arXiv:2302.01342v2 [cs.CL] 8 Feb 2023

课程引导的抽象式摘要

Sajad Sotudeh ♡*，**Hanieh Deilamsalehy**♠，**Franck Dernoncourt**♠，以及**Nazli Goharian**♡

♡IRLab，乔治城大学
{sajad,nazli}@ir.cs.georgetown.edu♠Adobe Research
{deilamsa,franck.dernoncourt}@adobe.com

摘要

近期基于 Transformer 的摘要模型为抽象式摘要提供了一种有前景的方法。它们不再局限于句子选择和抽取式策略，而是能够处理新词生成和句子改写等更复杂的任务。然而，这些模型存在两个缺点：(1) 它们在内容选择方面通常表现不佳；(2) 其训练策略不够高效，从而限制了模型性能。在本文中，我们探索了两种相互独立的方式来弥补这些不足。首先，我们在解码器中为 Transformer 网络增添句子交叉注意力模块，以鼓励模型对显著内容进行更多抽象。其次，我们引入课程学习方法，对训练样本重新加权，从而实现高效的学习过程。与第一种方法相比，我们用于改进 Transformer 网络训练策略的第二种方法带来了更显著的提升。我们将模型应用于极端摘要数据集 Reddit TIFU 帖子。随后，我们进一步研究了三个跨领域摘要数据集（*Webis-TLDR-17*、*CNN/DM* 和*XSum*），评估课程学习应用于摘要任务时的有效性。此外，我们还开展了人工评估，以展示所提方法在流畅度、信息量和整体质量等定性标准方面的有效性。

1 引言

文本摘要系统旨在将一段文本压缩成更短的形式，同时保留原文中的主要信息。这项任务主要通过两种方式完成：(1) 抽取式（Nallapati 等人，2017年；Xiao 和 Carenini，2019年；Xu 等人，2020b），直接从源文本中选取重要句子；(2) 抽象式（Ce- likyilmaz 等人，2018年；Lebanoff 等人，2018年；Liu 和 Lapata，2019年；Zou 等人，2020；Sotudeh 等人，2021年），涉及改写并生成源文本中未出现的新词。近期的研究还采用混合模型（See 等人，2017年；Hsu 等人，2018年；Chen 等人，2019年；MacAvaney 等人，2019年）。

过去十年间，基于循环神经网络（Hochreiter 和 Schmidhuber，1997）和 Transformer（Vaswani 等人，2017年）的神经摘要模型在文本摘要方面取得了可喜的成果。近期，预训练语言模型在包括摘要在内的各种下游任务中取得成功，推动最先进水平迈上了新的台阶。尽管这类方法生成的摘要流畅，但已有少数研究指出，内容选择是它们的一个缺陷（Gehrmann 等人，2018年；Narayan 等人，2020），这限制了模型生成信息丰富的摘要的性能。在本研究中，我们旨在通过引入句子级信息作为显著性信号来扩展解码器，以便在总结源文本时加以利用。

大规模深度神经模型往往难以训练；训练时通常依赖复杂的启发式设置，而这些设置需要耗费大量时间和成本进行调优（Gong 等人，2019年；Chen 等人，2021年）。对于 Transformer 来说尤其如此：研究表明，经过严格调优后，它们的表现始终优于 RNN 网络（Popel 和 Bojar，2018年），但也需要采用专门的学习率、大批量训练等启发式方法（Platanios 等人，2019年）。本文尝试通过为摘要模型的训练引入一种课程学习（*CL*）策略来解决上述问题，从而缩短收敛时间并提升性能。受人类教学方式的启发，课程学习主张将教学过程从较容易的样本逐步过渡到较困难的样本，其历史可追溯至 20 世纪 90 年代（Elman，1993年）。这一方法背后的核心思想是，网络能够更好地完成任务学习，当训练实例按照特定顺序呈现时，即从较容易的样本到较困难的样本（

* 部分工作是在 Adobe Research 实习期间完成的。

Chang 等人，2021年）。在神经网络领域，这一过程可以视为一种技术，使网络能够避免陷入局部最优；而在训练初期，这种情况更容易发生。据报道，采用课程学习的系统表现出较强的泛化能力、更快的收敛时间，甚至模型性能也有所提升（Platanios 等人，2019年）。

在发现基于 Transformer 的网络存在上述两项缺陷（即内容选择和训练效率低下）后，我们构建了自己的摘要框架来解决这些问题。为弥补第一项缺陷，我们提出为 Transformer 解码器增设一个句子交叉注意力层，鼓励解码器在生成抽象式摘要时更多关注源文本中的重要句子。对于第二项缺陷，我们为摘要模型提供课程学习目标。具体而言，我们采用 SUPERLOSS（Castells 等人，2020）函数，该函数属于置信度感知课程学习技术范畴，并为网络引入一个名为置信度的新参数（即σ）。尽管学习这一参数的效率较低，尤其是在摘要任务这类训练实例数量庞大的情况下，SUPERLOSS 通过直接使用特定学习阶段的置信度收敛值，克服了这一局限。我们在极限摘要任务（Narayan 等人，2018）上验证了模型，该任务旨在以极高的压缩率和抽象程度生成一句话摘要。为此，我们使用 Reddit TIFU（Kim 等人，2019）和 Webis-TLDR-17（Völske 等人，2017）数据集，分别包含 42k 和 4M 个实例，每个实例对均包含一篇 Reddit 帖子及其 TLDR[1] 摘要。为衡量模型的跨领域性能，我们还报告了模型在 CNN/DM（See 等人，2017）和 XSum（Narayan 等人，2018）这两个大规模新闻数据集上的表现。我们展示了引入课程学习能够显著提升神经网络式 Transformer 摘要系统的性能。此外，我们还进行了全面的人工评估，从流畅度、信息丰富度和整体质量三个定性指标考察模型的效果。

[1]TLDR 是“太长了，没读”的缩写。

2 相关工作

预训练语言建模。过去几年中，自监督预训练语言模型因其在多种自然语言处理任务上的显著改进，日益受到研究界的关注。这类模型的不同变体在大量无标注数据上进行预训练（Devlin 等人，2019年；an，2019年；Peters 等人，2018年），并采用各种预训练目标。尽管这类模型最初是为执行语言建模任务而提出的，但如今已可以对其进行微调，以适用于广泛的下游自然语言处理任务，其中就包括摘要。Liu andLapata （2019）率先将 BERT 微调用于摘要任务。他们具体提出了 BERTSUM 的三种变体：用于抽取式摘要的 BERTSUMEXT、用于抽象式摘要的 BERTSUMABS，以及采用两阶段微调方法、结合抽取式和抽象式目标的 BERTSUMEXTABS。沿着这一研究方向，Zhang 等人（2020a）提出了 PEGASUS，其预训练目标专为文本摘要设计，并在 12 个下游摘要任务上取得了最先进水平的结果。与此同时，Lewis 等人 （2020）提出了 BART，并展示了它在文本摘要等语言生成任务中的有效性。不同于仅使用预训练 BERT 编码器的 BERTSUM，PEGASUS 和 BART 在语言生成中同时利用预训练编码器和解码器。

句子引导摘要。将句子表示作为抽取式信号，与神经序列到序列模型中的词元嵌入结合使用，近年来已有相关研究。这一想法源于这样一个事实：尽管通用的编码器—解码器框架能够生成流畅的目标文本，但在内容选择方面往往表现不佳（Sotudeh 等人，2020）。一些研究指出，结合抽取式和抽象式目标可以解决这一问题（Gehrmann 等人，2018）。尽管已有大量工作尝试在传统的 RNN 网络中结合这类目标（See 等人，2017；Chen 等人，2019；Lebanoff 等人，2018），但只有少数研究探讨了它们在基于 Transformer 的网络中的有效性。例如，Liu and Lapata（2019）提出了 BERTSUMEXTABS，利用 BERTSUMEXT 模型中的抽取式目标，并进一步将其与 Transformer 解码器结合，以执行抽象式摘要。最近，Akiyama 等人（2021）提

出了 HIE-BART，该模型添加了一个自注意力层，以便在编码器中融入句子重要性。尽管他们在编码器端加入了句子级信息，但文献中尚未探讨在解码器中加入此类模块。据我们所知，我们的论文首次在基于 Transformer 的网络中探索这一方向。

课程学习。课程学习（CL）（Bengio 等人，2009）在过去十年间日益受到研究界的关注（Tay 等人，2019；MacAvaney 等人，2020；Xu 等人，2020a），尽管其教学方法（即从简单实例学习到较难实例）被称为增量学习，可追溯至 20 世纪 90 年代（Elman，1993）。该技术的核心思想是提供一种训练策略，使学习过程从简单样本逐步推进到较难样本，从而提升模型性能、缩短训练时间并增强模型的泛化能力（Chang 等人，2021）。Bengio 等人（2015年）率先将这一策略应用于使用 RNN 网络进行序列预测的情境，并提出了计划采样方法，在解码过程中逐步将训练过程从真实标记过渡到模型生成的标记。Platanios 等人（2019）提出了一种基于课程学习的神经机器翻译框架，根据训练样本的难度以及模型的能力状态来决定学习哪些训练样本。在这一方案中，样本难度是一个关键概念，用于区分简单样本和困难样本。研究人员采用了许多文本特征作为“难度度量”，包括 n 元语法频率（Haffari 等人，2009）、词语稀有度和句子长度（Platanios 等人，2019）。近期研究（Saxena 等人，2019；Cachola 等人，2020）采用了置信度感知方法，以学习训练样本的难度，并在训练过程中动态调整样本权重。

3 我们的方法

本节将详细介绍我们提出的模型，包括：（1）对 BART 模型的扩展，即在 BART 解码器中添加一个交叉注意力层；（2）在 BART 的 Transformer 架构框架上加入课程学习架构，对较容易的训练样本赋予更高权重，从而提高它们在学习阶段的贡献。这两项扩展均可添加到 BART 的 Transformer 网络中，并且可联合训练，也可单独训练。

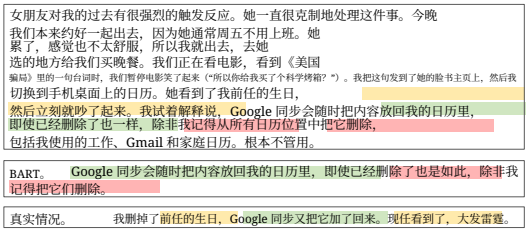


图 1：BART 在内容选择方面的不足。黄色：真实情况选中但 BART 跳过的内容；绿色：BART 和真实情况均选中的内容；红色：BART 选中但真实情况跳过的内容

3.1 句子引导的 BART

尽管 BART 凭借强大的预训练编码器和解码器，在生成抽象式摘要方面展现出良好前景，但它存在一个主要缺陷，限制了模型在内容选择上的效能。图 1 展示了一篇 Reddit 帖子以及 BART 生成的 TLDR 摘要和真实情况，说明了这一不足。正如观察到的那样，生成摘要虽然行文流畅、表达良好，却忽略了源文本中显著的部分，转而聚焦于不太重要的内容。鉴于结合抽取式目标与抽象式目标的有效性，我们通过添加一个交叉注意力层来扩展 BART，以便在解码时判断句子的重要性。为此，我们首先定义一个序列标注任务，目标是预测句子的相对重要性得分（即 $\mathbf{y}$），从而微调 BART，使其学习句子显著性。相对重要性 得分被定义为源句子相对于 TLDR 摘要的 ROUGE-2 和 ROUGE-L 得分的归一化均值：

$$\mathbf{y} = \text{relative importance}(s_i) = \frac{\text{RG}_{2+\text{L}}(s_i)}{\sum_{s_i \in R} \text{RG}_{2+\text{L}}(s_i)} \quad (1)$$

其中，s_i 是位于第 i个位置的句子，R 是帖子中句子的集合，而 $\text{RG}_{2+\text{L}}(\cdot)$ 是一个函数，它接收一个源句子，并输出该句子的 ROUGE-2 和 ROUGE-L 得分的平均值。为了将序列分类任务适配为序列句子标注问题，我们在每个输入句子的末尾插入 </s> 词元（即 BART 词汇表中的句子结束符），然后将其输入 BART 网络，这与Liu andLapata （2019年）对 BERTSUM 所采用的方法类似。BART 通过其网络对每个输入词元进行编码，与 </s>词元相关联的编码具体表示其前面的输入句子的特征。

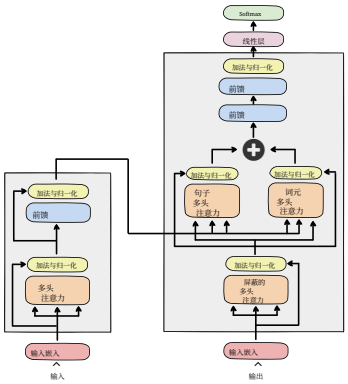


图2：我们摘要模型的概览

这是因为 BART 使用</s> 词元的表示作为分类头（Lewis 等人，2020）。获得与 </s> 词元相关联的表示后，我们通过带有 Sigmoid 分类器的线性层对其进行处理，以输出作为句子重要性分数的概率。形式化地说，设$\mathbf{P}$ 为一个包含句子$\mathbf{P} = [sent_1, sent_2, ..., sent_i, ..., sent_n]$和 $sent_i = [x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{im}]$的 Reddit 帖子。我们通过在输入$\mathbf{P}$ 末尾添加 </s> 词元，并在每个句子的开头添加 <s>，来构造输入。这样，输入 BART 网络的修改后内容为$\mathbf{P}' = [\texttt{<s>}sent_1\texttt{</s><s>}sent_2...\texttt{</s><s>}sent_n\texttt{</s>}]$，随后将其输入 BART 网络进行处理。该网络经过训练，用于预测重要性分数（即，方程1中的$\mathbf{y}$）。通过训练这样的序列标注网络，我们旨在向 BART 编码器和解码器注入归纳偏置，使其更充分地了解源句子的重要性，从而在生成抽象式摘要时发挥作用。

在下一阶段，我们通过为解码器添加一个交叉注意力层（即解码器内部的句子多头注意力），按照图2所示设计框架。为此，我们采用两阶段微调方法：首先，在序列标注任务上微调编码器模块和句子多头注意力；其次，进一步在抽象式摘要任务上对其进行微调。我们将预训练部分（即编码器和句子多头注意力模块）与解码器的优化器分开。也就是说，预训练部分应以较低的学习率进行微调，从而在解码器趋于稳定时，使用更准确的梯度进行训练。我们在实验中将该模型命名为 SENTSUM。

3.2 用于 BART 的课程学习器

课程学习（CL）（Bengio 等人，2009）是一种训练范式，其理念是在训练过程中先处理简单样本，再处理困难样本，以提升学习模型的性能和泛化能力（Castells 等人，2020）。这是因为模型在训练早期从较简单的样本开始学习，可以降低陷入局部最优的风险；深度神经网络中的大多数损失函数高度非凸（Chang 等人，2021年），因此难以收敛。考虑到课程学习在训练大规模网络中的适用性，我们计划将其用于摘要框架。在将课程学习策略融入模型训练阶段之前，我们首先需要定义难度指标，以区分样本的难易程度。

实际上，为每个样本估计先验难度被认为是一项复杂任务，因此我们提出在训练过程中利用渐进式信号区分样本——例如每次训练迭代中的相应样本损失。在此背景下，课程学习通过在每次训练迭代中以权重的形式预测每个样本的难度来实现，使得难样本在训练早期获得较低权重，反之亦然。为了对课程学习进行建模，我们提出使用 SUPERLOSS（Castells 等人，2020），这是一种基于任务损失函数构建的通用损失准则。具体而言，SUPERLOSS 是一种与任务无关的置信度感知损失函数，它包含两个参数：(1) 任务损失 $\mathcal{L}_i = \ell(y_i, \hat{y}_i)$，其中 y_i 是神经网络（即 BART 生成的摘要）的输出，而 $\hat{y}_i$ 是金标准标签（即真实摘要）；(2) σ_i，作为第 i 个样本的置信度参数。SUPERLOSS 表示为 $\mathrm{L}_\lambda(\mathcal{L}_i, \sigma_i)$，计算如下，

$$\mathrm{L}_\lambda(\mathcal{L}_i, \sigma_i) = (\mathcal{L}_i - \tau)\sigma_i + \lambda(\log \sigma_i)^2 \quad (2)$$

其中，λ 是正则化参数，τ 是训练期间任务损失的动态平均值或静态平均值（即 $\mathcal{L}$）。虽然 SUPERLOSS 为课程学习策略提供了一种定义明确的方法，但对于文本摘要等拥有大量训练实例的任务，学习 σ 参数并不可行。为解决这一问题并避免引入新的可学习参数，SUPERLOSS 建议使用 σ_i 在极限情况下的收敛值，

$$\sigma_\lambda^*(\ell_i) = \arg\min_{\sigma_i} \mathrm{L}_\lambda(\ell_i, \sigma_i)$$
$$\mathrm{SL}_\lambda(\ell_i) = \mathrm{L}_\lambda(\ell_i, \sigma_\lambda^*(\ell_i, \sigma_i)) = \min_{\sigma_i} \mathrm{L}_\lambda(\ell_i, \sigma_i), \tag{3}$$

采用这种技术后，训练期间无需学习置信度参数。Castells 等人(2020) 发现 $\sigma_\lambda^*(\ell_i)$ 存在闭式解，计算如下：

$$\sigma_\lambda^*(\ell_i) = e^{-W(\frac{1}{2}\max(-\frac{2}{e}, \beta))}, \beta = \frac{\ell_i - \tau}{\lambda} \tag{4}$$

其中， W 是 Lambert W 函数。有鉴于此，SUPERLOSS 会在训练期间动态增加较容易样本的权重，从而为摘要任务提供一种课程学习方法。我们将该模型称为 CURRSUM。在实验中，我们还尝试结合 SENTSUM 和 CURRSUM 模型，并将其命名为 CURRSENTSUM。

4 实验设置

4.1 数据集

在实验中，我们使用两个 Reddit 摘要数据集：Reddit TIFU 数据集（Kim 等人，2019年）和 Webis-TLDR-17（Völske 等人，2017年）；此外还使用两个知名的新闻摘要数据集：CNN/DM（See 等人，2017年）和 XSum（Narayan 等人，2018年）。Reddit 数据集来自 Reddit 讨论论坛，其中包含 42k 个（Reddit TIFU）和 4M 个（Webis-TLDR-17）实例，每个实例均包含源文本（即帖子正文）以及人工撰写的 TLDR 摘要。对于 Reddit TIFU 和 Webis-TLDR-17，我们分别采用随机训练集-验证集-测试集划分：80%（33,705）-10%（4,214）-10%（4,214），以及 99%（3,771,432）-1%（38,483）-1%（38,484）。对于新闻摘要数据集，我们采用原始论文建议的划分方式。

4.2 比较

我们将模型与各种最先进的抽取式和抽象式基线系统进行比较。以下概述了每种基线系统的说明。

BERTSUMEXT (Liu and Lapata，2019年)：BERTSUM 的抽取式变体，会在句子之间插入外部[CLS] 词元。随后使用 Sigmoid 分类器处理这些表示，以计算句子抽取概率。

BERTSUMABS (Liu andLapata，2019年)：BERTSUM 的抽象式变体，采用基于 Transformer 的编码器-解码器架构。编码器为 BERT，而解码器则从头开始预训练。

BERTSUMEXTABS (Liu andLapata, 2019年)：一种两阶段微调方法。BERT-SUMEXTABS 首先在抽取式摘要任务上对编码器进行微调，然后将其与解码器一起针对抽象式摘要任务进行微调。

MATCHSUM (Zhong 等人, 2020)：一种抽取式框架，在语义空间中将源文本与候选摘要进行匹配。与常用的抽取式摘要系统不同，MATCHSUM 不依赖于逐句抽取，而是选择一组与真实摘要语义相似度最高的句子（即候选摘要）。

PEGASUS (Zhang etal., 2020a)：一种抽象式模型，将间隙句生成（GSG）定义为一项新的预训练任务。在该任务中，关键源句子会被遮蔽，网络则学习生成缺失的句子。

BART (Lewis 等人，2020)：一种抽象式模型，采用预训练的编码器—解码器架构，不同于仅使用预训练编码器的 BERT。

NEUTOPICSUMM (Nguyen 等人，2021年)：一种抽象式基线方法，将神经主题模型融入摘要框架。

BART+R3F (Aghajanyan 等人，2021b)：一种摘要方法，提出了新的微调技术；R3F方法用参数噪声取代先前提出的对抗目标，并在微调过程中抑制表征变化，同时不降低性能。

BART+MUPPET (Aghajanyan 等人，2021a)：一种架构，在预训练和微调阶段之间引入一个额外的大规模学习阶段（即预微调），并使用一个涵盖不同任务的大规模数据集集合。MUPPET旨在促进表征学习，从而提升语言模型的泛化能力。

4.3 实现细节

我们扩展了 Huggingface 的 Transformer 库[2] (Wolf 等人，2020)，以实现我们的模型，并公开提供代码，以加快未来的研究

[2]https://github.com/huggingface/transformers

工作 [3]。我们对所有模型进行训练，其中 Reddit TIFU 训练 8 个 epoch，Webis-TLDR-17、CNN/DM 和 XSum 训练 5 个 epoch，并使用在验证集上 ROUGE-L 分数最高的检查点进行推理。所有摘要模型以及 BART 均使用 AdamW 优化器（Loshchilov 和 Hutter，2019），其学习率初始化为$3e-5$，$(\beta_1, \beta_2) = (0.9, 0.98)$，权重衰减为 0.01。具体而言，训练 SENTSUM 时，我们对预训练部分使用较低的学习率 $1e-5$。除预训练 SENTSUM 使用均方误差（MSE）损失函数外，所有模型均使用交叉熵损失。对于 BERTSUM 变体，我们使用主代码库 [4] 以及原论文（Liu 和 Lapata，2019）建议的默认超参数。为了跟踪训练过程，我们使用 Weights & Biases（Biewald，2020）工具包。

5 结果

自动评估。 表 1报告了基线模型以及我们的模型在（a）Reddit TIFU 数据集上的 ROUGE 分数变体（Lin，2004）表现。如表所示，我们最优的模型是 CURRSUM，它直接在 BART 模型之上使用 SUPERLOSS，并且在所有指标上都明显优于大多数基线系统。具体而言，在 Reddit TIFU 数据集上，与不使用课程学习的基准模型（即 BART）相比，CURRSUM 在 ROUGE-1、ROUGE-2 和 ROUGE-L 上的相对提升分别为 5.2%、11.22% 和 5.4%。尽管 CURRSUM 在 ROUGE-1 方面取得了与 BART+R3F 和 BART+MUPPET 具有竞争力的性能，但在 ROUGE-2 和 ROUGE-L 上仍落后于它们。这些差异可能是因为 BART+R3F 的计算开销高于我们的基于课程学习的模型 [5]，而 BART+MUPPET 的表现最好，可能是由于其额外的大规模学习阶段（即预微调），这一阶段未在 BART 和 CURRSUM 中考虑 [6]。有趣的是，虽然通过句子交叉注意力模块增强解码器（即 SENTSUM）略优于 BART

模型	RG-1	RG-2	RG-L
BERTSUMEXT (2019年)	20.32	4.81	13.77
BERTSUMABS (2019年)	21.92	4.99	14.21
BERTSUMEXTABS (2019年)	22.14	6.01	14.66
MATCHSUM (2020)	25.09	6.17	20.13
PEGASUS (2020a)	26.63	9.01	21.60
BART (2020)	28.80	9.02	23.02
神经主题摘要 (2021年)	27.96	9.43	23.08
BART+R3F (2021b)	**30.31**	10.98	24.74
BART+MUPPET (2021a)	**30.30**	**11.25**	**24.94**
SENTSUM（我们的方法）	29.09	9.14	23.39
CURRSUM（我们的方法）	**30.32**	10.16	24.27
CURRSENTSUM（我们的方法）	29.57	9.81	23.61

表 1：在 Reddit TIFU 数据集测试集上的 ROUGE 结果。我们将最佳数值以及与最佳数值相差不超过 0.15 的数值加粗。

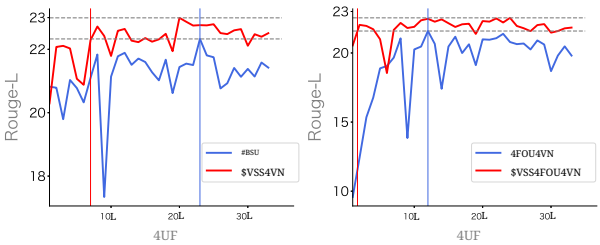


图 3：展示 Reddit TIFU 数据集上不同模型随着训练步数增加而变化的验证集性能。蓝线表示未使用课程学习时的性能，红线表示加入课程学习后的性能。竖线标示模型达到基线系统收敛时 ROUGE-L 分数所需的训练步数。

以基线系统为基础，使用课程学习 SUPERLOSS 函数进行训练（即 CURRSENTSUM）可进一步提升性能。这一发现有力地证明了将课程学习训练策略应用于摘要任务的有效性。比较抽象式模型与抽取式模型，我们观察到明显的性能差距，这表明 Reddit TIFU 中的摘要更偏向抽象式，而非抽取式。

我们进一步研究了在采用和不采用课程学习策略的情况下，摘要模型的验证集性能。如图 3所示，采用课程学习策略训练的模型（即 CURRSUM 和 CURRSENTSUM）往往收敛得更快，且与各自未采用课程学习的基线系统（即 BART 和 SENTSUM ）相比，性能更好。观察图 3 可见，考虑到得分和收敛步数（即纵向红

[3] *HTTP* [4] https://github.com/nlpyang/PreSumm [5] BART+R3F 会额外进行一次前向传播（FP），以计算对称 KL 散度项，从而衡量参数噪声。[6] BART+撰写本研究时，MUPPET 尚未发布；因此，我们无法报告应用课程学习后的性能。

模型	RG-1	RG-2	RG-L
BART (Lewis 等人，2020)	20.73	6.79	17.22
CURRSUM（我们的方法）	**22.12**	**6.91**	**17.97**

（a）Webis-TLDR-17

模型	RG-1	RG-2	RG-L
BART (Lewis 等人，2020)	44.16	21.28	40.90
CURRSUM（我们的方法）	**44.49**	**21.53**	**41.19**

（b）CNN/DM

模型	RG-1	RG-2	RG-L
BART (Lewis 等人，2020)	45.14	22.27	37.25
CURRSUM（我们的方法）	**45.41**	**22.39**	**37.41**

(c) XSum

表2：在跨领域数据集上对 BART 应用课程学习的结果。

课程学习模型的行）。

课程学习对跨领域数据集有何影响? 为了衡量采用课程学习策略的模型在跨领域数据集上的性能，我们比较了 BART的性能（作为一种基于 Transformer 的最先进摘要模型）与采用课程学习策略的 BART 变体（即我们的 CURRSUM 模型）在三个摘要数据集上的表现：Webis-TLDR-17、CNN/DM 和 XSum。表2中的结果有力地表明，课程学习对于来自不同领域（即社交媒体和新闻）的数据集是有效的，因为与 BART 基线相比，我们观察到了一致的性能提升。

人工评估。 一些研究已经认识到，广泛采用的 ROUGE 指标存在局限性，因为它偏重表层词汇相似度（Ng 和 Abrecht，2015年；Cohan 和 Goharian，2016年）。为了深入了解我们提出的模型的质量，我们对随机抽取的一组系统生成的摘要进行了人工评估。为此，我们从 Reddit TIFU 的测试集中随机抽取了 200 个样本，每个样本都包含帖子的源文本，以及经过匿名处理的 TLDR，分别来自：（1）作者撰写的参考摘要；（2）BART 基线；（3）我们的 CURRSENTSUM 模型。选择 CURRSENTSUM 的原因在于，我们希望在人工评估过程中考察句子注意力和课程学习各自的影响。为避免潜在偏见，我们随机打乱了提供给评估者的 TLDR 顺序。遵循先前的研究（Grusky 等人，2018年；Zhang 等人，2020b；Cho 等人，2021年），我们定义了三项指标：（1）**流畅度：** TLDR 是否写得流畅且易于理解？（2）**信息量：** 它是否

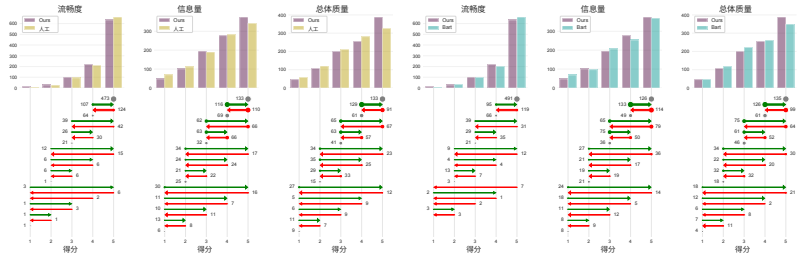
系统	流畅度	信息。	总体
BART	4.48	3.76	3.75
CURRSENTSUM	4.45	**3.91**	**3.98**
人工	**4.50**	3.72	3.70

表 3：对三个系统在流畅度、信息量和整体质量方面进行人工评估的结果。获胜得分以粗体显示。

TLDR 是否提供了有关源文本的有用信息？（3）**整体质量：** 总体而言，就内容（包括流畅度和信息量）及准确性而言，TLDR 的质量是否良好？随后，我们请五位人工评估者按照上述标准，对每个示例进行评分，评分范围为 [1-5]（从最差到最好）。这些评估者熟悉数据科学和标注，并通过 Upwork [7] 自由职业平台聘请。

表3 展示了各系统在上述定性标准上的平均得分。将我们的模型与其他系统进行比较后，我们发现：(1) 在流畅度方面，我们的模型与 BART 表现相当，而人工撰写的摘要最佳；(2) 我们计算了测试集中 TLDR 的平均长度（以词元计），结果为 22.9（人工）、23.6（BART）和 19.8（CURRSENTSUM）。有趣的是，尽管生成的摘要更短，我们的模型在信息量方面优于 BART 和人工撰写的 TLDR，这表明它能高效地从原文中筛选有用信息，并用相对更短的文本加以呈现。进一步考察信息量方面表现更优的案例后，我们注意到，如果摘要能以简洁的文本提供最重要的信息（即只提供切题的信息，符合 TLDR 的定义），标注员通常会给出更高的得分，而我们的模型通常能够做到这一点。(3) 在整体质量方面，与 BART 和人工撰写的摘要相比，我们的模型也更受青睐，且优势较为明显。总体而言，摘要模型在信息量和整体质量方面已能与人工撰写的 TLDR相媲美（有时甚至更受青睐），这印证了预训练语言模型近期取得的成功，Fabbri 等人（2021年）也展示了这一点。我们还计算了定性指标的 Fleiss' Kappa（Fleiss，1971）评定者间一致性，流畅度、信息量和整体质量的相关性得分分别为 10%、27% 和 22%。

[7] https://www.upwork.com



(a) (b) (c) (d) (e) (f)

图4：展示200条人工评分TLDR摘要之间得分变化的直方图和箭头图。(a–c)和(d–f)分别展示了我们的系统与人工及BART 基线系统的比较。尽管在流畅度指标方面与基线相当，并达到与人工相当的水平，但我们的模型在提升摘要的信息量和整体质量方面取得了显著提升。

系统	流畅度	信息量。	整体
BART	12%	26%	21%
CURRSENTSUM	**14%**	**33%**	**25%**
人工	11%	24%	20%

表 4：各系统之间的 Fleiss' Kappa 一致性

这些相关性得分分别对应流畅度、信息量和整体质量。根据 Fleiss' Kappa 的解释范围（Landis 和 Koch，1977），流畅度的相关性被视为"轻微"，信息量和整体质量则被视为"一般"。表 4 展示了各系统在三项指标上的 Fleiss' Kappa 一致性，表明对于我们系统生成的摘要，评定者之间的一致性高于其他系统。

受 MacAvaney 等人（2019年）；Sotudeh 等人（2020）的启发，我们在图 4 中绘制了人工评估的直方图和箭头图。直方图展示了各模型获得的得分分布，箭头图则展示了得分变化（即得分如何改变）在所提供样本中的情况。每个箭头的起点表示我们系统的得分，箭头指向另一系统（人工或 BART）获得的得分。箭头尾部旁标出了发生特定分数变化的样本数。如图所示，与另外两个系统相比，我们的系统在信息量和整体质量指标上取得了显著提升，同时在流畅度方面仍具竞争力。在信息量和整体质量指标上，将得分从 4 提升至 5 的效果尤为显著。

定性分析。为了深入了解我们的模型相较于基线系统的表现，我们进一步评估了标注样本。我们发现：（1）当源文本中包含大量重要信息时，我们的模型在从源文本中提取关键信息方面表现更出色，而大多数标准 TLDR 仅包含几句（少于 3 句），只涵盖最重要的信息。我们模型的这一特性使其能够在信息量和整体质量指标上占据优势。（2）在人工作者撰写的 TLDR 包含源文本中并未出现的蕴含关系或结论时，其信息量和整体质量得分相较于我们的模型生成的 TLDR 较低。（3）有趣的是，系统生成的 TLDR 越长，标注员在信息量和整体质量指标上给出的得分往往越低。这可能是因为较长的摘要包含了较高比例的源文本信息，而不论这些信息是否显著。

6 结论

尽管神经 Transformer 摘要模型展现出良好的前景，但仍存在两个缺陷，即内容选择和训练过程效率低下。本文探讨了两种解决这些问题的方法。首先，我们提出通过添加句子交叉注意力层来增强解码器，从而解决内容选择问题，使解码器能够感知句子显著性。其次，我们将置信度感知课程学习方法纳入摘要框架，希望借此提升模型的泛化能力、加快收敛，最终改善模型性能。我们在来自不同领域的多种数据集上进行的自动评估，以及人工评估，均表明了该模型的有效性。

参考文献

Armen Aghajanyan、AnchitGupta、AkshatShrivastava、Xilun Chen、Luke Zettlemoyer 和 Sonal Gupta。2021a。Muppet：通过预微调实现的大规模多任务表征。载于*2021*年自然语言处理实证方法会议论文集，第 5799–5811 页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Armen Aghajanyan、AkshatShrivastava、Anchit Gupta、Naman Goyal、Luke Zettlemoyer 和 Sonal Gupta。2021b。通过减少表征坍塌实现更好的微调。载于国际学习表征会议。

Kazuki Akiyama、Akihiro Tamura 和 Takashi Ni-nomiya。2021年。Hie-BART：利用层次化 BART 进行文档摘要。载于*2021*年计算语言学协会北美分会会议论文集：学生研究研讨会，第159–165页，线上。计算语言学协会。

Yinhan Liu an。2019年。RoBERTa：一种稳健优化的 BERT 预训练方法。*ArXiv* 预印本，abs/1907.11692。

Samy Bengio、Oriol Vinyals、Navdeep Jaitly 和 Noam Shazeer。2015年。用于循环神经网络序列预测的计划采样。载于神经信息处理系统进展*28*：神经信息处理系统年度会议，*2015*年*12*月*7*日至*12*日，加拿大魁北克省蒙特利尔，第1171–1179页。

Yoshua Bengio、Jérôme Louradour、Ronan Collobert 和 Jason Weston。2009。课程学习。载于第 *26* 届国际机器学习年会论文集（*ICML 2009*），加拿大魁北克省蒙特利尔，*2009* 年 *6* 月 *14* 日至 *18* 日，第 382 卷*ACM* 国际会议论文集系列，第 41–48 页。ACM。

Lukas Biewald。2020。使用 Weights & Biases 进行实验跟踪。软件可从 wandb.com 获取。

Isabel Cachola、Kyle Lo、Arman Cohan 和 Daniel Weld。2020。TLDR：科学文档的极限摘要。载于计算语言学协会研究成果：*EMNLP 2020*，第 4766–4777 页，线上。计算语言学协会。

Thibault Castells、Philippe Weinzaepfel 和 J érôme Revaud。2020。Superloss：一种用于鲁棒课程学习的通用损失函数。载于神经信息处理系统进展 *33*：神经信息处理系统年会论文集 *2020*，*NeurIPS 2020*，*2020* 年 *12* 月 *6* 日至 *12* 日，线上。

Asli Celikyilmaz、AntoineBosselut、Xiaodong He 和 Yejin Choi。2018年。用于抽象式摘要的深度通信智能体。收录于《计算语言学协会北美分会 *2018*年年会论文集：人类语言技术，第 *1* 卷（长篇论文），第 1662–1675 页，新奥尔良，路易斯安那州。计算语言学协会。

Ernie Chang、Hui-Syuan Yeh 和 Vera Demberg。2021年。训练样本的顺序是否重要？通过课程学习改进神经数据到文本生成。载于第*16*届计算语言学协会欧洲分会会议论文集：主会场，第727–733页，线上。计算语言学协会。

Xiaohan Chen、Yu Cheng、Shuohang Wang、Zhe Gan、Zhangyang Wang 和 Jingjing Liu。2021年。Early-BERT：利用早鸟彩票票据实现高效的 BERT 训练。载于计算语言学协会第*59*届年会暨第*11*届国际自然语言处理联合会议论文集（第*1*卷：长篇论文），第2195–2207页，线上。计算语言学协会。

Yangbin Chen、Yun Ma、Xudong Mao 和 Qing Li。2019年。用于抽象式和抽取式摘要的多任务学习。《数据科学与工程》，4：14–23。

Sangwoo Cho、Franck Dernoncourt、Tim Ganter、Trung Bui、Nedim Lipka、Walter Chang、Hailin Jin、Jonathan Brandt、Hassan Foroosh 和 Fei Liu。2021年。StreamHover：直播转录文本摘要与标注。载于*2021*年自然语言处理实证方法会议论文集，第6457–6474页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Arman Cohan 和 Nazli Goharian。2016年。重新审视科学文章的摘要评估。载于第十届语言资源与评估国际会议（*LREC'16*）论文集，第806–813页，斯洛文尼亚波尔托罗日。欧洲语言资源协会（ELRA）。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向变换器预训练。载于*2019*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（长篇与短篇论文），第4171–4186页，明尼阿波利斯，明尼苏达州。计算语言学协会。

J. Elman。1993年。神经网络中的学习与发展：从小规模开始的重要性。认知，48:71–99。

A. R. Fabbri、Wojciech Kryscinski、Bryan McCann、Richard Socher 和 Dragomir Radev。2021年。《Summeval：重新评估摘要评估》。计算语言学协会会刊，9：391–409。

Joseph L. Fleiss。1971。衡量多位评分者之间名义尺度的一致性。

Sebastian Gehrmann、Yuntian Deng 和 Alexander Rush。2018年。自底向上抽象式摘要。载于《*2018*年自然语言处理经验方法会议论文集》，第4098–4109页，比利时布鲁塞尔。计算语言学协会。

Linyuan Gong、Di He、Zhuohan Li、Tao Qin、Liwei Wang 和 Tie-Yan Liu。2019年。通过渐进式堆叠高效训练 BERT。载于《第*36*届国际机器学习会议论文集，*ICML 2019*，*2019*年*6*月*9*日至*15*日，美国加利福尼亚州长滩》，第97卷《机器学习研究论文集》，第2337–2346页。PMLR。

Max Grusky、Mor Naaman 和 Yoav Artzi。2018年。Newsroom：一个包含130万篇摘要、采用多种抽取式策略的数据集。载于《计算语言学协会北美分会*2018*年会议论文集：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。

Gholamreza Haffari、Maxim Roy 和 Anoop Sarkar。2009年。面向统计式短语机器翻译的主动学习。载于《人类语言技术：*2009*年北美计算语言学协会分会年会论文集》，第415–423页，科罗拉多州博尔德。计算语言学协会。

S. Hochreiter 和 J. Schmidhuber。1997年。长短期记忆。《神经计算》，9:1735–1780。

Wan-Ting Hsu、Chieh-Kai Lin、Ming-Ying Lee、Kerui Min、Jing Tang 和 Min Sun。2018年。使用不一致性损失的抽取式与抽象式摘要统一模型。载于《第*56*届计算语言学协会年会论文集（第*1*卷：长篇论文）》，第132–141页，澳大利亚墨尔本。计算语言学协会。

Byeongchang Kim、Hyunwoo Kim 和 Gunhee Kim。2019年。使用多层记忆网络对 Reddit 帖子进行抽象式摘要。载于*2019*年计算语言学协会北美分会人类语言技术会议论文集：第*1*卷（长篇与短篇论文），第2519–2531页，明尼苏达州明尼阿波利斯。计算语言学协会。

J. R. Landis 和 G. Koch。1977。分类数据观察者一致性的测量。《生物统计学》，33 1:159–74。

Logan Lebanoff、Kaiqiang Song 和 Fei Liu。2018年。将神经编码器-解码器框架从单文档摘要调整为多文档摘要。载于*2018*年自然语言处理实证方法会议论文集，第4131–4141页，比利时布鲁塞尔。计算语言学协会。Mike Lewis、Yinhan Liu、Naman Goyal、Mar- jan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020。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。载于计算语言学协会第*58*届年会论文集，第7871–7880页，线上。计算语言学协会。Chin-Yew Lin。2004。ROUGE：用于自动评估摘要的工具包。载于文本摘要新进展，第74–81页，西班牙巴塞罗那。计算语言学协会。Yang Liu 和 Mirella Lapata。2019。使用预训练编码器进行文本摘要。载于*2019*年自然语言处理实证方法会议暨第*9*届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第3730–3740页，中国香港。计算语言学协会。Ilya Loshchilov 和 Frank Hutter。2019。解耦权重衰减正则化。载于第*7*届国际学习表征会议（*ICLR 2019*），美国路易斯安那州新奥尔良，*2019*年*5*月*6*日至*9*日。OpenReview.net。Sean MacAvaney、Franco Maria Nardini、Raffaele Perego、Nicola Tonellotto、Nazli Goharian 和 Ophir Frieder。2020。用于开放领域答案重排序的训练课程。载于第*43*届国际*ACM SIGIR*信息检索研究与发展会议（*SIGIR 2020*），线上会议，中国，*2020*年*7*月*25*日至*30*日，第529–538页。ACM。Sean MacAvaney、Sajad Sotudeh、Arman Cohan、Nazli Goharian、Ish A. Talati 和 Ross W. Filice。2019。本体感知临床抽象式摘要。载于第*42*届国际*ACM SIGIR*信息检索研究与发展会议（*SIGIR 2019*）论文集，法国巴黎，*2019*年*7*月*21*日至*25*日，第1013–1016页。ACM。Ra mesh Nallapati、Feifei Zhai 和 Bowen Zhou。2017年。SummaRunner：一种基于循环神经网络的序列模型，用于文档抽取式摘要。载于第*31*届*AAAI*人工智能会议论文集，*2017*年*2*月*4*日至*9*日，美国加利福尼亚州旧金山，第3075–3081页。AAAI出版社。Shashi Narayan、Shay B. Cohen 和 Mirella Lapata。2018年。别给我细节，只要摘要！用于极限摘要的主题感知卷积神经网络。载于*2018*年自然语言处理实证方法会议论文集，第1797–1807页，比利时布鲁塞尔。计算语言学协会。

Shashi Narayan、Joshua Maynez、Jakub Adamek、Daniele Pighin、Blaz Bratanic 和 Ryan McDon-ald。2020。使用结构化 Transformer 进行逐步抽取式摘要与规划。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第4143–4159页，线上。计算语言学协会。

Jun-Ping Ng 和 Viktoria Abrecht。2015年。利用词嵌入改进 ROUGE 摘要评估。载于*2015*年自然语言处理实证方法会议论文集，第 1925–1930 页，葡萄牙里斯本。计算语言学协会。

Thong Nguyen、Anh Tuan Luu、Truc Lu 和 Tho Quan。2021年。丰富并控制文本摘要的全局语义。载于*2021*年自然语言处理实证方法会议论文集，第 9443–9456 页，线上及多米尼加共和国蓬塔卡纳。计算语言学协会。

Matthew E. Peters、Mark Neumann、Mohit Iyyer、Matt Gardner、Christopher Clark、Kenton Lee 和 Luke Zettlemoyer。2018年。深度上下文化词表示。载于*2018*年北美分会计算语言学协会会议论文集：人工语言技术，第 *1* 卷（长篇论文），第 2227–2237 页，美国路易斯安那州新奥尔良。计算语言学协会。

Emmanouil Antonios Platanios、Otilia Stretcu、Graham Neubig、Barnabas Poczos 和 Tom Mitchell。2019年。用于神经机器翻译的基于能力的课程学习。载于计算语言学协会北美分会*2019*年会议论文集：人类语言技术，第*1*卷（长篇与短篇论文），第1162–1172页，明尼阿波利斯，明尼苏达州。计算语言学协会。

Martin Popel 和 Ondrej Bojar。2018年。Transformer 模型的训练技巧。《布拉格数学语言学通报》，110：43–70。

Shreyas Saxena、Oncel Tuzel 和 Dennis DeCoste。2019年。数据参数：用于学习可微课程的一类新参数。载于神经信息处理系统进展*32*：神经信息处理系统*2019*年年会，*NeurIPS 2019*，*2019*年*12*月*8*日至*14*日，温哥华，不列颠哥伦比亚省，加拿大，第11093–11103页。

Abigail See、Peter J. Liu 和 Christopher D. Manning。2017年。切中要点：使用指针生成网络进行摘要。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，温哥华，加拿大。计算语言学协会。

Sajad Sotudeh、Hanieh Deilamsalehy、Franck Dernoncourt 和 Nazli Goharian。2021年。TLDR9+：用于社交媒体帖子的极限摘要大型资源。载于第三届摘要新前沿研讨会论文集，第142–151页，线上及多米尼加共和国。计算语言学协会。

Sajad Sotudeh、Nazli Goharian 和 Ross Filice。2020。关注医学本体：临床抽象式摘要的内容选择。载于 第*58*届计算语言学协会年会论文集，第1899–1905页，线上。计算语言学协会。

Yi Tay、Shuohang Wang、Anh Tuan Luu、Jie Fu、Minh C. Phan、Xingdi Yuan、Jinfeng Rao、Siu Che-ung Hui 和 Aston Zhang。2019年。用于长篇叙事文本阅读理解的简单高效课程指针生成网络。载于 第*57*届计算语言学协会年会论文集，第4922–4931页，意大利佛罗伦萨。计算语言学协会。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。2017年。注意力就是你所需要的一切。载于 《神经信息处理系统进展》第*30*卷：神经信息处理系统年会，*2017*年*12*月*4*日至*9*日，美国加利福尼亚州长滩，第5998–6008页。

Michael Völske、Martin Potthast、Shahbaz Syed 和 Benno Stein。2017年。TL;DR：挖掘 Reddit 以学习自动摘要。载于 摘要新前沿研讨会论文集，第 59–63 页，丹麦哥本哈根。计算语言学协会。

Thomas Wolf、Lysandre Debut、Victor Sanh、J ulien Chaumond、Clement Delangue、Anthony Moi、Pier- ric Cistac、Tim Rault、Remi Louf、Morgan Funtow- icz、Joe Davison、Sam Shleifer、Patrick von Platen、Clara Ma、Yacine Jernite、Julien Plu、Canwen Xu、Teven Le Scao、Sylvain Gugger、Mariama Drame、Quentin Lhoest 和 Alexander Rush。2020。Transformers：最先进水平的自然语言处理。载于*2020* 年自然语言处理实证方法会议论文集：系统演示，第 38–45 页，线上。计算语言学协会。

Wen Xiao 和 Giuseppe Carenini。2019年。结合全局和局部上下文的长文档抽取式摘要。载于*2019*年自然语言处理经验方法会议和第九届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第 3011–3021 页，香港，中国。计算语言学协会。

Benfeng Xu、Licheng Zhang、Zhendong Mao、Quan Wang、Hongtao Xie 和 Yongdong Zhang。2020a。

自然语言理解的课程学习。载于计算语言学协会第*58*届年会论文集，第6095–6104页，线上。计算语言学协会。

Jiacheng Xu、Zhe Gan、Yu Cheng 和 Jingjing Liu。2020b。具备话语感知能力的神经抽取式文本摘要。载于第 *58* 届计算语言学协会年会论文集，第 5021–5031 页，线上。计算语言学协会。

Jingqing Zhang、Yao Zhao、Mohammad Saleh 和 Peter J. Liu。2020a。PEGASUS：通过抽取间隔句进行预训练，以实现抽象式摘要。载于 *ICML*。

Yuhao Zhang、Derek Merck、Emily Tsai、Christopher D. Manning 和 Curtis Langlotz。2020b。优化摘要的事实正确性：一项关于放射学报告摘要的研究。载于第 *58* 届计算语言学协会年会论文集，第 5108–5120 页，线上。计算语言学协会。

Ming Zhong、Pengfei Liu、Yiran Chen、Danqing Wang、Xipeng Qiu 和 Xuanjing Huang。2020。将抽取式摘要视为文本匹配。载于第*58*届计算语言学协会年会论文集，第6197–6208页，线上。计算语言学协会。

Yanyan Zou、Xingxing Zhang、Wei Lu、Furu Wei 和 Ming Zhou。2020。通过重新引入源文本进行抽象式文档摘要预训练。载于*2020*年自然语言处理实证方法会议（*EMNLP*）论文集，第3646–3660页，线上。计算语言学协会。