端到端语音摘要》，载于 Interspeech 2024，2024，第 1950–1954 页。[5] W. Kang 和 D. Roy，《利用音频提示大语言模型进行通用语音摘要》，载于 Interspeech 2024，2024，第 1955–1959 页。[6] R. G. Reddy 等，《Smartbook：借助 AI 为情报分析师生成情况报告》，2024。[7] D. Chand 和 H. O˘gul，《一种基于提取的语音内容进行讲座视频分段的框架》，2021 IEEE SAMI，第 000 299–000 304 页，2021。[8] A. Janin 等，《ICSI 会议语料库》，载于 ICASSP，第 1 卷，2003，第 I–I 页。[9] I. Mccowan 等，《AMI 会议语料库》，《行为研究方法与技术国际会议》，2005。[10] O. Berlage、K. Lux 和 D. Graus，《利用音频嵌入改进广播节目自动分段》，CoRR，第 abs/2002.05194 卷，2020。[11] L. Zhang 和 Q. Zhou，《对话流的主题分割》，载于 2019 年亚太信号与信息处理协会年度峰会暨会议（APSIPA ASC），2019，第 1036–1043 页。[12] A. Clifton、S. Reddy、Y. Yu、A. Pappu、R. Rezapour、H. Bonab、M. Eskevich、G. Jones、J. Karlgren、B. Carterette 和 R. Jones，《十万播客：英语口语文档语料库》，载于 COLING 会议论文集，2020，第 5903–5917 页。[13] E. Garmash 等，《Cem mil podcasts：用于多模态、多语言和多方言信息获取研究的葡萄牙语口语文档语料库》，载于《实验性信息检索与多语言、多模态及交互》，2023，第 48–59 页。[14] F. Retkowski 和 A. Waibel，《从文本分段到智能章节划分：一种用于组织视频转录文本的新型基准》，载于 EACL 会议论文集，Y. Graham 和 M. Purver 编，2024，第 406–419 页。[15] M. A. Hearst，《TextTiling：将文本分割为多段落子主题片段》，《计算语言学》，第 23 卷，64，1997 1，33第。–。[16] A. Solbiati、K. Heffernan、G. Damaskinos、S. Poddar、S. Modi 和 J. Cali，《利用 BERT 嵌入对会议进行无监督主题分段》，2021。[17] O. Koshorek、A. Cohen、N. Mor、M. Rotman 和 J. Berant，《将文本分段视为监督学习任务》，载于 NAACL 会议论文集。ACL，2018，第 469–473 页。[18] M. Lukasik、B. Dadachev、G. Sim˜oes 和 K. Papineni，《通过跨分段注意力进行文本分段》，CoRR，第 abs/2004.14535 卷，2020。[19] R. Ghosh、H. S. Kajal、S. Kamath、D. Shrivastava、S. Basu 和 S. Srinivasan，《真实场景中的主题分段：迈向半结构化和非结构化聊天的分段》，2022。[20] E. R. Soares 和 E. Barr´ere，《一种利用 word2vec 和声学特征进行视频讲座时间分段的优化模型》，载于 WebMedia 会议论文集，2019，第 513–520 页。[21] P.-Y. Hsueh、J. D. Moore 和 S. Renals，《多方对话的自动分段》，载于 EACL 会议论文集，D. McCarthy 和 S. Wintner 编，2006，第 273–280 页。[22] J. Eisenstein，《基于多尺度词汇衔接的层次化文本分段》，载于 NAACL 会议论文集，M. Ostendorf、M. Collins、S. Narayanan、D. W. Oard 和 L. Vanderwende 编，2009，第 353

[23] A. Kazantseva 和 S. Szpakowicz，“基于亲和传播的层次化主题分段”，载于*COLING*论文集，J. Tsujii 和 J. Hajic 编，2014 年，第 37–47 页。[24] D. Li、T. Chen、A. Tung 和 L. B. Chilton，“长篇口语对话的层次化摘要”，载于第 *34* 届 *ACM* 用户界面软件与技术年度研讨会，2021 年，第 582–597 页。[25] S. R. K. Branavan、P. Deshpande 和 R. Barzilay，“生成目录”，载于*ACL*论文集，A. Zaenen 和 A. van den Bosch 编，2007 年，第 544–551 页。[26] N. Erbs、I. Gurevych 和 T. Zesch，“用于自动生成目录的层级识别”，载于*RANLP*论文集，R. Mitkov、G. Angelova 和 K. Bontcheva 编，2013 年，第 252–260 页。[27] V. C. Nguyen、L. M. Nguyen 和 A. Shimazu，“一种用于生成目录的半监督方法”，载于*RANLP*论文集，G. Angelova 和 R. Mitkov 编，2009 年，第 312–317 页。[28] K. Ohtsuki、N. Hiroshima、Y. Hayashi、K. Bessho 和 S. Matsunaga，“用于会议索引的主题结构提取”，载于*Interspeech 2004*，2004 年，第 305–308 页。[29] A. Ghazimatin 等，“PODTILE：借助自动生成章节促进播客单集浏览”，载于*CIKM*论文集，2024 年，第 4487–4495 页。[30] Y. Fan、F. Jiang、P. Li 和 H. Li，“发掘 ChatGPT 在对话话语分析中的潜力：一项实证研究”，载于*LREC-COLING*论文集，N. Calzolari、M.-Y. Kan、V. Hoste、A. Lenci、S. Sakti 和 N. Xue 编，2024 年，第 16 998–17 010 页。[31] H. Yu、C. Deng、Q. Zhang、J. Liu、Q. Chen 和 W. Wang，“通过增强连贯性建模改进长文档主题分段模型”，载于*EMNLP*论文集，H. Bouamor、J. Pino 和 K. Bali 编，2023 年，第 5592–5605 页。[32] S. Freisinger、F. Schneider、A. Herygers、M. Georges、T. Bocklet 和 K. Riedhammer，“视频讲座的无监督多语言主题分段：层次标签能告诉我们哪些性能信息？”，载于*SLaTE*论文集，2023 年，第 141–145 页。[33] E. J. Hu、Y. Shen、P. Wallis、Z. Allen-Zhu、Y. Li、S. Wang、L. Wang 和 W. Chen，“LoRA：大语言模型的低秩适配”，载于*ICLR*，2022 年。[34] T. Dettmers、A. Pagnoni、A. Holtzman 和 L. Zettlemoyer，“QLoRA：量化大语言模型的高效微调”，载于*NIPS*，2023 年。[35] A. Radford、J. W. Kim、T. Xu、G. Brockman、C. McLeavey 和 I. Sutskever，“通过大规模弱监督实现稳健的语音识别”，2022 年。[36] M. McAuliffe、M. Socolof、S. Mihuc、M. Wagner 和 M. Sonderegger，“Montreal Forced Aligner：使用 Kaldi 进行可训练的文本-语音对齐”，载于*Interspeech*，2017 年，第 498–502 页。[37] C. Fournier，“使用边界编辑距离评估文本分段”，载于*ACL*论文集，H. Schuetze、P. Fung 和 M. Poesio 编，2013 年，第 1702–1712 页。[38] N. Reimers 和 I. Gurevych，“利用知识蒸馏将单语句子嵌入转换为多语言句子嵌入”，载于*EMNLP*论文集，2020 年。[39] W. Wang、F. Wei、L. Dong、H. Bao、N. Yang 和 M. Zhou，“MiniLM：面向任务无关压缩的深度自注意力蒸馏预训练Transformer”，载于*NIPS*，2020 年。[40] J. Su、M. Ahmed、Y. Lu、S. Pan、W. Bo 和 Y. Liu，“RoFormer：采用旋转位置嵌入的增强型Transformer”，载于*Neurocomput.*，第 568 卷，C 期，2024 年。