arXiv:2304.03548v3 [cs.CL] 9 Dec 2023

GEMINI：控制抽象式文本摘要中的句子级摘要风格

Guangsheng Bao[1,2],**Zebin Ou**[2],**和 Yue Zhang**[*,2,3,1] 浙江大学[2] 工程学院，西湖大学[3] 先进技术研究院，西湖高等研究院[2]
{baoguangsheng, ouzebin, zhangyue}@westlake.edu.cn

摘要

人类专家使用不同的方法撰写摘要，包括从文档中提取句子并进行改写，或融合文档中的各种信息对其进行摘要。这些方法具有灵活性，因此难以由任何单一方法加以模仿。为了解决这一问题，我们提出了一种自适应模型 GEMINI，该模型整合了改写器和生成器，分别模仿句子改写和摘要生成的方法。GEMINI 能够自适应地选择改写文档中的特定句子，或从头开始生成摘要句子。实验表明，在三个基准数据集上，我们的自适应方法优于纯抽象式和改写式基线方法，并在 WikiHow 上取得了最佳结果。有趣的是，实证结果表明，给定摘要句子的上下文，其人类摘要风格始终是可预测的。我们在 https://github.com/baoguangsheng/gemini 发布了我们的代码和模型。

1 引言

文本摘要旨在为给定的文本文档自动生成流畅且简洁的摘要（Maybury，1999；Nenkova和McKeown，2012；Allahyari等，2017年）。两种主流方- 法，即抽取式和生成式摘要技术，已被采用。抽取式方法（Nallapati 等，2017年；Narayan等，2018；Liu 和Lapata，2019；Zhou 等，2020；Zhong 等，2020）从输入中识别出显著的文本片段，并将其组合成输出摘要，从而生成忠实但可能存在冗余且不连贯的摘-要（Chen 和 Bansal，2018；Gehrmann 等，2018；Cheng 和 Lapata，2016）；在此过程中，可以使用重写技术进一步减少冗- 余并增强连贯性（Bae 等，2019；Bao 和 Zhang，2021年）。相比之下，生成式方法（Rush 等，2015；Nallapati 等，2016；

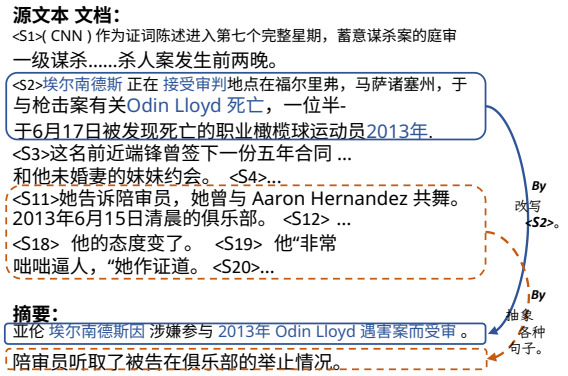


图 1：来自 CNN/DM 的人工摘要。

See 等人，2017年；Lewis 等人，2020）采用自然语言生成（NLG）技术合成输出，从而获得更加简洁连贯的摘要，但代价是忠实度下降（Huang 等人，2020；Maynez 等人，2020）。

抽取式和抽象式方法的有效性取决于摘要的摘要风格。研究表明，人们在撰写每个摘要句时会采用不同的风格（Jing 和 McKeown，1999），我们将其大致分为抽取式风格和抽象式风格。抽取式风格主要直接传达文章句子中的观点，而抽象式风格则传达从文章中的多个句子推导出的新观点。图1展示了一个例子，其中摘要由两个句子组成。第一个句子是通过改写文档中的句子 $<S_2>$ 生成的。相比之下，第二个摘要句是通过概括多个句子生成的。这些风格灵活多变，因此难以通过单一方法加以模仿。

在本文中，我们旨在模仿人工摘要的风格，我们相信这能够提升我们控制摘要风格的能力，并加深我们对人工摘要生成方式的理解。为了更好地适应人工摘要中的摘要风格，我们提出了一种自适应模型 **GEMINI**，该模型

** 通讯作者。

包含一个改写器和一个生成器，分别用于模仿抽取式风格和抽象式风格。对于 改写器，我们采用结合了内部句子抽取器的上下文改写方法（Bao 和 Zhang，2021）。对于生成器，我们使用标准的 seq2seq 摘要器（Lewis 等人，2020）。

改写器和生成器集成在单个解码器中，使用风格控制器 切换生成风格。风格控制器会分配不同的组标签，使输入中的相关句子能够用于引导解码器。在抽象式风格下，解码器不会聚焦于特定句子；而在抽取式风格下，解码器会通过更多关注某个特定的改写后句子来获得引导。为了训练这种自适应摘要器，我们使用句子级摘要风格的自动检测来生成预言机抽取式/抽象式风格。

我们在三个具有代表性的基准数据集上评估了我们的模型。结果表明，GEMINI通过区分摘要风格，使模型能够更好地拟合训练数据。我们的自适应改写器-生成器网络在这些基准数据集上以显著优势超越了强大的抽象式基线和近期的改写器模型。有趣的是，实验还表明，在测试集阶段可以始终如一地预测句子的摘要风格，这说明在上下文中，人类对摘要句摘要风格的选择具有内在的一致性。据我们所知，我们是首个明确控制每个摘要句风格的研究，同时还取得了更高的 ROUGE 分数。我们的自动风格检测指标为未来对摘要风格进行进一步的定量分析提供了可能。

2 相关工作

抽象式摘要器通过在预训练的 seq2seq 模型上进行微调，取得了具有竞争力的结果（Lewis 等人，2020；Zhang 等人，2020）。我们的工作与此一脉相承，贡献在于对摘要句进行新颖的风格控制，使模型能够更好地拟合“句子改写”和“长距离抽象”风格的人工摘要。我们的生成器是一个标准的 seq2seq 模型，架构与 BART 相同（Lewis 等人，2020），而我们的改写器则与以往的单句改写工作（Chen 和 Bansal，2018；Bae 等人，2019；Xiao 等人，2020）以及语境化改写工作（Bao 和 Zhang，2021）相关。

我们的改写器采用上下文改写机制，该机制考虑每个改写句子的文档上下文，从而可以召回上下文中的重要信息并保持跨句连贯性。然而，不同于 Bao 和 Zhang(2021)，其依赖外部提取器来选择句子，我们使用指针机制（Vinyals 等人，2015），类似于 NeuSum（Zhou 等人，2020），以自回归方式选择句子。据我们所知，我们是首个将改写器和生成器整合到独立抽象式摘要器中的团队。

摘要风格	CNN/DM	XSum	WikiHow
抽取式	88.6%	11.8%	61.1%
抽象式生成	11.4%	88.2%	38.9%

表 1：人工评估句子级摘要风格在三个数据集上的表现。百分比表示属于该风格的摘要句子所占的比例。

我们的模型 GEMINI 也可以视为一种专家混合模型（Jacobs 等人，1991），它会在改写器和生成器之间动态切换。相关工作包括用于序列到序列建模的指针生成网络（See 等人，2017年），也可以将其视为一种专家混合模型。另一项相关工作是 HYDRASUM，它在解码时使用两个专家。这些模型可以视为一种软专家混合模型，通过学习潜在专家并整合其输出作出决策。相比之下，我们的模型可以视为一种硬专家混合模型，在作出决策时会选择改写器或生成器。两种模型的一个显著区别在于，我们的 GEMINI 针对每个句子作出决策，而以往的工作是在词元层面作出决策。这些模型的目标截然不同。

3 句子级摘要风格

3.1 摘要风格的人工评估

我们对每个摘要句的摘要风格进行人工评估：如果摘要句可以由文章中的某个句子推知，则标注为抽取式风格；如果需要结合文章中的多个句子才能推知该摘要句，则标注为抽象式风格。我们从每个数据集中抽取 100 个摘要句，请三位标注者进行标注，并将至少获得两票的风格作为最终标签。

风格的平均分布如表1所示。CNN/DM主要采用抽取式风格，其中 88.6% 的摘要句子以抽取式风格撰写。相比之下，XSum 主要是

数据集	新颖 n 元语法			片段		Ours
	1 元语法	2 元语法	3 元语法	覆盖率	密度	融合指数
CNN/DM	0.62	0.55	0.56	0.59	0.56	**0.76**
XSum	0.25	0.17	0.24	0.15	0.16	**0.46**
WikiHow	0.47	0.49	0.46	0.31	0.30	**0.56**

表2：我们的自动指标与以往指标在衡量融合程度方面的比较。数字表示该指标与人工标注的融合程度之间的皮尔逊相关系数。

抽象式风格，其中88.2%的摘要句- 子以抽象式风格撰写。WikiHow的摘要风格分布更加均衡，约60%的摘要句子采用抽取式风格。结果表明，真实摘要是多种风格的混合体，即使对于广为人知的抽取式数据集CNN/DM和抽象式数据集XSum也是如此。

3.2 摘要风格的自动检测

由于人工标注的成本高昂，我们转而采用自动方法，在句子级检测摘要风格。以往的研究主要在摘要级衡量摘要风格。例如，Grusky 等人(2018) 提出使用抽取片段的覆盖率和密度来衡量抽取性。See 等人(2017年) 提出将新颖 n 元语法（即出现在摘要中但未出现在输入文档中的 n 元语法）的比例作为抽象性的指标。我们将这些指标应用于句子，并将其作为我们方法的基线。

我们提出了融合指数，用于衡量上下文信息融合的程度（Barzilay 和 McKe- own, 2005），并考虑两个因素：1）摘要句中的信息有多少可以从一个文档句子中召回？如果召回率较高，该句子更可能属于抽取式风格，因为只需改写文档句子即可生成该句子。2）摘要句与多少个文档句子相关？相关文档句子的数量越多，表明融合程度越高，摘要句也越可能属于抽象式风格。我们的融合指数根据这两个因素计算。

召回率。我们通过将摘要句匹配回文档句子，来衡量可召回信息的百分比。给定摘要句 S 和源文档 $D = \{S_1, S_2, ..., S_{|D|}\}$，我们找到最佳匹配

$$RC(S \mid D) = \max_{1 \le i \le |D|} R(S \mid S_i), \tag{1}$$

其中，$R(S|S_i)$ 是给定 S_i 时 S 的 ROUGE-1/2/L 召回率的平均值，表示句子 S 的信息被句子 S_i 覆盖的百分比。

散布度。我们通过将摘要句与所有文档句子进行匹配，来衡量摘要句内容的散布度。如果匹配分数在所有文档句子之间均匀分布，则散布度较高；如果除一个文档句子外，其他文档句子的匹配分数均为零，则散布度较低。我们使用根据匹配分数得出的分布熵来计算散布度。

$$SC(S \mid D) = -\sum_{j=1}^{K} p_j \log p_j / \log K, \tag{2}$$

其中，p_j 是根据前K 个最佳匹配项$\{r_j\}|_{j=1}^{K}$ 计算出的概率估计值，用于判断 S 是否由相应的文档句子 S_i生成，在线a

$$\begin{aligned} p_j &= r_j / \sum_{j=1}^{K} r_j, \\ \{r_j\}|_{j=1}^{K} &= \mathrm{Top}(\{R(S|S_i) \mid 1 \le i \le |D|\}, K). \end{aligned} \tag{3}$$

超参数 K 通过在人工评估集上进行经验搜索确定。

融合指数。我们根据召回率 *(RC)*和散 (FI) 布度 *(SC)*计算融合指数。

$$FI(S \mid D) = (1 - RC(S \mid D)) * SC(S \mid D), \tag{4}$$

其表示融合程度，其中 0 表示没有融合，1 表示极度融合，其他值介于两者之间。

我们评估了我们的指标与先前研究中的候选指标，并报告其与人工标注的摘要风格之间的皮尔逊相关系数，如表2所示。我们提出的融合指数与摘要风格的相关性最高。相比之下，在以往的指标中，只有新颖 1-gram 的相关性最接近，但平均仍比融合指数低约 0.14。结果表明，融合指数是更适合句子级的抽取式—抽象式度量。

3.3 OracleLabelforSummary Style

我们使用自动融合指数生成预言机抽取式/抽象式标签，以便训练一个

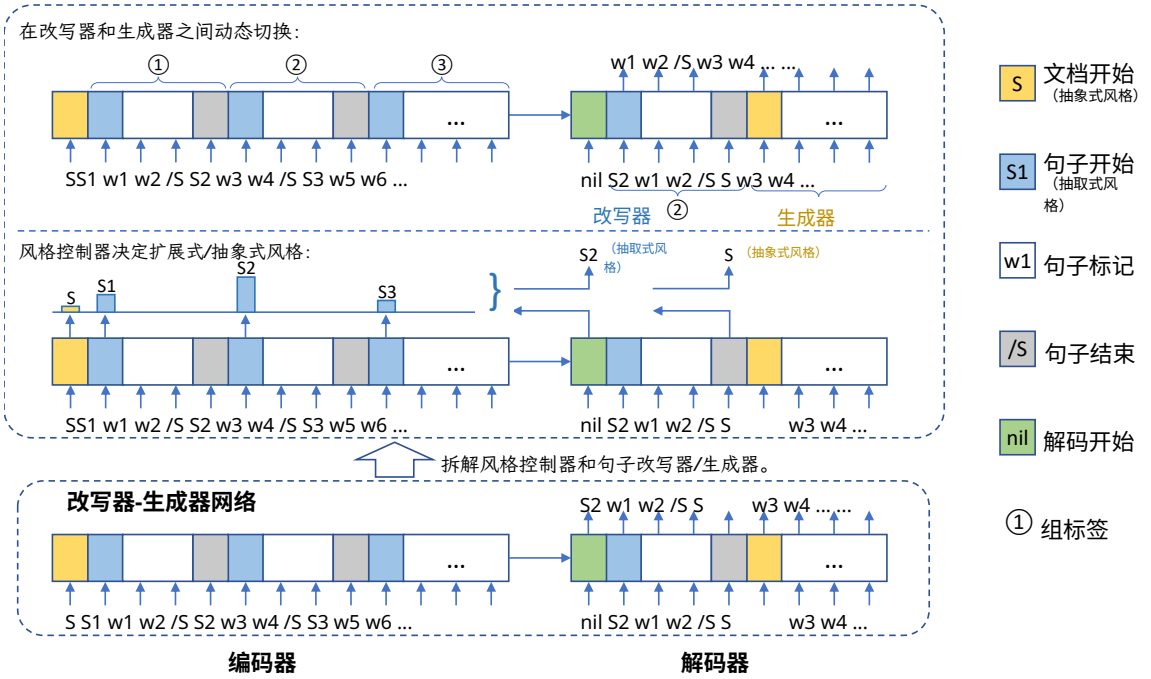


图2：GEMINI 使用控制器来决定 ext/抽象式风格，并进一步相应地在改写器和生成器之间切换解码器。我们将标识符令牌“<S>”和“$<S_k>$”以简洁形式“S”和“S_k”表示，并使用“w1 w2”表示第一个句子中的令牌，使用“w3 w4”表示第二个句子中的令牌。组标签被转换为嵌入，并添加到编码器和解码器的输入令牌嵌入中。解码器预测一个标识符令牌，以确定后续时间步的组标签，例如使用“<S2>”开始组标签“2”，直到句子“</S>”结束。

能够在句子层面进行显式风格控制的摘要器。如果融合指数高于某个阈值，我们就将该句子视为抽取式风格。如果摘要句的融合指数低于该阈值，我们就将该句子视为抽象式风格。我们通过开发实验为每个数据集寻找最佳阈值。

4 GEMINI：改写器-生成器网络

如图2 所示，GEMINI 采用预训练的 BART（Lewis 等人，2020），并使用风格控制器来确定摘要风格。根据风格，激活改写器或生成器来生成每个句子。

直观而言，当 GEMINI 处理风格平衡的数据集时效果最佳，因为这能让改写器和生成器相辅相成；同时，预言机风格的准确率也要足够高，才能确保训练监督具有较高质量。

4.1 输入与输出

我们引入特殊的标识符令牌，包括“<S>”——文档的开头、“$<S_k>$”——句子 k 的开头，以及“</S>”——句子的结尾。

我们将输入文档表示为“<S> $<S_1>$ 文档句子一. </S> $<S_2>$ 句子二. </S> $<S_3>$句子三. </S> ...”，其中序列以标识符令牌“<S>”开头，并使用“$<S_k>$”和“</S>”包围每个句子。我们将输出摘要表示为“$<S_2>$摘要句一 . </S><S> 摘要句二. </S> ...”，其中每个句子以“$<S_k>$”或“<S>”开头，并以“</S>”结尾。如果摘要句以“$<S_k>$”开头，解码器将根据第k个文档句子生成该句子。如果摘要句以“<S>”开头，解码器将根据整个文档生成该句子。

4.2 文档编码器

我们扩展嵌入表，将标识符标记纳入其中，以便风格控制器能够匹配这些标记的嵌入，从而确定摘要风格。我们遵循上下文改写方法（Bao 和 Zhang，2021），为每个输入句子分配一个组标签，以便改写器能够借助多头注意力依靠这些组标签定位改写后的句子，正如 ① ② ③ 图2所示。

具体来说，第一个摘要句和第二个文档句子具有相同的组

数据集	源文本	训练集	验证集	Test	文档 词数	句子数	摘要 词数	句子数	#W/S
CNN/DM	News	287,227	13,368	11,490	788	36.9	56	3.8	15
XSum	News	204,045	11,332	11,334	431	18.7	23	1.0	23
WikiHow	知识库	168,128	6,000	6,000	584	32.2	62	7.5	8

表3：评估数据集。词数、句子数和每句词数（平均每句词数）均为平均值。

标签 ②，它们会被转换为组标签嵌入，并添加到句子中的词元嵌入中。通过在编码器和解码器之间共享组标签嵌入表，可以训练解码器在生成第一个摘要句时关注第二个文档句子。

形式化地，我们将组标签嵌入表表示为 EMB^{tag}，并根据每个句子的索引从 X 中唯一生成组标签序列G_X，如下所示：

$$G_X = \{k \text{ if } w_i \in \text{S}_k \text{ else } 0\}|_{i=1}^{|X|}, \tag{5}$$

对于 k句中的词元 w_i S_k，我们为其分配编号为 k的组标签。我们将 G_X转换为嵌入，并将其注入 BART 编码器。

我们采用标准的 Transformer 编码层，其中包含一个自注意力模块和一个前馈模块：

$$\begin{aligned} x^{(l)} &= \text{LN}(x^{(l-1)} + \textsc{SelfAttn}(x^{(l-1)})), \\ x^{(l)} &= \text{LN}(x^{(l)} + \textsc{FeedForward}(x^{(l)})), \end{aligned} \tag{6}$$

其中，LN 表示层归一化（Ba 等人，2016）。最后一层 L 输出最终的编码器输出 $x_{out} = x^{(L)}$。向量 x_{emb} 和 x_{out} 被传递给解码器进行预测。

4.3 摘要解码器

我们为 BART 解码器扩展了一个风格控制器和一个改写器，而对于生成器，我们使用默认的解码器。

风格控制器。我们使用注意力指针（Vinyals 等人，2015）来预测风格，其中我们对每个输入句子（注意力集中于“<S$_k$>”）或整个文档（注意力集中于“<S>”）进行加权。如果“<S>”获得最大的注意力分数，我们选择抽象式风格；如果 <S$_k$> 获得最高的注意力分数，我们选择抽取式风格。

在摘要开头或摘要句末尾，我们预测下一个摘要句的风格。我们将 tokenoutput 与编码器输出进行匹配，以决定选择。

$$y_{match} = y_{out} \times (x_{out} * \alpha + x_{emb} * (1 - \alpha))^T, \tag{7}$$

其中，α 是一个可训练的标量，用于混合编码器输出 x_{out} 和词元嵌入 x_{emb}。我们将这些混合嵌入与解码器输出 y_{out}进行匹配，得到指针分布的 logits y_{match}。我们仅保留句子标识符对应的 logits，包括“<S>”和“<S$_k$>”，并使用 softmax 函数预测分布。

改写器和生成器。我们使用标准解码器作为改写器和生成器的骨干网络。对于改写器，我们遵循 Bao 和 Zhang(2021)，将组标签嵌入应用于解码器的输入。对于生成器，我们不应用组标签嵌入，因此它不对应任何文档句子。

形式化地，给定摘要 $Y = \{w_j\}|_{j=1}^{|Y|}$，我们根据标识符 token“<S$_k$>”和“<S>”，从 Y 唯一生成 G_Y：对于以“<S$_k$>”开头的句子中的每个 token w_i，其组标签为 k；而对于以“<S>”开头的句子中的每个 token，其组标签为 0。例如，如果 Y 是序列“<S$_2$>w_1 w_2 </S> <S> w_3 w_4 </S> <S$_7$> ...”，那么 G_Y 将为 $\{2, 2, 2, 2, 0, 0, 0, 0, 7, ...\}$。

我们不更改解码层，每层都包含一个自注意力模块、一个交叉注意力模块和一个前馈模块：

$$\begin{aligned} y^{(l)} &= \text{LN}(y^{(l-1)} + \textsc{SelfAttn}(y^{(l-1)})), \\ y^{(l)} &= \text{LN}(y^{(l)} + \textsc{CrossAttn}(y^{(l)}, x_{out})), \\ y^{(l)} &= \text{LN}(y^{(l)} + \textsc{FeedForward}(y^{(l)})), \end{aligned} \tag{8}$$

其中 LN 表示层归一化（Ba 等人，2016）。最后一层 L 输出最终的解码器输出 y_{out}。解码器输出 y_{out} 随后与词元嵌入进行匹配，以预测下一个词元。

4.4 训练与推理

对于词元预测和风格预测，我们均使用 MLE损失。整体损失的计算方式如下

$$\mathcal{L} = \mathcal{L}_{token} + \kappa * \mathcal{L}_{style}, \tag{9}$$

其中，$\mathcal{L}_{style}$ 是样式预测的 MLE 损失，$\mathcal{L}_{token}$ 是词元预测的 MLE 损失。由于样式与

数据集	方法	R-1	R-2	R-L
CNN/DM	BRIO(Liu 等人, 2022)	47.78	23.55	44.57
	SimCLS（Liu 和 Liu，2021）	46.67	22.15	43.54
	GSum (Dou et al., 2021)	45.94	22.32	42.48
	PEGASUS（Zhang 等人，2020）	44.17	21.47	41.11
	GOLD（Pang 和 He，2020）	45.40	22.01	42.25
	BERT-Rewriter（Bao 和 Zhang，2021）	43.52	20.57	40.56
	BERT+复制/改写（Xiao 等人，2020）	42.92	19.43	39.35
	BERT-Ext+Abs+RL（Bae 等人，2019）	41.90	19.08	39.64
	公平设置			
	BART-Rewriter(Bao 和 Zhang，2023)	44.26	21.23	41.34
	BART（Lewis 等人，2020）	44.16	21.28	40.90
	GEMINI（我们的方法）	**45.27***	**21.77***	**42.34***
XSum	BRIO(Liu 等，2022)	49.07	25.59	40.40
	GSum (Dou 等,2021)	45.40	21.89	36.67
	PEGASUS（Zhang 等人，2020）	47.21	24.56	39.25
	GOLD（Pang 和 He，2020）	45.85	22.58	37.65
	公平设置			
	HYDRASUM (Goyal 等，2022)	44.61	20.91	36.17
	BART (Lewis 等, 2020)	45.14	22.27	37.25
	GEMINI（我们的方法）	**45.86***	**22.55***	**37.68***
WikiHow	RefSum (Liu 等人，2021)	42.12	18.13	40.66
	GSum (Dou 等人，2021)	41.74	17.73	40.09
	PEGASUS (Zhang 等人，2020)	41.35	18.51	33.42
	公平设置			
	BART (Lewis 等人，2020)	41.46	17.80	39.89
	GEMINI（我们的方法）	**42.43***	**19.36***	**41.12***

表 4：基于 ROUGE 分数的<总自动评估>，其中 * 表示通过 t 检验，与 BART 基线相比的显著性 $p < 0.01$。为确保公平比较，我们主要列出基于 BART 的模型作为基线，将其他预训练模型、集成方法和重排序技术排除在比较范围之外。

对于标记预测，我们使用超参数 κ 来协调它们的收敛速度。在实践中，我们根据开发集上的最佳性能选择 κ。

在<推理>过程中，首先激活风格控制器，以决定采用抽取式风格还是抽象式风格。如果选择抽取式风格，则会使用匹配的句子标识符“<S_k>”为后续词元生成组标签。如果选择抽象式风格，则将词元设置为特殊的组标签 0。

5 实验设置

我们使用三个英语基准数据集，如表3所示，它们代表了不同的风格和领域。

CNN/DailyMail（Hermann 等人，2015）是最受欢迎的单文档摘要数据集，由线上新闻文章和人工撰写的摘要组成。

XSum (Narayan 等人，2018）是一种抽象式风格摘要数据集，基于新闻文章构建，每篇文章都有一条由专业作者撰写的一句摘要。

WikiHow (Koupaee 和 Wang，2018) 是一个从由人类作者撰写的在线知识库 WikiHow 中提取的多样化摘要数据集。

为了生成预言机风格，我们分别为 CNN/DM、XSum 和 WikiHow 使用融合指数阈值 $\gamma = 0.7$、$\gamma = 0.7$ 和 $\gamma = 0.3$。

我们采用两阶段策略训练 GEMINI：先预微调新参数，再联合微调所有参数。由于我们为预训练的 BART 引入了额外结构和新参数，直接联合微调这两类参数可能会导致预训练参数的性能下降。我们采用预微调 来准备随机初始化的参数：冻结预训练参数，并先微调 8 个轮次，再进行联合微调。两个阶段使用相同的最大似然估计（MLE）损失，并采用收敛协调参数 $\kappa = 1.1$。

6 结果

6.1 自动评估

如表4 所示，我们在三个基准数据集上评估了我们的模型，并在公平的设置下与其他基于BART 的模型进行比较。我们报告自动评估指标ROUGE-1/2/L（林，2004）。

与抽象式 BART 基线相比，GEMINI 在 CNN/DM、XSum 和 WikiHow 上的 ROUGE 分数平均分别提高了 1.01、0.48 和 1.25。各项提升在

方法	信息。	浓度。	阅读。	忠实度。
BART	3.78	3.10	3.25	4.74
BART-Rewriter	**3.90**	3.45	3.12	**4.85**
GEMINI	**3.93**	**4.27**	**3.86**	**4.82**

表 5：在 CNN/DM 上对信息量、简洁性、可读性和忠实度进行的人工评估 。与其他指标保持不变相比，GEMINI 大幅提升了简洁性和可读性。

CNN/DM 的ROUGE-L和 Wiki-How的 ROUGE-2 尤为显著，分别达到 1.44 和 1.56。结果表明，自适应方法明显优于纯抽象式模型。

与改写器基线 BART-Rewriter（依赖外部抽取器提供待改写句子）相比，GEMINI 在 CNN/DM 上的 ROUGE-1/L 分数提高了约 1.0，证明了自适应方法相较于纯改写器的有效性。与 HYDRASUM 相比，后者通过专家混合隐式表达摘要风格，但 ROUGE 分数低于 BART 基线；GEMINI 借助显式风格控制取得了更高的 ROUGE 分数。结果表明，与纯改写器和隐式风格控制相比，采用自适应风格的模型更具优势。

表 4 列出了采用更大规模预训练、组装、重排序和强化学习技术的近期工作。然而，这些模型与我们的模型无法直接比较。例如，PEGASUS（large）有 5.68 亿个参数，比拥有 4 亿个参数的 BART（large）多 42%。BRIO 使用 4 块 NVIDIA RTX 3090 GPU，在 CNN/DM 上训练时每个训练轮次需要 20 小时，总共需要 15 个训练轮次；而我们的模型在相同设备上每个训练轮次仅需 2 小时，总共训练 11 个轮次，所用计算资源仅为 BRIO 的 7%。我们不将这些模型作为基线，因为它们属于不同的研究方向，采用的技术与本研究并不相同。更重要的是，我们关注句子级摘要风格及其控制，而非仅仅提升 ROUGE。

6.2 人工评估

我们开展了人工评估，以定量衡量生成的摘要质量。我们将GEMINI 与BART 和 BART-Rewriter 基线进行比较，研究四项质量指标，包括 信息丰富度、 简洁性、 可读性和 忠实性。我们遵循近期的 HYDRASUM（Goyal 等人，2022）以及 SummaReranker（Ravaut 等人，2022) 从CNN/DM 的测试集中抽取50 份文档进行评估。我们聘请三名具有专业英语水平（托福成绩为 120分中的100 分以上）的研究生，将每份候选摘要从 1 分（最差）到 5 分（最佳）进行标注，并报告三位标注者的平均分。

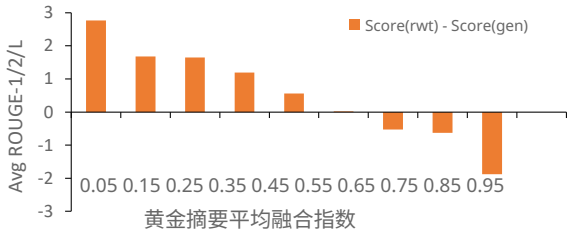


图3：在 CNN/DM 上评估时，融合指数较低的分桶中，改写器（rwt）的平均 ROUGE 得分高于生成器（gen）。

如表5所示，GEMINI 的总体得分最佳，尤其是在简洁性和可读性方面。与 BART-Rewriter 相比，GEMINI 在信息量和忠实度方面的得分接近，但在简洁性和可读性方面得分更高。与 BART 相比，GEMINI 在四项指标上的得分均更高。

GEMINI 的显式风格控制充当了粗略规划器的角色，限制了内容的生成。在 CNN/DM 上，实证结果表明，与BART（3.9 句子数/摘要）和 BART-Rewriter（3.7句子数/摘要）相比，GEMINI 生成的摘要包含更少的句子（3.3句子数/摘要）。因此，GEMINI 生成的摘要更加简洁，其长度比 BART短 20%，比 BART-Rewriter 短 10%；但它仍然获得了更高的 n-gram 召回率，正如表4 所示的更高 ROUGE-1/2，这表明 GEMINI生成的摘要往往包含更密集的信息。我们推测，GEMINI的高可读性源于其对风格预测和句子生成的自回归建模，通过这种方式，风格转换得到了优化。我们在附录A 中列出了两个案例，以说明 GEMINI 在简洁性和可读性方面的优势。

6.3 消融实验

改写器与生成器。我们通过分别观察改写器和生成器的贡献，进一步研究自适应模型。我们根据平均融合指数将 CNN/DM 中的测试样本划分为 10 个桶，并计算每个桶的平均 ROUGE-1/2/L 分数。我们将 con-

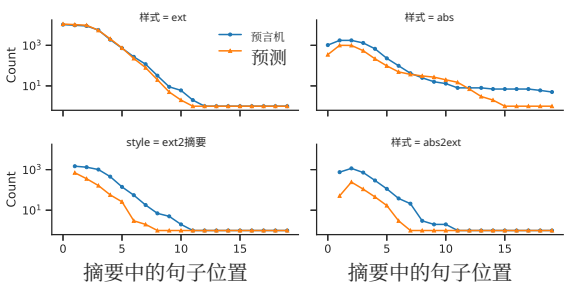


图4：在 CNN/DM 上评估的摘要中每个句子位置的风格和风格转换分布。上面的两个子图关于风格分布，而下面的两个子图关于风格转换。句子位置表示摘要中句子的序号。

方法	R-1	R-2	R-L
CNN/DM			
GEMINI	45.27	21.77	42.34
– 使用随机风格	42.94	18.77	39.93
– 使用预言机风格	**46.09**	**22.09**	**43.16**
XSum			
GEMINI	45.86	22.55	37.68
– 使用随机风格	45.71	22.43	37.56
– 使用预言机风格	**45.94**	**22.70**	**37.79**
WikiHow			
GEMINI	42.43	19.36	41.12
– 使用随机风格	41.26	17.93	39.93
– 使用预言机风格	**42.70**	**19.81**	**41.49**

表 6：风格预测与随机风格及预言机风格的贡献对比。

对比改写器（rwt）和生成器（gen）的 ROUGE 分数，如图 3所示。我们可以看到，改写器在融合指数较低的区域占据性能优势，而生成器在融合指数较高的区域占据主导。改写器和生成器的 ROUGE 分数分布说明了这两种风格的专业化。

预微调。预微调会改变随机初始化参数的分布。以 CNN/DM 上的 GEMINI 为例。句子标识符和组标签嵌入的初始平均范数均为 0.06，这是特意设定的较小值，以减少对预训练网络的负面影响。如果直接微调模型，训练后的模型中句子标识符和组标签的平均范数分别为 0.44 和 0.17。然而，如果对模型进行 8 个周期的预微调，这些平均范数会分别升至 0.92 和 0.63。经过后续微调后，平均范数分别收敛至 0.66 和 0.50，远高于直接微调的模型。

如果移除预微调，GEMINI 在 CNN/DM 上的 ROUGE-1/2/L 得分将分别从 45.27、21.77 和 42.34 降至 44.76、21.60 和 41.71，这表明预微调阶段必不可少。

7 分析

我们尝试回答两个与句子级摘要风格控制相关的关键问题，如下所示。

7.1 句子级摘要风格是否可预测?

我们使用预言机风格来构建模型的训练集，这使模型更容易进行训练。然而，在测试期间，每个句子的摘要风格可以任意选择。关于测试集的风格在一定程度上是否可预测，仍然是一个有趣的研究问题。

我们尝试通过对 CNN/DM 上的风格分布进行定量分析来回答这一问题。在风格预测方面，我们获得了 0.78 的 F1 值，更详细的分布如图 4上方的子图所示。预测风格的分布与预言机风格的分布相匹配，但在句子位置高于 15的一个狭窄范围内，预测的抽象式风格出现的可能性低于预言机抽象式风格。对于风格转换的分布，正如下方子图所示，ext-to-abs和 abs-to-ext在预测结果与预言机结果之间存在一些差异：预测的风格转换总体上不如预言机风格转换频繁，但趋势相同。这些图展示了预测风格与预言机风格在分布上的一致性，表明这些风格在一定程度上是可预测的。

我们进一步通过比较使用随机选择风格和预言机风格进行 GEMINI解码时的性能，评估预测风格的作用。如表6所示，当我们用随机风格替代预测风格时，CNN/DM 上的性能平均下降 2.58 个 ROUGE 点，WikiHow 上则平均下降 1.26 个点。性能下降表明，模型对风格的预测为生成高质量摘要提供了有用信息。预测能力源自对输入文档的理解。这也说明，在给定摘要句的现有上下文时，选择摘要风格存在一致的倾向。

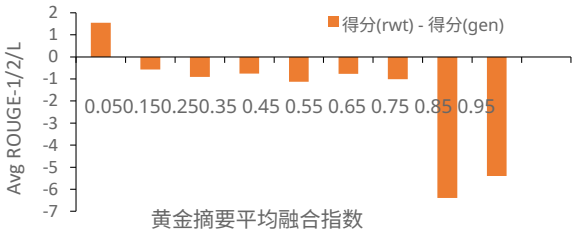


图5：在 WikiHow 上评估时，融合指数较低的分桶中，改写器（rwt）的平均 ROUGE 分数高于生成器（gen）。

7.2 自适应模型何时表现最佳?

直观而言，GEMINI 在风格平衡的数据集上效果最佳，这样改写器和生成器便能互为补充。更重要的是，预言机风格必须准确，才能获得高质量的监督信号。

首先，数据集的分布很重要。在这三个数据集中，WikiHow 的风格最为平衡，有 61.1% 的摘要句子倾向于使用改写器。因此，与纯抽象式基线相比，GEM-INI 在 WikiHow 上的性能提升最大，ROUGE 平均提高了 1.25 分，这证实了我们关于数据集分布与模型性能之间关系的直觉。

其次，互补的改写器和生成器是不可或缺的先决条件。如图3 所示，在 CNN/DM 上，改写器和生成器的能力相对均衡：当融合指数低于 0.55 时，优先选择改写器；高于 0.55 时，则优先选择生成器。相比之下，如图 5所示，WikiHow 上的改写器和生成器能力存在偏差：改写器能力较弱，仅当融合指数低于 0.15 时才会被优先选择。因此，GEMINI 仅对 19.3% 的摘要句子采用抽取式风格，而人工评估得出的比例为 61.1%。分析表明，若采用更好的改写器，WikiHow 上的 GEMINI 还可以进一步提升；而改进句子抽取器或许能够实现这一点。

最后，预言机风格的质量决定了两个生成器的专业化程度。如表2所示，融合指数在 WikiHow 上的皮尔逊相关系数仅为 0.56，远低于 CNN/DM 上的 0.76。这表明，通过开发更好的自动指标，预言机风格仍有进一步提升的空间。

8 讨论

本文在句子级模拟人工摘要风格，并进一步理解人工摘要的生成方式。基于这一理解，研究人员可以将 GEMINI 用于不同目的。首先，自动指标融合指数可用于在开发新数据集时分析风格分布。其次，GEMINI 可用于控制句子级的摘要风格。由于抽取式风格的摘要句子天然比抽象式风格的摘要句子更不容易产生幻觉，我们可以通过确定抽象式风格摘要句子的比例来控制忠实性风险。最后，GEMINI 会为每个摘要句生成明确的风格，因此，在关键应用中，我们甚至可以标记这些抽象式风格的摘要句，以提示忠实性风险。

局限性。GEMINI 可以适配特定数据集的摘要风格，使改写器和生成器能够分别擅长处理不同情况，从而提升改写器和生成器的质量。然而，我们尚不清楚这种自适应方法是否确实能提高摘要生成模型的“抽象”能力。目前，我们没有可靠的方法来衡量抽象能力，因此无法评估训练后的生成器。我们将在未来的研究中进一步探讨这一问题。

9 结论

我们研究了人工撰写摘要的句子级摘要风格，并评估了三个基准数据集的风格分布。我们提出了一种融合指数度量方法，用于自动检测句子级摘要风格，使-用该方法，我们生成预言机风格来训练自适应模型。实验表明，我们的GEMINI在基准测试中显著优于纯抽象式和改写器基线，证-明了摘要风格中细粒度控制的有效性。

致谢

我们感谢匿名审稿人提出的建设性意见。我们衷心感谢浙江省尖兵领雁研发攻关计划（编号：2022SDXHDX0003）以及中国国家自然科学基金（NSFC，编号：62161160339）提供的资助。

参考文献

Mehdi Allahyari、Seyedamin Pouriyeh、Mehdi Assefi、Saeid Safaei、Elizabeth D Trippe、Juan B Gutierrez 和 Krys Kochut。2017年。文本摘要技术：简要综述。国际先进计算机科学与应用期刊（*ijacsa*），8(10):397。

Jimmy Lei Ba、JamieRyan Kiros和GeoffreyE Hin- ton。2016 年。层归一化。*arXiv* 预印本 *arXiv:1607.06450*。

SanghwanBae、Taeuk Kim、JihoonKim和S ang- goo Lee。2019 年。面向生成式摘要的句子重写摘要级训练。在关于摘要新前沿的第 *2* 届研-讨会论文集中，第 10–20 页，香港，中国。计- 算语言学协会。

Guangsheng Bao 和 Yue Zhang。2021。用于文本摘要的上下文改写。载于 *AAAI*人工智能大会论文集，第 35 卷，第 12544–12553 页。

Guangsheng Bao 和 Yue Zhang。2023。《一种用于文本摘要的通用上下文重写框架》。*IEEE/ACM* 音频、语音与语言处理汇刊。

Regina Barzilay 和 Kathleen R McKeown。2005年。多文档新闻摘要的句子融合。计算语言学，31(3)：297–328。

Yen-ChunChen 和 Mohit Bansal。2018。通过强化选择的句子重写实现快速的抽象式摘要。在计算语言学协会第 *56* 届年会论文集（第*1* 卷：长篇论文）中，第 675–686 页，澳大利亚墨尔本。计算语言学协会。

JianpengCheng 和Maria Lapata。2016 年。通过抽取句子和词进行神经摘要。在计算语言学协会第*54*届年会，第 484–494 页。计算语言学协会。

Zi-YiDou、Pengfei Liu、Hiroaki Hayashi、Zhengbao Jiang 和Graham Neubig。2021 年。GSum：一个用于引导式神经抽象摘要的通用框架。载于 *2021*年北美计算语言学协会分会会议论文集：人类语言技术，第 4830–4842 页。

Sebastian Gehrmann、Yuntian Deng 和 Alexander M Rush。2018。自底向上的生成式摘要。载于*2018*年自然语言处理实证方法会议论文集，第 4098–4109 页。

Tanya Goyal、Nazneen Rajani、Wenhao Liu 和 Wo- jciech Kry´sci´nski。2022。Hydrasum：借助多解码器模型解耦文本摘要中的风格特征。载于《*2022*年自然语言处理实证方法会议论文集》，第464–479页。

Max Grusky、Mor Naaman 和 Yoav Artzi。2018。Newsroom：一个包含130万个摘要、采用多种抽取式策略的数据集。载于《*2018*年计算语言学协会北美分会会议论文集：人类语言技术，第*1*卷（长篇论文）》，第708–719页，美国路易斯安那州新奥尔良。计算语言学协会。

Karl Moritz Hermann、Tomás Kociský、Edward Grefen- stette、Lasse Espeholt、Will Kay、Mustafa Suleyman 和 Phil Blunsom。2015年。教机器阅读和理解。载于第*28*届神经信息处理系统进展：*2015*年神经信息处理系统年会，*2015*年*12*月7日至*12*日，加拿大魁北克省蒙特利尔，第1693–1701页。

Dandan Huang、Leyang Cui、Sen Yang、Guangsheng Bao、KunWang、Jun Xie 和 YueZhang。2020 年。我们在文本摘要方面取得了哪些成就？载于*2020*年自然语言处理实证方法会议论文集（*EMNLP*），第 446–469 页。

Robert A Jacobs、Michael I Jordan、Steven J Nowlan和Geoffrey E Hinton。1991 年。局部专家自适应混合。神经计算，3(1)：79–87。

Hongyan Jing 和 Kathleen R McKeown。1999。人工撰写的摘要句子的分解。载于 第*22*届 *ACM SIGIR* 信息检索研究与发展国际年会论文集，第129–136页。

Mahnaz Koupaee 和 William Yang Wang。2018。WikiHow：一个大规模文本摘要数据集。*arXiv* 预印本 *arXiv:1810.09305*。

Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020 年。BART：用于自然语言生成、翻译和理解的去噪序列到序列预训练。载于 第 *58* 届计算语言学协会年度会议论文集，第 7871–7880 页，线上。计算语言学协会。

Chin-YewLin. 2004. ROUGE：一个用于摘要自动评估的软件包。载于 文本摘要的拓展，第74–81页。

Yang Liu 和 Mirella Lapata。2019年。使用预训练编码器进行文本摘要。在*2019*年自然语言处理经验方法会议暨第*9*届自然语言处理国际联合会议（*EMNLP-IJCNLP*），第3730–3740页，中国香港。计算语言学协会。

YixinLiu、Zi-YiDou和 Pengfei Liu。2021年。Refsum：重构神经摘要。载于计算语言学协会北美分会 *2021* 年会议论文集：人类语言技术，第 1437–1448 页。

Yixin Liu 和 Pengfei Liu。2021。Simcls：用于生成式摘要对比学习的简单框架。载于 计算语言学协会第 *59* 届年会暨第 *11* 届自然语言处理国际联合会议论文集（第 *2* 卷：短论文），第 1065–1072 页。

Yixin Liu、Pengfei Liu、Dragomir Radev和GrahamNeubig。2022 年。BRIO：为生成式摘要带来秩序。载于计算语言学协会第*60* 届年会论文集（第*1* 卷：长篇论文），第 2890–2903 页。

Mani Maybury.1999. 自动文本摘要技术进展。MITpress。

Joshua Maynez、Shashi Narayan、Bernd Bohnet和 RyanMcDonald。2020年。论生成式摘要的忠实性和事实性。载于第*58*届计算语言学协会年会论文集，第1906–1919页。

Ramesh Nallapati、Feifei Zhai 和 Bowen Zhou。2017年。Summarunner：一种基于循环神经网络的序列模型，用于文档的抽取式摘要。载于第三十一届*AAAI*人工智能大会。

Ramesh Nallapati、Bowen Zhou、Cicero dos Santos、Ça˘glarGulçehre和BingXiang。2016 年。使用序列到序列 RNN 及更进一步的抽象式文本摘要。在第 *20* 届*SIGNLL* 计算自然语言学习会议论文集中，第 280–290 页。

Shashi Narayan、ShayB Cohen 和 MirellaLapata。2018 年。别告诉我细节，只给我摘要！面向主题的卷积神经网络用于极端摘要。在*2018*年自然语言处理实证方法会议论文集中，第 1797–1807 页。

Ani Nenkova 和 Kathleen McKeown。2012。《文本摘要技术综述》。载于文本数据挖掘，第 43–76 页。Springer。

Richard Yuanzhe Pang 和 He He。2020。通过从演示中学习进行文本生成。载于国际学习表征会议。

Mathieu Ravaut、Shafiq Joty 和 Nancy Chen。2022。Summareranker：一种用于生成式摘要的多任务混合专家重排序框架。载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第4504–4524页。

Alexander M Rush、Sumit Chopra 和 Jason Weston。2015年。一种用于抽象式句子摘要的神经注意力模型。载于*2015*年自然语言处理实证方法会议论文集，第379–389页。

Abigail See、Peter J. Liu 和 Christopher D. Manning。2017年。切中要点：采用指针生成器网络进行摘要。载于计算语言学协会第*55*届年会论文集（第*1*卷：长篇论文），第1073–1083页，加拿大温哥华。计算语言学协会。

Oriol Vinyals、Meire Fortunato 和 Navdeep Jaitly。2015。指针网络。神经信息处理系统进展，28：2692–2700。

Liqiang Xiao、Lu Wang、Hao He 和 Yaohui Jin。2020。复制或改写：结合层次化强化学习的混合式摘要。*AAAI*人工智能大会论文集（*AAAI*）。

Jingqing Zhang、Yao Zhao、Mohammad Saleh 和 Peter Liu。2020。PEGASUS：通过抽取的间隔句进行预训练，以实现生成式摘要。载于国际机器学习大会，第11328–11339页。PMLR。

Ming Zhong、Pengfei Liu、Yiran Chen、Danqing Wang、Xipeng Qiu 和 Xuan-Jing Huang。2020。《抽取式摘要即文本匹配》。载于第*58*届计算语言学协会年会论文集，第6197–6208页。

Qingyu Zhou, Nan Yang, Furu Wei, ShaohanHuang, MingZhou, 和TiejunZhao。2020。用于神经抽取式文档摘要的联合句子评分与选择框架。*IEEE/ACM*音频、语音与语言处理汇刊，28:671– 681。

源文本： 弗洛伊德·梅威瑟开了一家快闪店，售卖背景印有菲律宾国旗的T恤。“我喜欢这个，”帕奎奥笑着说，随后对金钱团队的领队说道：“欢迎加入曼尼团队。”在即将进行徒手格斗的最后几天里，我们习惯看到拳击手面容严肃、反应粗鲁——他们称之为比赛表情。曼尼·帕奎奥正式抵达曼德勒海湾酒店时向人群致意。帕奎奥与传奇教练弗雷迪·罗奇自拍，同时向人群致意。印有菲律宾国旗背景的金钱团队T恤正在拉斯维加斯出售。帕奎奥有了一台新相机，他遇到每个人都会用相机对着他们。你认为弗洛伊德·梅威瑟与曼尼·帕奎奥之间的比赛谁会赢？.....

目标： 曼尼·帕奎奥抵达了拉斯维加斯的曼德勒海湾酒店。这位菲律宾人将于星期六在米高梅大酒店迎战弗洛伊德·梅威瑟。这场比赛将成为有史以来奖金最高的赛事，收入超过3亿美元。帕奎奥有信心终结梅威瑟的不败纪录。

BART： 弗洛伊德·梅威瑟开了一家快闪店，出售背景印有菲律宾国旗的T恤。曼尼·帕奎奥是拉斯维加斯最快乐的人，面带笑容、妙语连珠，热情欢迎每一个人。他最新的玩具是一台相机，遇到谁就把镜头对准谁，还说：“我不太懂这些东西，但我喜欢拍肖像。”“现在是我和弗洛伊德交手的好时机。如今的我比五年前更有经验，速度和力量依然在线。”

BART-Rewriter： 弗洛伊德·梅威瑟开了一家快闪店，出售背景印有菲律宾国旗的T恤。帕奎奥笑着说：“我喜欢这个。”随后对金钱团队的领头人说：“欢迎加入曼尼团队。”帕奎奥有了一台新相机，遇到谁就把镜头对准谁。

GEMINI: 随着拉斯维加斯比赛周持续进行，数千名粉丝热烈欢迎曼尼·帕奎奥。这位菲律宾拳击手信心十足，认为自己能在星期六晚上击败弗洛伊德·梅威瑟。在这场有史以来奖金最高的比赛中，帕奎奥被视为对阵梅威瑟的弱势一方。

源文本： 剑桥公爵夫人预计现在随时可能临产，但一名软件设计师是否预测了王位第四顺位继承人的长相？程序员Charanjeet Kondal使用一款应用程序创建了新王子或公主的图像。如果他的预测正确，这位继承人在两至四岁时将拥有飘逸的金色头发、深棕色眼睛和一个小鼻子。下拉查看视频。王室宝宝？一名软件程序员预测，这位继承人在两至四岁时将拥有飘逸的金色头发、深棕色眼睛和一个小鼻子。公爵夫人预计现在随时可能临产，但是否有人预测了宝宝会长什么样。新王子或公主的图片是使用 Kondal 先生的 Morph Thing 应用程序创建的。 ...

目标： 所以 软件设计师 CharanjeetKondal使用应用程序创作了新王室成员的图像。如果他是对的，继承人将会有金发头发，深色 棕色眼睛和一个小鼻子。投注公众相信王室宝宝将在星期三降临人世。
BART： 程序员 Charanjeet Kondal 使用一款应用程序制作了新王子或公主的图像。如果他说得对，这位继承人将有一头飘逸的金色头发、深棕色眼睛和一个小鼻子。凯特首次怀上乔治王子时宣布怀孕后，他也运行了同一款软件。

BART-Rewriter： 程序员 Charanjeet Kondal 使用一款应用程序生成了图像。他预测继承人会有一头飘逸的金色头发、深棕色眼睛和一个小鼻子。这些图片是使用 Kondal 先生的 Morph Thing 应用程序生成的。凯特宣布怀孕时，他也运行了同一款软件。

GEMINI： 程序员 Charanjeet Kondal 使用应用程序创建了新王子或公主的图像。如果他的判断正确，这位继承人在两到四岁时将有一头纤细的金色头发、深棕色的眼睛和一个小鼻子。

表7：在 *CNN/DM* 上评估的 BART、BART-Rewriter 和 GEMINI 生成的摘要对比。

案例研究

如表 7 所示，GEMINI 生成的摘要通常比
BART 和 BART-Rewriter 生成的摘要更易读、
更简洁。