DiscoSum：意识到语篇的新闻摘要

Alexander Spangher[*1], **Tenghao Huang**[*1], **Jialiang Gu**[*2], **Jiatong Shi**[3], **and Muhao Chen**[4]

[1]University of Southern California Information Sciences Institute
[2]School of Computer Science, Wuhan University
[3]Independent Contributor
[4]University of California, Davis
spangher@usc.edu

Abstract

近年来，文本摘要的进展主要依赖于大型语言模型来生成简明摘要。然而，语言模型往往无法维持长期的篇章结构，特别是在新闻文章中，组织流程显著影响读者的参与度。我们引入了一种将篇章结构整合到摘要过程中去的新方法，专注于各种媒体中的新闻文章。我们呈现了一个新的摘要数据集，其中新闻文章在不同的社交媒体平台（例如 LinkedIn、Facebook 等）上以不同方式被多次摘要。我们开发了一种新的新闻篇章模式来描述摘要结构，并提出了一种新算法 DiscoSum，其采用束搜索技术以结构感知的方式进行摘要，使得新闻故事可以根据不同的风格和结构要求进行转化。人类和自动评估结果均表明，我们的方法在维持叙述忠实度和满足结构需求方面的有效性。

1 介绍

近年来，文本摘要取得了显著进展，这得益于基础大型语言模型，它们能够生成简明而具有丰富上下文的长文档概览 (Li and Chaturvedi, 2024; Peper et al., 2024; Zhang et al., 2024)。然而，尽管取得了这些进步，目前的摘要方法很少考虑文本组织的一个基本方面：话语结构 (Cohan et al., 2018)。

现代新闻机构，如《纽约时报》，越来越多地通过多种媒体（如印刷报纸、移动应用程序、播客和社交媒体）发布新闻摘要，每种媒体都有不同的受众期望和内容格式 (Kalsnes and Larsson, 2018; Ngoc, 2022)。例如，像《纽约时报》这样的机构可能会制作适合儿童的播客版本，使用简化语言和柔和框架；在 Instagram 上发布简明、视觉上引人入胜的片段；以及在 LinkedIn 或报纸本身的网站上发布更长、更详细的文章，以迎合专业或学术读者。将一则新闻转化为多种风格和长度，同时保持其核心叙述和重点，需要对话语结构进行细致的控制 (Shen et al., 2017; Hu et al., 2017)。

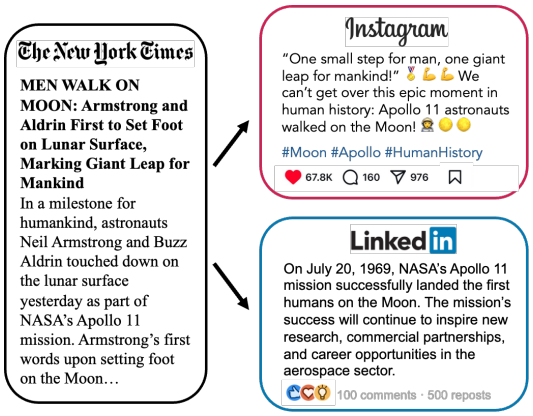

Figure 1:《纽约时报》在多个平台上对阿波罗 11 号登月新闻的比较呈现。这个例子展示了内容格式和语言适应的多样性，以迎合不同的观众：一篇详细的传统印刷文章、一篇简洁且视觉导向的 Instagram 帖子，以及一个专业取向的 LinkedIn 摘要。每个平台都反映出特定的编辑策略，以有效地吸引其独特的受众，强调了意识到语境的新闻摘要重要性。

尽管对自动新闻摘要的兴趣日益增加 (See et al., 2017; Zhang et al., 2020; Beltagy et al., 2020; He et al., 2020; Zhao et al., 2022b)，现有的数据集方法已经忽略了这一需求[1]。为了弥合这些差距，我们提出了一项新的关注话语结构的摘要任务，该任务强调在表面级别的摘要连贯性或事实正确性之外的结构性话语建模。

首先，我们介绍了 DiscoSum：一个话语感知新闻摘要数据集。DiscoSum 是最大和最多样化的专业撰写的跨平台新闻摘要集合，包含来自 10 个国家的 23 个不同新闻媒体的 20,000 篇新闻文章，并与来自 Facebook、Instagram、Twitter 和新闻简报这 4 个不同平台的超过 100,000 篇人工撰写的摘要多次配对。接下来，我们开发了一种新颖的话语模式，用于描述新闻摘要的结构组成部分，包含五个句子级话语标签。最后，我们还提出了一种新颖的话语驱动解码方法，该方法利用束搜索技术来评估和选择摘要

[*]Equal contribution.

[1]关于更深入的比较，请参见附录 C 和 Grusky et al. (2020)。

中包含的最佳后续句子。我们通过开发表面层和结构指标来评估我们的模型在生成结构感知摘要方面的有效性。我们的人工和自动评估证实了我们的方法有效地保持了叙述的忠实性并遵循结构要求。总之，我们做出了以下贡献：

1. 新任务：我们将结构感知摘要引入新闻领域。

2. 新数据集：我们引入了一个大规模的语料库，其中包含 2 万篇新闻文章，与超过 10 万篇不同的人撰写的总结在 Facebook、Twitter、Instagram 和新闻简报中形成一对多的配对。我们引入了一种新的用于结构化摘要的论述模式。

3. 基准测试结果：我们展示了基准模型和评估，证明了 NLP 系统在改进结构感知新闻摘要方面的可行性和潜力。

2 相关工作

News Summarization. 新闻摘要一直是自然语言处理研究的重点 (Barzilay and McKeown, 2005; Hong et al., 2014; Paulus et al., 2017; Goyal et al., 2022) 。传统的方法通常依赖于抽取技术，比如选择“导言”句子来近似新闻的“引言” (Fabbri et al., 2019; Wang et al., 2020) ，但最近神经生成模型的进步使得摘要更具连贯性和语境丰富 (Li and Chaturvedi, 2024; Peper et al., 2024; Zhang et al., 2024) 。大型数据集，包括专门为新闻文章整理的那些，进一步推动了模型性能的发展，通过提供多样化和具有代表性的训练样本 (Grusky et al., 2020; Chen et al., 2016) 。然而，许多这些方法没有明确地对新闻文章的内在结构进行建模，导致摘要虽然流畅，但可能省略了关键的结构组件 (Grenander et al., 2019; Zhao et al., 2022b) 。

Controllable Generation and Test-Time Alignment. 可控生成已经成为一种有前景的方法，以确保输出满足某些风格、语气或长度要求 (Yang et al., 2019; Yang and Klein, 2021; Zhao et al., 2022a) 。可控生成研究的一个显著领域是测试时对齐，即模型在推理时结合约束或偏好，以更好地符合用户或任务特定的准则。提示工程或解码时间门控等技术在引导模型输出向所需属性靠拢方面显示了前景 (Meng et al., 2022; Huang et al., 2023; Liu et al., 2024) 。然而，这些方法往往侧重于表面层次的约束——如字数或风格——可能未能考虑新闻文章的深层话语结构特征。

Discourse-Aware Language Modeling. 越来越多的研究强调话语结构的重要性，例如识别文档的转折点、来源和总结陈述，以改进文本生成任务 (Zhai et al., 2003; Tian et al., 2024; Spangher et al., 2024c, 2025) 。具备话语意识的方法利用话语元素 (Spangher et al., 2022a) 或新闻指南 (Spangher et al., 2022b) 等理论来解析和利用文本生成过程中的结构组件。虽然一些努力在特定领域任务中纳入修辞角色或话语解析 (Wang and Cardie, 2013; Wang and Ling, 2016) ，但它们在新闻文章中的应用仍处于初期阶段。这些方法通过与新闻文本的自然组织对齐，显示出生成既能告知又能吸引人的摘要的前景，缩小了事实一致性与面向受众的设计之间的差距。

我们的工作与之前在其他领域中尝试的结构化摘要有所不同。STRONG 框架 (Zhong and Litman, 2023) 利用结构来解析法律文件，以确定在摘要中包括哪些元素，而不是控制生成结构本身。同样，关于对话摘要的工作 (Chen and Yang, 2023) 使用结构控制如实体元组和对话行为分布，但侧重于局部连贯性而非全局结构。与我们的方法最相似的是关于元评论摘要的研究 (Shen et al., 2022) ，尽管它依赖于手工制作和手动标记的、一对一文章到摘要映射的文章。我们的工作引入了无需手工标记的结构化摘要方法，并允许一对多的映射。

在本节中，我们描述了结构化摘要（§2.1 ）的任务表述和评估指标。我们介绍了我们提出的数据集，包括其组成和标注过程（**??** ）。

2.1 任务表述

设 D 表示原始新闻文档，它可以由多个段落或句子组成。我们定义期望的论述标签序列为 $\boldsymbol{T} = (t_1, t_2, \ldots, t_n)$ ，其中每个 t_i 代表一个论述标签（例如，“上下文细节”或“介绍性元素”等），i -th 句子的摘要应满足这些标签。目标是生成一个摘要 $\boldsymbol{S} = (s_1, s_2, \ldots, s_m)$ ，其中每个 s_i 都是与 D 相关且连贯的句子。Note: 在结构化摘要任务中，我们假设用户预先提供目标标签序列 $\boldsymbol{T}$ 。对于新输入预测最佳结构的工作将留待将来进行。

我们使用一个分类函数 $C(\cdot)$ ，该函数在给定一个句子的情况下预测其话语标签。设 $\boldsymbol{L} = (l_1, l_2, \ldots, l_m)$ 为由 $C(s_i)$ 对 $\boldsymbol{S}$ 中的每个句子 s_i 预测出的标签序列。我们要求 $\boldsymbol{L}$ 按顺序与 $\boldsymbol{T}$ 对齐，以便对每个位置 i 满足 $l_i = t_i$ 。尽管最直接的情形设置为 $m = n$ ，使得总结正好包含 n 句子，但更灵活的变体可能允许稍微的偏差，同时仍然保证核心位置与目标标签匹配。

我们寻求构建一个大型多样的数据集，该数据集包括新闻文章及其在不同社交媒体平台和新闻简报上由记者撰写的多个不同摘要。我们从 10 个不同的国家（美国、中国、印度、英

Category	Count
# of Outlets	23
# of News Articles	20,811
# of Facebook Posts	18,275
# of Instagram Posts	66,030
# of Twitter Posts	8,977
# of Newsletters	10,506

Table 1: 不同类别的总体计数。

Types	Counts
Overall	45,195
News Article Tweet	12,516
News Article Facebook Post	15,645
News Article Instagram Post	7,738
News Article Newsletter Post	9,296

Table 2: 新闻文章到摘要图的统计数据，显示帖子类型之间的边数。

国、德国等）收集了 23 个主要的国家和国际新闻媒体，以捕捉不同写作风格中的多种不同话语风格。

社交媒体收集我们从 23 家新闻机构中收集了两年的 Twitter、Facebook 和 Instagram 上的社交媒体帖子。为此，我们构建了半自动滚动代理，可以滚动每个新闻机构媒体页面的动态消息。我们收集每个帖子的完整 HTML，包括每个帖子的文本以及任何链接的 URL。总共，我们收集了 8,977 个 Twitter 帖子，18,275 个 Facebook 帖子，和 66,030 个 Instagram 帖子（更多详情见 Tab. 2 ）。为了识别结构性摘要，我们进一步过滤这些帖子，保留那些包含 50 个或更多字符的帖子。这消除了我们数据的大约 30 %。

邮件简报收集我们选择了由新闻媒体发布的 7 个邮件简报品牌，[2] 专门搜索那些在其品牌下的所有过往邮件简报在线上存档可用的。我们构建了爬虫来收集每份邮件简报的完整 HTML，并收集了 2 年的数据，总共超过 20,000 份邮件简报（详见表格 Tab. 1 ）。

通常一个简报会同时总结许多新闻文章，而我们的任务是单篇文档的摘要任务。因此，我们需要解析每份简报的文本，以便简报文本块对应单独的新闻文章。这涉及具有重叠段的文本分割，因为在简报中可能需要较大的文本片段来包含链接。为此，我们提示了 LLMs [3] ，并基于先前的研究工作展示了 LLM 在文本分割任务中的有效性 (Nayak; Zhao et al., 2024; Fan et al., 2024; Jiang et al., 2023) 。我们选择了一种提示配置，指导一个 LLM 去：（1）识别所有新闻内容链接，（2）提取每个链接的周围文本上下文，（3）排除模板化内容，以及（4）保持原始文本的准确性。为减轻潜在的偏差或幻觉，我们实施了一个验证程序，最大提取块会在多个迭代中与 LLM 自己的输出进行交叉检查，任何不一致都会被标记以供人工审查。人工检查证实了 LLM 在这项任务中的能力，其分割质量在我们随机抽样的一组 100 个简报的审计中超出 95 % 的准确度。总共，我们从收集的简报中生成了 10,506 个摘要。

新闻文章集合我们从以上描述的所有社交媒体帖子和新闻简报中收集新闻文章 URL 的超集。根据 Spangher et al. (2024a) 的方法，我们从 Wayback Machine 中抓取每篇新闻文章的 HTML。我们使用一个大型语言模型（GPT-4）来清理 HTML，以提取完整的新闻文章（我们发现现有的库 [4] 是不够的）。我们的提示策略指导模型过滤掉非新闻段落（例如登录提示、广告和冗余内容），而仅保留文章内容。

新闻文章和摘要匹配对于许多社交媒体帖子，我们在帖子中有一个 URL 可以让我们明确匹配；然而，对于其他帖子则不行（例如，Instagram 不允许在帖子中使用 URL）。为了发现尽可能多的连接，我们决定将任何渠道的新闻文章与任何社交媒体帖子或新闻简报摘要进行匹配。为此，我们采用两步的排序和检查方法。具体来说，我们首先使用 SBERT (Reimers, 2019) 来嵌入新闻文章和摘要；对于每一篇新闻文章，我们找出十个最接近的摘要作为候选。然后，我们使用 GPT-4 对每个候选项进行严格的成对比较，仅返回关于它们是否描述同一新闻故事的二元“是”或“否”的判断，遵循在 Spangher et al. (2024b) [5] 中验证的方法。在手动审计中，这一步匹配的准确率超过 95 %，证明了我们多步程序的稳健性。这种方法不仅帮助我们恢复了每篇文章由单一新闻渠道所产生的所有摘要，还可以看到其他新闻渠道如何报道同一新闻事件。

数据集划分在所有实验中，我们使用 DiscoSum 数据集的 70%/20%/10% 的训练/验证/测试（14k/4k/2k 文章-摘要对）划分。此划分是在文章层面进行的，以防止信息泄漏，因此同一篇文章的所有摘要都保持在同一划分中。

在本节中，我们概述了生成结构感知摘要的方法。首先，我们描述两个必要的组成部分：（1）我们用来进行结构化摘要的语篇框架，以及（2）一句话级别的标注器，它预测语篇标签，我们用其指导生成（§2.2 - §2.3 ）。然

[2] Axios ‘‘终点线”; 纽约时报 ‘‘早报”; 洛杉矶时报 ‘‘今日加州”; The Skimm ‘‘每日概要”; The Daily Beast ‘‘小抄”; Semafor ‘‘新闻简报”; CNN ‘‘可靠来源”

[3] 附录 D.4 中显示的提示。

[4] https://newspaper4k.readthedocs.io/zh/latest/

[5] 作者发现，大型语言模型可以用于高性能地验证跨文档事件共指。

后，我们提出两种算法来生成符合目标语篇序列 $\boldsymbol{T}$ 的摘要：（1）基于编辑的方法（§2.4.1）和（2）束搜索方法（§2.4.2）。

2.2 语篇图式生成

为了形式化“结构化”摘要的概念，我们寻求构建一个低维的、创新的话语模式来描述社交媒体和新闻通讯的摘要。首先，我们使用自动化过程生成一个模式，这与之前的工作通过手动分析来开发模式的方法形成对比，通常基于 O(10) 个例子[6]。受到 Pham et al. (2024) 的启发，我们首先请求一个 LLM 为我们所有摘要中的每个句子的语篇角色生成描述性标签（O(100k) 个句子）。然后，我们使用 SBERT 嵌入模型 (Reimers, 2019) 对这些标签进行嵌入，并使用 k-means 对这些嵌入进行聚类。

通过这个嵌入过程，我们识别出代表不同叙述角色的五个不同簇：Introductory Elements、Contextual Details、Event Narration、Source Attribution 和 Engagement Directive。每个话语角色的定义请参见 Tab. 4。我们通过邀请两位专业记者评估质量并构思缺失的角色标签，确认了该框架的有效性。选择具体的五个话语标签是经过大量实验得出的。在我们的聚类方法中，虽然其他参数选择（例如，k=7、13 或 23）也是可行的，我们基于人类评估试验选择了一个五维框架，该试验显示了高水平的标注者间一致性（$\kappa = 0.615$）来验证这些标签的有效性。虽然五维框架在捕捉新闻话语结构的全部复杂性上可能显得有限，特别是在跨文化或小众新闻场景中，但它为本次话语感知摘要的试点研究提供了坚实的基础。

2.3 话语标注器

接下来，为了指导我们的结构感知生成（部分 2.4.2），我们根据 Spangher et al. (2021, 2022a) 构建了一个句子级别的分类器，用于给句子分配话语标签。分类器是在 DiscoSum 的训练集上训练的。为了验证验证集的质量，我们请两位专家注释员独立标记了一个包含 500 个句子的子集。训练的标注器在验证集上实现了超过 90 % 的高准确率，如 Fig. 3 所示。如此高水平的准确性对于摘要过程中的作用至关重要，因为它后续将被用作奖励指导机制，确保生成的摘要符合所需的话语结构。附录中的完整混淆矩阵展示了标注器在所有五个话语类别上的强劲表现，其中每类的 F1 得分最低也超过 0.85。

[6]例如，Van Dijk (1988) 建立了基于对 12 篇新闻文章的分析的架构。

Algorithm 1 句子级别的基于语篇的波束搜索，波束大小为 k

Require: Source text X , target label sequence $t_1, \ldots, t_N$, beam width k
Ensure: Best summary $S = \langle s_1, s_2, \ldots, s_N \rangle$

Initialize beam: $\mathcal{B} \leftarrow \{[]\}$ ▷ Start with an empty sequence
for $i \leftarrow 1$ to N **do**
  $\mathcal{B}' \leftarrow \emptyset$
  for $s \in \mathcal{B}$ **do**
    $\mathcal{C} \leftarrow \text{LLM}(s, X, k)$ ▷ Generate k candidate
    for $c \in \mathcal{C}$ **do**
      $s' \leftarrow \text{append}(s, c)$
      score $\leftarrow C(s', t_i)$
      $\mathcal{B}' \leftarrow \mathcal{B}' \cup \{(s', \text{score})\}$
    end for
  end for
  $\mathcal{B} \leftarrow \text{selectTopK}(\mathcal{B}', k)$
end for
$S \leftarrow \text{argmax}_{(s,\text{score}) \in \mathcal{B}}$ score
return S

2.4 生成方法

2.4.1 迭代编辑

我们的第一个策略将摘要生成视为一个迭代细化的过程。我们首先提示大型语言模型生成一个完整的初始摘要，然后反复“编辑”任何不符合其预期话语标签的句子。在生成初始摘要后，我们使用我们的标注器 $C(\cdot)$ 来识别哪些句子的标签是错误的。然后，我们移除这些“标签不匹配”的句子，并生成新的候选句子。经过几次迭代，摘要逐渐“演变”以匹配序列 $\boldsymbol{T}$。

通过仅关注个别问题句子，这种方法保留了摘要中已经正确的部分。它还可以适应复杂的标签序列，而无需在每次发现不匹配时重新开始整个生成过程。

2.4.2 句子级别束搜索

与逐步修正错误相反，我们的第二种策略采用波束搜索样式，从头开始逐句构建符合标签的摘要句子 (Lowerre, 1976)。

我们从一个空的摘要开始，一次考虑一个位置（例如，首先是应该具有“介绍性元素”标签的句子，然后是应该具有“背景细节”标签的句子，依此类推）。在每个步骤 i 中，LLM 生成多个候选句子（形成句子级的“束”），然后由 $C(\cdot)$ 进行评估。我们选择与目标标签 t_i 最匹配的候选句子。这个句子被添加到当前的部分摘要中。通过在每个步骤评估多个选项并选

择与期望标签最匹配的句子，这种方法确保每个摘要句子都遵循预定的标签序列。详细的过程如 Alg. 1 所述。

3 实验

在本节中，我们展示了实验设置（§3.1 ）和用于目标话语标签的结构化摘要的评估框架（§3.2 ）。我们介绍了基准测试的基线模型和方法（§3.3 ）。接下来，我们展示实证结果（§3.4 ）、人工偏好评估（§3.5 ）以及对不同束大小影响的分析（§3.6 ）。

3.1 实现细节

对于普通生成，我们基于自动话语标签器从 16 次试验中采样最佳输出。在句子级束搜索中，我们采用了 BeamSize = 16 。我们使用 PEFT 方法在 DiscoSum 的训练集上微调了 LLaMa-3-8B 模型。这种微调方法在 20 个 epoch 后显著降低了验证损失。关键超参数包括一个学习率为 5e-05 的多 GPU 分布式训练设置，使用八个 Nvidia 4090。在我们的实验中，每次生成我们都会随机生成一个结构标签列表，以模拟最广泛的用户输入集。这也防止了我们在常见观测到的话语结构上过拟合。

3.2 评估协议

为了量化生成新闻摘要的内容准确性，我们使用了几个指标：

- ROUGE-L。(Lin, 2004) ROUGE-L，最初设计用于摘要，测量生成的摘要与参考摘要之间的最长公共子序列的词元。
- FactCC。(Kryscinski et al., 2020) FactCC 是一种基于模型的度量，旨在分类每个生成句子是否在事实上一致于源文档。
- AlignScore。AlignScore 是一种一致性指标，用于直接衡量源文本与摘要之间的事实对应关系。

为了评估生成的摘要 $\boldsymbol{S}$ 与预期话语结构 $\boldsymbol{T}$ 之间的一致性，我们从 $\boldsymbol{S}$ 中推导出一个预测标签序列 $\boldsymbol{L}$ ，形式为

$$\boldsymbol{L} = \text{Labeler}(s_1, s_2, \ldots, s_n), s_i \in S$$

，其中 Labeler 表示设计用于识别话语结构的人类标注者或自动化模型。

我们采用三种指标来量化 $\boldsymbol{L}$ 与目标标签序列 $\boldsymbol{T}$ 的接近程度，后者表示摘要中句子的理想结构角色：

- 最长公共子序列（LCS）。LCS 衡量的是 $\boldsymbol{L}$ 和 $\boldsymbol{T}$ 共有的最长子序列的长度。更高的 LCS 值表示预测标签更紧密地保持了预期的标签顺序。
- 匹配得分。匹配得分评估了 $\boldsymbol{L}$ 和 $\boldsymbol{T}$ 之间逐位准确匹配的数量。该指标反映了在序列中每个标签预测到正确位置的精确度。
- Levenshtein 距离。(Levenshtein, 1965) 这个指标计算将 $\boldsymbol{L}$ 转换为 $\boldsymbol{T}$ 所需的最小单元素编辑次数（插入、删除或替换）。较低的 Levenshtein 距离表示更高的序列相似度。

鉴于人工评估的潜在高成本，我们提供了自动化和人工评估的协议：

我们与两名人工标注者合作，手动评估生成摘要中每个句子的语篇结构。在这项研究中，我们要求标注者对每个模型的 100 个摘要进行评估。

3.3 基线

为了评估我们所提出方法的有效性，我们将其与一系列在架构、训练范式和优化目标上有所不同的基线模型进行对比。这些模型包括专有系统和开源替代方案，提供了当前文本摘要及相关任务的最新能力的全面概述。

这些模型，例如 DeepSeek-V3 [7] 、Claude-3-5-sonnet [8] 和 GPT-4o [9] ，主要被纳入是为了帮助我们评估我们的方法与尖端技术相比的表现，即使这些模型不是我们评估的主要焦点。

Open-Source LLMs. 模型如 Qwen-2.5 和各种配置的 LLaMa-3-8B 代表了在学术研究中广泛使用的更易获得的选项。每个版本的 LLaMa-3-8B——无论是普通版本、基于编辑的修改还是经过微调的迭代——都用于展示在开源框架中不同的潜在改进和权衡。

3.4 主要结果

Content Accuracy Evaluation. 表 3 显示了多种模型的表面级和结构评估。尽管在不同系统中 ROUGE-L、FactCC 和 AlignScore 存在波动，我们的方法——特别是 LLaMa-3-8B 的束搜索变体——在表面级指标中保持了竞争力的性能。值得注意的是，我们的束搜索方法实现了最高的 AlignScore（0.3890），相比于专有和其他开源模型，展示了与源文档的卓越的事实一致性。这特别重要，因为它表明可以在不牺牲事实上与源内容对齐的情况下实现结构改进，甚至可以增强这种对齐。我们还包括了以推理为中心的模型 O1，它在几个指标上优于

[7] https://api-docs.deepseek.com/news/news1226
[8] https://www.anthropic.com/claude/sonnet
[9] https://openai.com/index/hello-gpt-4o/

Models	Content Accuracy			Auto Struct.			Human Struct.		
	R-L (%) ↑	FactCC ↑	AlignScore ↑	MS ↑	Lev ↓	LCS ↑	MS ↑	Lev ↓	LCS ↑
Proprietary Models									
DeepSeek-V3	47.15	0.47	0.3886	0.26	0.64	0.65	0.24	0.65	0.65
Claude	34.30	0.70	0.3882	0.25	0.68	0.64	0.20	0.49	0.75
GPT-4o	29.51	0.63	0.3884	0.11	0.80	0.62	0.15	0.58	0.68
O1	44.65	0.50	-	0.28	0.66	0.54	-	-	-
Open-sourced Models									
Qwen-2.5	40.82	0.58	0.3888	0.24	0.66	0.65	0.15	0.52	0.64
LLaMa-3-8B	47.18	0.50	0.3496	0.21	0.77	0.36	0.24	0.49	0.65
– Finetuned	22.01	0.61	0.3495	0.14	0.77	0.45	0.18	0.55	0.72
– Edit-based	15.28	0.59	-	0.51	0.48	0.56	0.24	0.65	0.36
– Beam Search	42.98	0.64	0.3890	0.72	0.32	0.68	0.55	0.17	0.87

Table 3: 模型在各种指标上的比较。指标分为内容准确性和结构评估，两者都有自动和人工标注。指标包括 ROUGE-L（% ）、FactCC、AlignScore（用于事实一致性）、匹配分数（MS）、Levenshtein 距离（Lev）和最长公共子序列（LCS）。↑表示数值越高越好，↓表示数值越低越好。加粗数字突出最佳表现，而带下划线的数字表示各类别中的显著但次要表现。

GPT-4o，但在我们的 LLaMa-3-8B 束搜索变体之后仍然落后。

值得注意的是，我们的方法在自动和人工结构评估中都表现出色，在这两个方面相较于开源基线和更复杂的专有模型都有显著提升。LLaMa-3-8B 的束搜索变体 consistently 更加接近指定的话语标签序列，其表现为更高的匹配得分和更小的 Levenshtein 距离。这种在结构对齐上的增强突出了模型在不显著损失表层准确性的情况下严格遵循特定修辞结构的能力。通过在文本重叠和结构忠实性之间实现有效平衡，我们的方法显著提升了生成文本的可控性和连贯性。

Performances of Edit-based and Finetuned Methods. 基于编辑的方法在增强生成摘要的结构与期望的语篇标签对齐方面展示了有希望的能力，这在其在结构评估中的强表现中得到了证明。然而，这种结构的忠实性是以内容的准确性和流畅性为代价的，其 ROUGE-L 分数明显低于其他方法。这一下降表明，尽管基于编辑的方法有效地塑造了摘要的结构，但可能会显著偏离原文的语义和句法性质。

另一方面，经过微调的 LLaMa-3-8B 模型变体在任务适应性上表现不够令人满意。尽管微调能够将模型行为严格定制为特定的数据集或任务要求，但观察到的性能指标表明，该模型未能捕捉到针对这一具体的基于话语的摘要任务所需的更深层次和结构性细微差别。低得分意味着仅仅进行微调可能不足以应对需要深入理解和根据复杂标记方案转换文本的任务。这种低性能强调了我们任务需要更高级的方法。

3.5 人类对摘要质量的评估

我们招募了两名标注者，根据内容准确性和结构遵从性对三种摘要生成方法——Vanilla LLaMA-3-8B、其微调版本以及我们的束搜索方法——进行排名。我们的结果如图 **??** 所示，表明束搜索方法显著优于其他方法，其平均倒数排名 (MRR) 达到 0.71，而 Vanilla 和微调方法分别为 0.55 和 0.58。

3.6 波束大小的影响

我们的分析纳入了从 2 到 16 的各种束大小。随着束大小的增加，我们观察到 LCS 分数的整体提升，表明与目标话语结构的对齐增强。相反，Levenshtein 距离（用于测量将预测序列与目标对齐所需的编辑距离）在束大小增加时呈现出总体下降趋势，表明更大的束大小改善了结构对齐。

观测到的趋势为未来的研究开辟了几条途径。一个潜在的领域是探索可以根据文本的复杂性或文档不同部分的特定话语结构需求动态调整的自适应束大小。此外，尽管束搜索技术在推理过程中提高了摘要的质量和相关性，将这些高质量的摘要整合到训练中可能会提升模型的整体表现。未来的研究可以探讨如何利用这些精炼的输出来促进训练过程。

在这项研究中，我们介绍了一种结构性摘要方法，将话语组织整合到新闻文章的摘要中，强调叙事的真实性和结构的对齐。我们的新数据集 DiscoSum 和评估指标突出了我们方法的有效性，特别是束搜索技术，确保摘要在上下文相关性和结构精确性上都达到要求。结果显示比传统方法有显著改进，表明我们的方法可以增强跨多种媒体平台的自动新闻摘要。

我们的贡献突出了新闻摘要领域中尚未探索

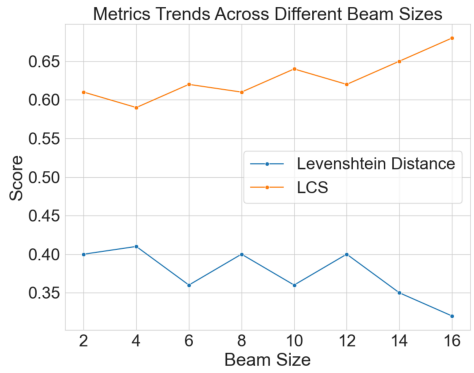

Figure 2: 在结构化摘要中，编辑距离和最长公共子序列（LCS）得分作为束大小的函数。图表显示了编辑距离的一般降低和 LCS 得分的逐步提高，表明随着束大小的增大，结构对齐有所改善。

的研究问题。我们的 DiscoSum 数据集和相应的评估指标为进一步探索如何系统地将话语元素纳入摘要模型奠定了基础。这种向深入理解话语结构的转变不仅挑战现有模型，还为新闻叙述重构的更复杂方法开辟了路径。通过强调结构而非表层连贯性，我们邀请研究界探索可能改变如何在不同媒体环境中总结新闻内容的新方法论 (Caswell and Dörr, 2018; Spangher et al., 2022b; Caswell, 2024; Welsh et al., 2024) 。

4

限制 **Focus of the Study.** 尽管我们使用标准指标（例如，FactCC，ROUGE-L）来衡量内容准确性并承认其重要性，但我们的主要目标是确保与话语标签的结构对齐，而不是优化事实正确性。因此，事实精确度或内容覆盖率的提升是偶然的而非有意的。未来的工作可以研究结合更强大事实核查和检索增强生成技术的方法，以补充结构上的忠实性，尤其是在事实准确性至关重要的应用中。

Trade-offs in Decoding Efficiency. 虽然我们的束搜索方法显著改善了结构的遵从性，但与更简单的生成技术相比，它在计算上可能更加昂贵。这一开销可能对实时应用或大规模部署构成挑战。未来的研究可以探讨自适应束策略或混合方法，这些方法在解码速度与需要严格话语控制之间取得平衡。

Potential Data Biases. 我们的数据收集方法涉及 LLMs 用于多个关键任务，包括 HTML 清理、时事通讯分段和文章-摘要匹配。尽管我们采取了广泛的措施来验证这些过程，但这些模型可能引入影响数据集构成和结果方案的偏见。为减轻这种担忧，我们收集了一个跨越来自 10 个不同国家的 23 个主要新闻媒体，并涵盖 4 种不同分发方式的多样化数据集，这有助于平衡不同写作风格和媒体偏好的潜在偏见。

此外，虽然我们的论述框架有意设置为粗粒度以增强普适性，我们承认结构中仍可能存在偏差。尽管我们的主要关注点是结构而非词汇方面，之前研究中识别的实体或性别偏见 (Spangher et al., 2024a) 可能会渗透到结构模式中。我们的数据集的规模和多样性有助于降低这些问题的影响，但未来的研究应探讨词汇偏见与论述结构之间的关系，特别是对于需要跨文化或特定领域适应的应用。

References

Regina Barzilay and Kathleen R McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328.

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

David Caswell. 2024. Telling every story: Characteristics of systematic reporting. In *Journalism and Reporting Synergistic Effects of Climate Change*, pages 266–283. Routledge.

David Caswell and Konstantin Dörr. 2018. Automated journalism 2.0: Event-driven narratives: From simple descriptions to real stories. *Journalism practice*, 12(4):477–496.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Association for Computational Linguistics (ACL)*.

Jiaao Chen and Diyi Yang. 2023. Controllable conversation generation with conversation structures via diffusion models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7238–7251.

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

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

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2024. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16998–17010.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. 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 6019–6024, Hong Kong, China. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2020. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *Preprint*, arXiv:1804.11283.

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

Kai Hong, John M Conroy, Benoit Favre, Alex Kulesza, Hui Lin, Ani Nenkova, et al. 2014. A repository of state of the art and competitive baseline summaries for generic news summarization. In *LREC*, pages 1608–1616.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Tenghao Huang, Ehsan Qasemi, Bangzheng Li, He Wang, Faeze Brahman, Muhao Chen, and Snigdha Chaturvedi. 2023. Affective and dynamic beam search for story generation. *ArXiv*, abs/2310.15079.

Feng Jiang, Weihao Liu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Haizhou Li. 2023. Advancing topic segmentation and outline generation in chinese texts: The paragraph-level topic representation, corpus, and benchmark. *arXiv preprint arXiv:2305.14790*.

Bente Kalsnes and Anders Olof Larsson. 2018. Understanding news sharing across social media: Detailing distribution on facebook and twitter. *Journalism studies*, 19(11):1669–1688.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Haoyuan Li and Snigdha Chaturvedi. 2024. Rationale-based opinion summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8274–8292, Mexico City, Mexico. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Qin Liu, Fei Wang, Nan Xu, Tianyi Yan, Tao Meng, and Muhao Chen. 2024. Monotonic paraphrasing improves generalization of language model prompting. *ArXiv*, abs/2403.16038.

Bruce T. Lowerre. 1976. The harpy speech recognition system.

Tao Meng, Sidi Lu, Nanyun Peng, and Kai-Wei Chang. 2022. Controllable text generation with neurally-decomposed oracle. *ArXiv*, abs/2205.14219.

Kota Shamanth Ramanath Nayak. Does chatgpt measure up to discourse unit segmentation? a comparative analysis utilizing zero-shot custom prompts.

Nguyen Minh Ngoc. 2022. Journalism and social media: The transformation of journalism in the age of social media and online news. *European Journal of Social Sciences Studies*, 7(6).

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *Preprint*, arXiv:1705.04304.

Jospeh J. Peper, Wenzhao Qiu, and Lu Wang. 2024. Pelms: Pre-training for effective low-shot multi-document summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. Topicgpt: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *Preprint*, arXiv:1704.04368.

Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. Mred: A meta-review dataset for structure-controllable text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2521–2535.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Preprint*, arXiv:1705.09655.

Alexander Spangher, Tenghao Huang, Philippe Laban, and Nanyun Peng. 2025. Creative planning with language models: Practice, evaluation and applications. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 1–9, Albuquerque, New Mexico. Association for Computational Linguistics.

Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 498–517.

Alexander Spangher, Yao Ming, Xinyu Hua, and Nanyun Peng. 2022a. Sequentially controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6848–6866, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexander Spangher, Nanyun Peng, Sebastian Gehrmann, and Mark Dredze. 2024a. Do llms plan like human writers? comparing journalist coverage of press releases with llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21814–21828.

Alexander Spangher, Xiang Ren, Jonathan May, and Nanyun Peng. 2022b. NewsEdits: A news article revision dataset and a novel document-level reasoning challenge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 127–157, Seattle, United States. Association for Computational Linguistics.

Alexander Spangher, Serdar Tumgoren, Ben Welsh, Nanyun Peng, Emilio Ferrara, and Jonathan May. 2024b. Tracking the newsworthiness of public documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14150–14168, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Spangher, James Youn, Matt DeButts, Nanyun Peng, Emilio Ferrara, and Jonathan May. 2024c. Explaining mixtures of sources in news articles. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15837–15859, Miami, Florida, USA. Association for Computational Linguistics.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Teun A Van Dijk. 1988. *News as discourse*. Routledge.

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

Lu Wang and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1395–1405, Sofia, Bulgaria. Association for Computational Linguistics.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, California. Association for Computational Linguistics.

Ben Welsh, Naitian Zhou, Arda Kaz, Michael Vu, and Alexander Spangher. 2024. Newshomepages: Homepage layouts capture information prioritization decisions. *Preprint*, arXiv:2501.00004.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Shuai Yang, Zhangyang Wang, Zhaowen Wang, Ning Xu, Jiaying Liu, and Zongming Guo. 2019. Controllable artistic text style transfer via shape-matching gan. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4442–4451.

ChengXiang Zhai, William W. Cohen, and John D. Lafferty. 2003. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*.

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.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Chao Zhao, Faeze Brahman, Tenghao Huang, and Snigdha Chaturvedi. 2022a. Revisiting generative commonsense reasoning: A pre-ordering approach. *ArXiv*, abs/2205.13183.

Chao Zhao, Tenghao Huang, Somnath Basu Roy Chowdhury, Muthu Kumar Chandrasekaran, Kathleen McKeown, and Snigdha Chaturvedi. 2022b. Read top news first: A document reordering approach for multi-document news summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 613–621, Dublin, Ireland. Association for Computational Linguistics.

Jihao Zhao, Zhiyuan Ji, Yuchen Feng, Pengnian Qi, Simin Niu, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Meta-chunking: Learning efficient text segmentation via logical perception. *arXiv preprint arXiv:2410.12788*.

Yang Zhong and Diane Litman. 2023. Strong–structure controllable legal opinion summary generation. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 431–448.

Part I
Appendix

Table of Contents

A 话语图式定义

如前所述，我们的话语框架旨在捕捉跨不同平台和格式新闻内容的结构组织。该框架被开发用于捕获那些承担叙事角色的常见话语元素——从建立背景和引入关键事件到提供背景细节和吸引读者。通过模拟这些话语角色，我们使摘要系统能够保持新闻内容中使其连贯和吸引人的基本结构组件，而不是仅仅关注表层的语言特征。有关框架元素及其定义的完整列表，请参见表 4 。

我们模式中的话语标签反映了句子在新闻内容的更广泛叙述结构中所扮演的高层次功能角色。这些包括例如 Event Narration （描述主要新闻事件）、Contextual Details （提供必要的背景信息）、Introductory Elements （设定故事背景）、以及 Engagement Directives （旨在吸引读者注意的元素）。每个标签代表一种独特的沟通目的，促进新闻叙述整体的连贯性和有效性，使我们的系统不仅能够理解应该包含哪些信息，还能理解不同信息在叙述结构中的作用。

B 语篇标注器的混淆矩阵

在本节中，我们进一步详细介绍了我们的话语标注器的准确性。具体来说，我们呈现了经过训练的话语标注器与真实标注标签的混淆矩阵。整体准确率为 90.90 %，F-1 得分为 0.9087。结果显示了我们训练的话语标注器的稳健性。

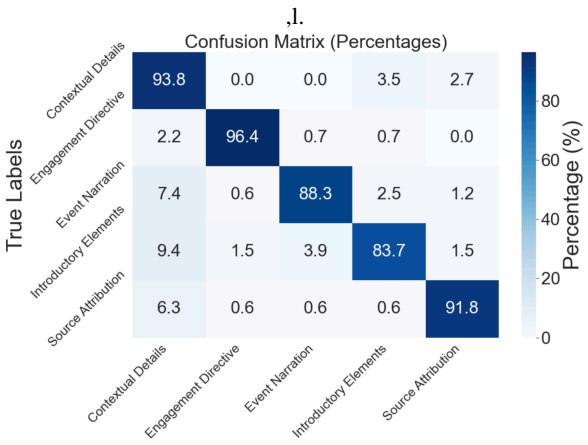

Figure 3: 话语标签的混淆矩阵。

C 与 Newsroom 的比较

NEWSROOM 数据集 (Grusky et al., 2020) 是一个广泛用于摘要研究的资源，与我们的工作类似，包含新闻文章及其摘要。然而，我们的 DiscoSum 数据集在收集方法、内容丰富性和研究重点上有着根本的不同。

C.1 收集方法

NEWSROOM 的收集机制从 HTML 元数据中提取摘要，具体来说是从嵌入文章 URL 的 `<meta property="description">...</meta>` 标签中提取。这种方法可以高效地收集大量摘要，但每篇文章只限于一个摘要，该摘要针对搜索引擎或链接预览上下文而设计。

相比之下，DiscoSum 收集的是记者为特定社交媒体平台和新闻通讯编写的实际帖子。现代新闻编辑室通常会雇用专门的社交媒体团队来制作特定平台的内容，因此，不同渠道对同一篇文章产生了多个不同的摘要。这些摘要很少会在文章的 HTML 元数据中显示，因为它们是直接写入每个平台的发布界面的。

为了说明这一差异，我们展示了一个案例研究，内容是《纽约时报》关于独特形状的洋基棒球棒的文章：

如表 5 所示，元描述（由 NEWSROOM 收集）简洁明了，专注于分析师的职业变动。相比之下，社交媒体帖子（由 DiscoSum 收集）提供了关于故事核心要素的更丰富信息——创新的球棒设计、背后的物理原理以及来自创作者的引述——在各个平台上的细节程度各不相同。

C.2 研究价值

DiscoSum 为摘要研究提供了几个优势：

1. 每篇文章有多个参考摘要：DiscoSum 为每篇文章提供多篇专业撰写的摘要，涵盖不同平台和格式。

2. 平台特定的结构模式：该数据集展示了相同内容是如何为不同平台（推特、脸书、Instagram、新闻通讯）而改编的，揭示了平台特定的结构模式。

3. 真实世界的受众定位：DiscoSum 中的摘要代表用户实际看到的内容，这些内容由专业记者撰写，考虑了特定的受众和平台。

4. 话语结构分析：通过用话语标签标注这些不同的摘要，DiscoSum 使得研究叙述结构如何在跨平台中适应成为可能。

尽管 NEWSROOM 在通用摘要研究中具有重要价值，但 DiscoSum 专门支持研究可适应不同平台和结构要求的语篇感知摘要策略——随着新闻消费在各种数字渠道中分化，这一能力变得越来越重要。

Label	Definition
Introductory Elements	Sets the stage for the summary by introducing the main topic, themes, or key points that will be covered.
Contextual Details	Provides additional background and setting information to help understand the main events or topics being summarized.
Engagement Directive	Directs the reader's attention or actions through calls to action, questions, or direct addresses to engage them with the content.
Event Narration	Describes specific events or occurrences in a narrative form, detailing what happened in a sequential or explanatory manner.
Source Attribution	Cites the origins of the information, giving credit to sources or clarifying the basis of the claims made in the summary.

Table 4: 用于结构化摘要的语篇标签定义

Source	Content
Article URL	https://www.nytimes.com/athletic/6241862/2025/03/30/yankees-bats-aaron-leanhardt-marlins/
Meta Description (NEWSROOM)	"Aaron Leanhardt was the Yankees lead analyst in 2024 before joining the Marlins coaching staff this offseason."
Facebook Post (DiscoSum)	"The New York Yankees' uniquely shaped bats have caught the attention of many and are the result of two years of research and experimentation."
Twitter Post (DiscoSum)	"From @TheAthletic: The New York Yankees' uniquely shaped bats have caught the attention of many and are the result of two years of research and experimentation. Meet the former MIT physicist behind the 'torpedo' bats."
Instagram Post (DiscoSum)	"The New York Yankees' uniquely shaped bat is the result of two years of research and experimentation with a former MIT physicist-turned-coach at the helm. Aaron Leanhardt, the brains behind the 'torpedo bats' making headlines, says the idea behind his innovation was simple —redistribute the weight of the bat to where it matters. The bats have been around for more than just this season. Players used them in 2024. But after last weekend's laser show in the Bronx, they have broken into the mainstream. 'Ultimately, it just takes people asking the right questions and being willing to be forward-thinking,' Leanhardt says."

Table 5: 对比 NEWSROOM 与 DiscoSum 收集的内容，针对同一篇纽约时报文章。请注意不同社交媒体平台之间的格式差异。

D 提示

D.1 叶节点提示

该提示用于通过总结一组带注释句子的共同模式来分析和分类新闻文章中的话语角色。它生成简洁的标签，以捕捉一组句子的基本话语功能。

```
You are a helpful assistant. I will give
    you a large set of notes about
    sentences in news articles that I
    wrote down.

Here are the notes:

{labels}

Please summarize them, focusing on the
    common discourse role each sentence
    plays, based on the notes. Ignore
    the topic.
Summarize them with a single, specific
    label for the entire group, being
    sure to concisely capture what they
    are about.
Make the label 2-3 words, max. Be
    descriptive but not too broad.
    Please return just one label and one
     description.
Make it in this format: ‘‘‘"Label":
    Description‘‘‘
```

D.2 中间树提示

这个提示用于对写作元素进行层级分类。它有助于创建中级标签，将相似的语篇角色分组在一起，关注共同的功能方面，同时忽略特定的主题。

```
You are a helpful assistant. I will give
    you a notes about different writing
    elements.

Here are the notes:
```

```
{labels}

Please summarize them with one label,
    focusing on the common discourse
    role each element plays, ignoring
    the topic.
Summarize them with a single, specific
    label for the entire group,
    concisely capturing what they are
    about.
Make the label 2-3 words, max. Be
    descriptive but not too broad.
    Please return just one label and one
     description.
Make it in this format: ‘‘‘"Label":
    Description‘‘‘
```

D.3 少样本示例选择提示

此提示用于识别每个论述标签的代表性示例。它选择多样化的、高质量的示例来最好地说明特定标签，这些示例可以稍后用于小样本学习或注释指南。

```
I am trying to find good examples to use
     for demonstrating a label.
Here is the label: {label}. The
    definition for the label is: {
    definition}.

Here are a large set of examples I have,
     alone with notes for each one:
[Start Examples]
{examples}
[End Examples]

Some examples are bad. Please choose 4
    examples that best represent this
    label. Try to pick diverse ones.
Return the examples and the notes, and
    copy them fully.
Return as a json. Be careful to format
    the quotes correctly.
```

这个提示为社交媒体帖子中的句子分配预定义的语篇角色标签。它利用完整帖子和特定注释中的上下文信息，从一个受控词汇中为句子匹配最合适的语篇标签。

```
I will give you a social media post and
    a single sentence from that post.
Your goal is to assign a label to that
    sentence with a general discourse
    role that best describes it's
    purpose in the overall script.
Each sentence also includes some notes I
     took about the very specific
    discourse role it plays, you can use
     them if it's helpful.

Choose from this list:
{discourse_labels}

Do NOT return any labels NOT in that
    list. Here are some shortened
    examples:
‘‘‘{examples}‘‘‘

Now it's your turn. Here is a social
    media post:
‘‘‘{full_document}‘‘‘

What discourse role is this sentence in
    it serving?
Sentence: ‘‘‘{sentence}‘‘‘
Notes: ‘‘‘{notes}‘‘‘
Answer:
```

D.4 通讯处理提示

此提示从新闻简报中提取并组织新闻内容。它识别并分离与特定链接相关的文本块，重点关注有意义的新闻内容，同时过滤掉常见的样板文字。

```
Look at the clean HTML of this
    newsletter.

Please separate the blocks of text into
    news content corresponding to each
    individual link.
This includes all the context
    surrounding the links.
Exclude links that do not pertain to
    news content.
The same text can be included in
    different chunks if it is relevant
    to a link.
Try to include all text in at least one
    chunk.
If a line doesn't end with a period,
    please add one.
Do not change the text otherwise, in any
     way.
Ignore text that is boilerplate and not
    related to news content.

Return a python dictionary mapping the
    link to each chunk of text. Don't
    return anything else. Copy the text
    exactly.

    ‘‘‘{html}‘‘‘
```

在本节中，我们展示带有语篇标签的帖子示例。

Label	sentence
Introductory Elements	Boston’s streets are changing.
Contextual Details	A growing number of them have bike lanes meant to protect bicyclists, slow down drivers, reduce the risk of crashes, and ultimately get more people to feel comfortable biking
Introductory Elements	The city is aiming to expand the bike lane network so that half of residents live within a 3-minute walk of a safe and connected bike route by the end of next year.
Contextual Details	The theory is that if there is a safe path for biking, more people will take it, in turn reducing climate change-causing emissions, traffic deaths, and mind-numbing congestion.
Engagement Directive	But challenges remain.
Introductory Elements	Many projects face vocal opposition to ceding valuable street real estate to bikes.
Introductory Elements	And other issues, such as the prevalence of large trucks, and lingering gaps in the bike network, make biking more dangerous than most would like.

Table 6: 带有句子级别标签的 Instagram 帖子示例

Label	sentence
Introductory Elements	Global upheaval has once again increased America’s geopolitical importance.
Event Narration	This years election campaign will shape the direction of U.S. policy.
Contextual Details	It is thus being closely watched around the world.

Table 7: 具有句子级标签的 Facebook 帖子示例

Label	sentence
Introductory Elements	Logan Edra, a 21-year-old American B-Girl, said the Olympics could provide young girls with a vision of the future.
Source Attribution	"Any type of representation is going to help people see what is possible."

Table 8: 一个带有句子级别标签的推文示例

Label	sentence
Event Narration	Disney began laying off thousands of staff members, its second round of layoffs, to save $ 5.5 billion in costs and cut 7,000 jobs.
Contextual Details	Employees at ESPN, Disney Entertainment, Disney Parks, and Experiences and Products will also be affected.
Engagement Directive	A third round of layoffs is expected before summer.
Introductory Elements	Meanwhile, Insider’s employees went on strike after about 10 % of its staff was laid off.
Contextual Details	Staffing cuts have also affected Buzzfeed, NPR, and other news organizations.

Table 9: 一个带有句子级标签的通讯示例

Original News Article	
There's overwhelming support for an age limit on the president and Congress, but it won't happen anytime soon. Polling by Insider and Morning Consult indicates that three in four Americans favor an age limit for members of Congress. More than four in 10 viewed the ages of political leaders as a "major" problem. Some lawmakers are concerned that politics has become a gerontocracy, and the country's oldest politicians have not just the incentive to stick around but the power to remain entrenched. The seniority system, a campaign-finance system that favors incumbents, gerrymandered congressional districts, and the ability to continue to seek office indefinitely are mechanisms that members of Congress are most eager to discuss. Age limits are not on the table, and Congress' youngest and oldest lawmakers seem to agree. The country's power class is also aging across the board; the current Congress is the oldest in the nation's history. At the start of the latest Congress, in January 2021, baby boomers controlled roughly 56 % of the seats in both chambers despite representing approximately just 21 % of the US population. Some lawmakers reflect on an underlying system that prioritizes tenure and seniority over other potential measures of political acumen, generating a political leadership that skews older. Term limits offer another potential solution, though that too would require going through the arduous constitutional-amendment process."	
Target Labels: Engagement Directive, Contextual Details, Introductory Elements, Introductory Elements, Event Narration	
Engagement Directive	Many are calling for a constitutional amendment to set a mandatory retirement age for lawmakers, citing the need for fresh perspectives and a more diverse range of experiences in Washington.
Contextual Details	Additionally, the proposed age limit aims to address concerns that the current age distribution in Congress may hinder its ability to effectively represent the diverse needs and values of the younger generation.
Introductory Elements	Despite the widespread support, opponents argue that imposing an age limit would violate the constitutional principle of equal protection, as it would unfairly target a specific group based on age.
Introductory Elements	Despite the ongoing debate, polling suggests that three in four Americans support an age limit for members of Congress, with over four in 10 viewing the ages of political leaders as a major problem.
Event Narration	A recent poll has found that three in four Americans favor an age limit for members of Congress, with over four in 10 viewing the ages of political leaders as a "major" problem.

Table 10: 示例生成 # 1

Original News Article
Are robot waiters the future? Some restaurants think so. MADISON HEIGHTS, Mich. (AP) ŏ2014 You may have already seen them in restaurants: waist-high machines that can greet guests, lead them to their tables, deliver food and drinks and ferry dirty dishes to the kitchen. Some have cat-like faces and even purr when you scratch their heads. But are robot waiters the future? Itŏ2019s a question the restaurant industry is increasingly trying to answer. Many think robot waiters are the solution to the industryŏ2019s labor shortages. Sales of them have been growing rapidly in recent years, with tens of thousands now gliding through dining rooms worldwide. Dennis Reynolds, dean of the Hilton College of Global Hospitality Leadership at the University of Houston, says, There's no doubt in my mind that this is where the world is going. The school's restaurant began using a robot in December, and Reynolds says it has eased the workload for human staff and made service more efficient. However, others say robot waiters aren't much more than a gimmick that have a long way to go before they can replace humans. They can't take orders, and many restaurants have steps, outdoor patios and other physical challenges they can't adapt to. Redwood City, California-based Bear Robotics introduced its Servi robot in 2021 and expects to have 10,000 deployed by the end of this year in 44 U.S. states and overseas. Shenzen, China-based Pudu Robotics, which was founded in 2016, has deployed more than 56,000 robots worldwide. Li Zhai, owner of Noodle Topia in Madison Heights, Michigan, says the robots give human servers more time to mingle with customers, which increases tips. And customers often post videos of the robots on social media that entice others to visit. But not all chains have had success with robots. Chili's introduced a robot server named Rita in 2020 and expanded the test to 61 U.S. restaurants before abruptly halting it last August. The chain found that Rita moved too slowly and got in the way of human servers. Eventually, experts expect that a certain percentage of restaurants maybe 30 % will continue to have human servers and be considered more luxurious, while the rest will lean more heavily on robots in the kitchen and in dining rooms.
Target Labels: Event Narration, Contextual Details, Event Narration, Engagement Directive, Event Narration
Event Narration — Robot waiters are here to stay, and their presence is redefining the future of the restaurant industry.
Contextual Details — The innovative technology, first introduced at a Michigan restaurant, has been gaining popularity worldwide, as they bring efficiency, convenience, and a touch of novelty to the dining experience.
Event Narration — Robot waiters are increasingly being adopted by restaurants worldwide, offering benefits such as reduced labor costs, improved accuracy, and enhanced customer engagement.
Engagement Directive — As the technology continues to evolve, we can expect to see robot waiters becoming more sophisticated, with advanced navigation systems and the ability to learn and adapt to individual customers' preferences.
Event Narration — With the potential to transform the dining experience, robot waiters are being hailed as a game-changer in the industry, enabling restaurants to adapt to changing consumer preferences and improve overall efficiency, as well as providing a unique and entertaining experience for customers.

Table 11: 例子生成 # 2

Original News Article
D̈on't fool yourself: Billions more needed to protect tropical forests, warns new report. At least $ 130bn a year is needed to protect the most at-risk areas of tropical forest by the end of the decade, alongside reductions in beef and dairy consumption and government bans on deforestation, a thinktank has warned. Currently, finance to protect forests averages between $ 2bn and $ 3bn a year. The report estimates that eliminating the economic incentive to destroy forests for cattle ranching, agriculture and other uses would cost at least $ 130bn a year. The money could come from carbon markets, wealthy governments and philanthropists, but there must also be urgent actions such as a ban on clearing forests, developing businesses that rely on standing forests and reducing demand for commodities linked to deforestation, such as palm oil, soya, beef and cocoa. Lord Turner, a former head of the CBI and ex-chair of the UK government's Committee on Climate Change, warned that governments should not delude themselves about the scale of the challenge, and that robust quantification of what you spend and what you get is much more difficult than anywhere else. Land use change is the second largest source of human greenhouse gas emissions, with deforestation accounting for about 15 % of the total. It is also a major driver of biodiversity loss and ecosystems degradation, and has continued at a relentless pace despite scientific warnings that ecosystems such as the Amazon could soon collapse.
Target Labels: Event Narration, Event Narration, Event Narration, Contextual Details, Event Narration
Event Narration: The new report emphasizes that protecting tropical forests is crucial for mitigating climate change, preserving biodiversity, and supporting local communities, but it requires a significant increase in funding to achieve these goals.
Event Narration: A significant increase in funding of at least $ 130 billion annually by the end of the decade is necessary, alongside reductions in beef and dairy consumption and government bans on deforestation, to effectively protect the most at-risk areas of tropical forest.
Event Narration: This funding increase is essential to protect the most at-risk areas of tropical forest, which are currently under threat due to economic incentives driving deforestation.
Contextual Details: Currently, finance to protect forests averages between $ 2bn and $ 3bn a year, which is a tiny fraction of the estimated $ 130bn needed annually.
Event Narration: Accordingly, the report warns that current funding levels are insufficient, and drastic measures are needed to address the issue.

Table 12: 例子生成 # 3

Sentence	Superset Discourse Label
Tina McMahon-Foley is now celebrating her 30th year as a naturalist for Cape Ann Whale Watch (@capeannwhalewatch).',	"Main Event": This sentence directly describes a primary event, noting a milestone in Tina McMahon-Foleyś career as a naturalist, which is the focal point of the document.',
But the story of how she found her way to Cape Ann begins in her former home in Albany, when she worked as a young science teacher in the early 1990s.',	"Previous Event": This sentence describes a specific event that occurred before Tina McMahon-Foley became a naturalist for Cape Ann Whale Watch, providing background information on how she found her way to Cape Ann.',
She was watching television one night when she flipped to the Discovery Channel.',	"Previous Event": This sentence describes a specific event that occurred before Tina McMahon-Foley became a naturalist for Cape Ann Whale Watch, providing context and background information on how she discovered her interest in whales.',
"What came next – a whale documentary, a drive to Massachusetts, a scientist's admiration of her gumption – stick with her today.",	"Consequence": This sentence describes a series of events that directly succeeded a previous event (watching the Discovery Channel) and had a lasting impact on the subjectś life, shaping her current situation as a naturalist.'
On a recent trip, she appeared just as excited to see a whale as she was the first day that scientist, Roger Payne, sent her to sea.',	"Anecdotal Event": This sentence describes a specific, personal experience of Tina McMahon-Foley that illustrates her enduring passion for whale watching, adding an emotional and relatable aspect to her story.',
As a calf breached near the ship, she spoke into the mic to those on board: "Have you caught your breath yet?'	"Anecdotal Event": This sentence describes a specific, personal moment in Tina McMahon-Foleyś experience as a naturalist, which is used to illustrate her enthusiasm and passion for her work, rather than to advance the main narrative of her 30-year career.',

Table 13: 用于构建我们的篇章模式的篇章标签超集的一个例子。