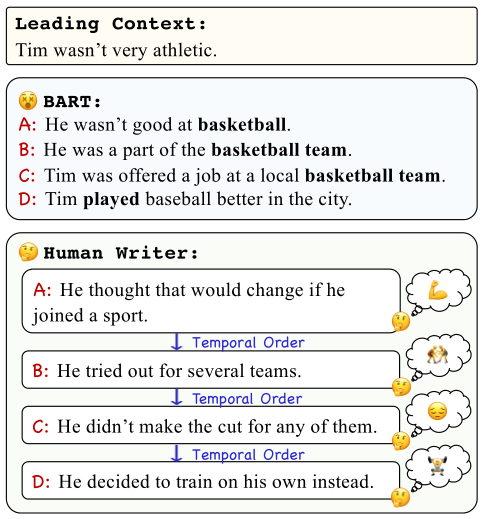


ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021.

生成长而连贯的文本是一项重要但具有挑战性的任务，特别是对于开放式语言生成任务，如**故事生成**。尽管在句子内连贯建模方面取得了成功，但现有的生成模型（如BART）仍然难以在生成的文本中保持连贯的事件序列。我们推测这是因为译码器很难在token级共现之外的上下文中捕捉高层语义和语篇结构。在本文中，我们提出了一个长文本生成模型，该模型可以在解码过程中在句子和语篇两个层面上表示前缀句。为此，我们提出了两个预训练目标，通过预测句间语义相似度和区分正常句序和无序句序来学习表征。大量实验表明，我们的模型可以生成比最先进的基线更连贯的文本。



生成连贯的长文本的能力在许多自然语言生成（NLG）应用程序中起着重要作用，特别是对于开放式语言生成任务，如故事生成，即从提示或引导上下文生成合理的故事。虽然现有的生成模型可以生成具有良好**句内**连贯性的文本，但仍然难以在**整个文本**中规划连贯的情节，即使使用强大的预训练模型BART。

如图所示，由经过微调的BART模型和一位人类作家根据ROC Stories提供的相同主要背景编写的故事示例。尽管有一些相关的概念（粗体），BART生成的故事仍然存在严重的不连贯问题。相比之下，人类作家能写出连贯的故事，因为他们充分考虑句子中的语境语义和话语关系（例如，时间顺序）。重复的情节（如句子B和C）、不相关的事件（如“打棒球更好”）和冲突的逻辑上（例如，“不擅长篮球”但“在篮球队”）。这些问题在其他NLG模型中也普遍存在。

预训练生成模型在各种NLG任务（如摘要和翻译）上表现出了最先进的性能。然而，这些任务需要在输入中提供足够的源信息以生成文本，而开放式生成任务需要从非常有限的输入信息中扩展合理的绘图。我们认为，现有的模型很少在标记级共现之外进行训练，因此它们可以很容易地生成相关的概念，但不能合理地安排它们。相比之下，人类作者总是首先完全理解语义（例如，一些关键事件，如“尝试”、“不成功”）以及在决定以下内容之前，已写句子之间的语篇关系（例如，时间顺序）。这样，即使相关概念很少，作者也能写出连贯的故事。因此，下一代在上下文中捕捉高级特征是很重要的。

本文：HINT，为长文本生成提供高级表示的生成模型。

典型的生成模型通常基于对所有前缀词的注意，通过下一个词的预测来训练从左到右的解码器。为了鼓励模型捕获高级特征，我们扩展了解码器，分别在句子和语篇层面上表示前缀信息，并在每个句子的末尾插入特殊标记。为了有效地学习这些表征，我们提出了两个预训练目标，包括：

（a）语义相似度预测，这需要使用句子级表征预测句子间的相似度，并使用强大的句子理解模型SentenceBERT作为教师模型；

（b）句子顺序区分，这需要使用语篇层面的表征来区分正常和无序的句子顺序。目的是帮助译码器捕获前缀的语义和语篇结构，这有助于在生成长文本时对长距离连贯性进行建模。我们将我们的贡献总结为两个方面：

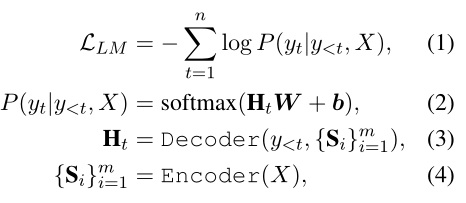
形式化定义：

输入，例如文章的开头；

输出多句子文本，有连贯性情节。

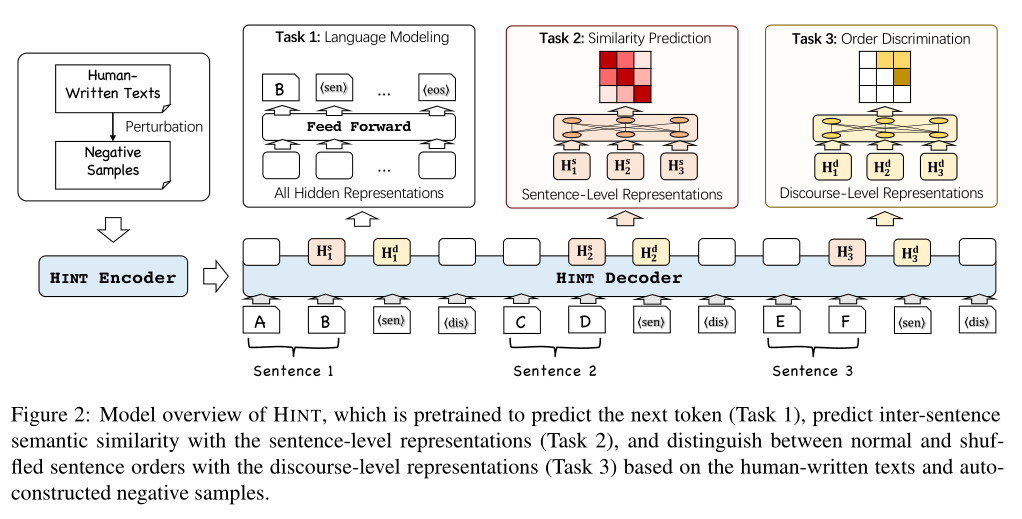
传统方法BART：

双向编码器和从左到右的解码器，最小化人类书面文本的-log概率：



H\_t是解码器的隐状态（仅关注前t-1个词），S\_i是上下文化的表示x\_i（双向）。W和b用于训练。

问题：由于解码器无法捕获前缀句子的高级特征，这些模型通常会生成不连贯的文本。



方法：我们使用**高级表示解码器**以收集前缀信息。具体来说，我们将人类书写的文本分割成连续的句子，并在每个句子的末尾添加特殊的标记，这些标记将用于在解码过程中聚合它们各自的语义以及它们之间的话语关系。

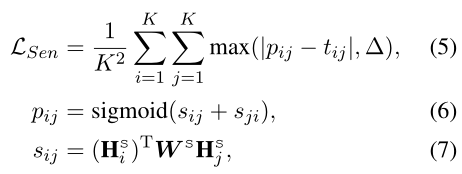
除了标准语言建模目标外，我们还设计了两个预训练任务，包括相似性预测和顺序判别，分别学习句子和语篇层面的表征。虽然在这篇文章中我们仅考虑把句子看作段（segment），但是我们的方法可以很容易地扩展到其他的句法层面，如短语或段落。

1. 句子级别的表示

在每个句子后面增加一个<sen>，用于聚合每个句子的语义。定义为解码器在第k个句子是正确句子（即原文本）的隐状态。我们期待是对输出的有意义的句子表示，也就是说“语义相似的句子在向量空间中有紧密的表示”。由于句子表征已经通过许多强大的模型（如SentenceBERT）进行了很好的语言理解研究，因此我们建议直接将其语义知识转移到我们的句子表征学习中。

具体而言，我们要求HINT解码器仅使用和预测两个句子和的相似性，使用SentenceBERT中的相似性作为实际值进行训练（SentenceBERT使用余弦相似性，在经过mean池化后的句子嵌入上执行，我们将结果线性映射到了[0,1]中）。

请注意：我们不直接学习每个句子的SentenceBERT表示，而是学习相似性分数，以避免不同模型偏差之间的差异。此外，为了减轻SentenceBERT固有bias，我们没有强制HINT完全符合真实值。相反，预测分数和黄金相似度之间的差值小于一个差值就足够了。从形式上讲，相似性预测任务的损失函数可以导出如下：

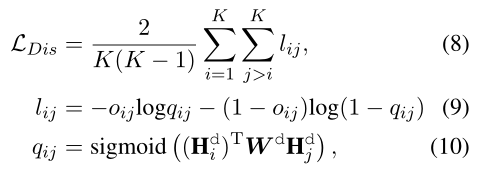


P是预测相似度，t是SentenceBERT相似度，s是中间变量（确保p的对称性），W可训练（从HINT空间到SentenceBERT空间）。该任务通过外部的监督学习，以学习句子级表示，增强HINT解码器充分理解前缀句子语义的能力。

1. 语篇级别的表示

与句子级表征学习类似，我们还在每个句子和相应的<sen>标记后插入一个特殊的语篇标记<dis>，以收集不同句子之间的语篇信息。让表示解码器的隐状态。这里的k表示在第k个话语标记是要预测的真实值。应该是一种有意义的表达，可以用来推导与他人的话语关系（例如，第k个句子在时间顺序上先于另一个句子）。

先前的研究表明，从混乱的句子中重建正确的顺序有助于理解话语关系（Lee等人，2020年）。然而，非缓冲任务并不直接适用于NLG，因为解码器应该学会在解码过程中动态建模语篇结构，而不是等到完成整个文本的解码。因此，我们建议通过区分两个句子的顺序是否正确，以成对的方式学习语篇层面的表征。



o是真实值（1表示句子i在j之前，0为之后），q是预测，W可训练。

相比于句子级别，语篇级别更关注语篇（句子和其它句子的关系），从而捕捉内容和顺序上的高层次特性。

预训练与微调：

目标：泛化HINT的性能

为了更有效地学习高级表征，我们通过从人类书面文本中自动构建否定样本来增强训练语料库，以便进行预训练。具体来说，对于顺序辨别任务，我们将人类书面文本中的句子随机洗牌作为否定样本。在相似性预测任务中，除了含有打乱句子的负样本外，我们还随机重复一个句子，或者用另一个句子替换一个句子以成为负样本。

损失函数：



LM用于人类书面文本，Dis用于人类文本和打乱句子的负样本，Sen是人类书面文本和全部负样本。预训练这两个目标可以帮助我们捕捉语义和文章结构。微调仅使用LM损失。

基本参数设置：

由于我们的方法可以适应所有具有自回归解码器的生成模型（例如GPT-2、UniLM等），因此我们使用BART作为HINT的基本框架，该框架已被证明对长文本生成具有很强的性能（GoldfarbTarrant et al.，2020）。我们还提供了文献中广泛使用的GPT-2的性能。由于计算资源有限，我们遵循BART-BASE的超参数，利用公共预训练checkpoint初始化提示。编码器和解码器的批大小设置为10，最大序列长度设置为512。∆=0.1，我们将结果与其他设置一起显示∆在附录中。和均设置为0.1。

我们基于实现了我们的模型，并使用了HuggingFace的Transformers的checkpoint。编码器和解码器都包含6个隐藏层和12个注意头。词汇表由50625个具有字节对编码的令牌组成。我们将原始词汇中的MASK和S分别视为句子标记SEN和语篇标记DIS。后期训练和微调的学习率均为3e-5，Adam为优化器。。HINT在BookCorpus上的后期培训大约需要32小时，而在ROC/WP上进行微调则需要7小时/8小时。结果基于1个NVIDIA TITAN X GPU。

预训练：

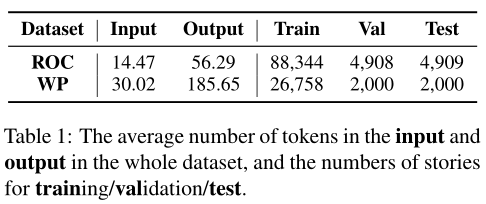
我们采用BookCorpus作为我们的预训练数据集，并使用NLTK将每个文本拆分为句子。我们以一个句子作为输入，十个句子作为目标输出，生成训练文本。此外，我们还构建了与人类书面文本相同数量的负样本（包括进行重复、替换或混洗）。我们在BookCorpus上预训练HINT为0.1M步。

BookCorpus数据集：



微调：

评估HINT在ROC（ROCStories）和WP（手写提示）上的性能。以下是数据集切分。



在ROC上使用特殊占位符屏蔽所有name来实现更好的泛化。WP最初包含303358个故事和写作提示，它们通常不受写作主题的限制。考虑到使用太多的示例进行微调可能会削弱后训练的影响，我们从WP的原始验证集和测试集中随机选择故事进行后续实验。我们将第一句话和提示作为输入，分别为ROC和WP生成文本。我们只保留WP中文本的前十个句子（使用NLTK分割）进行微调。对于ROC/WP，批次大小分别设置为10/4。其他超参数与训练前阶段相同。

Baseline：

Seq2Seq：它根据输入生成文本。为了获得更好的性能，我们通过在下游数据集上从头开始训练BART来实现基线，而无需预先训练。

Plan&Write：首先根据输入计划关键字序列；然后根据关键词生成文本（Yao等人，2019）。我们根据本文提供的代码实现了该模型。

GPT-2和BART：它们在具有语言建模目标的下游数据集上进行了微调。

BART-Post：首先在预训练数据集上进行Post训练，初始预训练目标为BART（文本填充和句子排列），步骤数与提示数相同；然后使用语言建模目标对下游数据集进行微调。

BART-MTL：该模型通过对下游数据集的多任务学习（MTL）微调BART进行训练，包括语言建模目标和辅助多标签分类目标，这需要区分人类书写文本和自动构建的负面样本。

此外，我们通过消融试验研究每个组件的影响。此外，我们还通过在下游数据集直接微调BART和HINT，以提出的两个目标作为辅助任务，展示了我们的方法对通用语言生成模型的适应性。为了公平比较，我们将所有预训练的模型都设置为基础版本。我们还将句子标记和话语标记插入到所有基线的每个训练文本中。

我们使用p=0.9的nucleus抽样生成文本，0.7的softmax temperature平衡多样性和流利性之间的权衡。如果最后一个标记是<sen>，我们设置生成<dis>的概率为1，以确保HINT可以获得每个句子的高级表示。在评估过程中，我们移除生成文本中的特殊标记。我们将这些设置应用于所有基线。

自动化评估矩阵：

PPL：困惑（Perplexity）分数越小，总体来说流畅程度越高。在句子或话语标记是真实值的位置，我们不计算概率值。

BLEU：我们使用n=1,2来评估生成文本和人类书面文本之间的重叠。

词法重复：计算所有生成文本中重复4-gram至少n次的百分比。ROC设置n=2，为WP设置n=5。

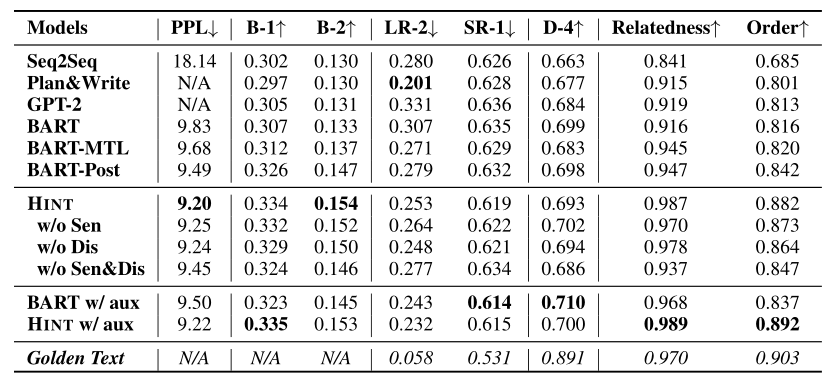
语义重复：计算平均前n个SentenceBERT相似度的平均值。ROC设置n=1，WP设置n=10。

Distinct-4（D-4）：Distinct 4-grams/All generated 4-grams，衡量世代多样性。

上下文相关性：这是一个可学习的自动度量。首先，我们使用RoBERTaBASE训练分类器，以区分人类书面文本和通过随机替换人类书面文本的单词、短语和句子构建的否定样本。然后，我们使用所有生成文本的平均分类器分数来度量上下文相关性。

句子顺序：与关联度测量类似，我们训练另一个分类器来区分人类书写的文本和句子随机乱排的否定样本。我们使用平均分类器得分来衡量句子顺序。我们根据下游数据集的训练集训练最后两个指标。

结果：



解释：

表中显示了ROC的结果。我们不提供Plan& Write和GPT-2的困惑分数，因为它们没有使用与BART相同的词汇来标记文本。Hint在易理解性方面高于所有基线，表明在测试集中对文本建模的能力更强。

HINT可以产生更多的单词与参考文本的重叠，这可以通过更好的BLEU分数来体现。这与之前的观察结果（Xu等人，2020年）一致，即Plan&Write比训练前HINT模型的词汇重复更少，可能是因为小模型更擅长学习短期统计数据（如n-gram），而不是长期依赖性。然而，与GPT-2和BART相比，HINT证明了这种情况，并且语义重复比所有基线都少，表明HINT捕捉语义特征的能力更好。

此外，我们的方法不会损害世代多样性（D4）。HINT在生成相关事件和安排适当顺序方面优于基线模型，如较高的相关性和顺序分数所示。此外，将所提出的目标作为辅助任务进行微调，可以进一步减少词汇和语义重复，并提高BART和HINT的关联性和顺序分数，这表明在句子和语篇层面建立长期连贯模型的总体好处。

此外，消融实验表明，句子层面和语篇层面的表征对于提高生成具有相关事件和合理顺序的文本的能力相对更为重要。它们都有助于减少语义冗余。当训练后仅以语言建模为目标时，几乎所有指标都会大幅下降，这表明建模高水平连贯性的重要性。

注意到有一些数据比黄金文本（Golden Text）还要好，我们总结了以下可能的原因：

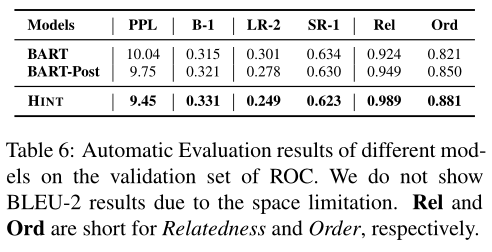
（a）在一些黄金文本中，学习分类器仍然难以判断内隐关联性，这可能需要很强的推理能力。

（b）黄金文本中存在着一些关联性差的嘈杂文本。

（c）系统倾向于生成一组有限的文本（与黄金文本相比，distinct-4低得多），带有通用图，这可能很容易获得较高的相关性分数。

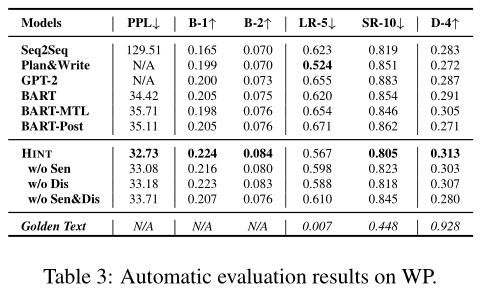
然而，我们认为，可学习性指标对于比较具有相似多样性的不同模型的上下文相关性仍然是有意义的。

附：除了在主要论文中报告的测试集上的性能外，我们还在表中提供了HINT和强基线的ROC验证集上的性能。



WP数据集的结果：

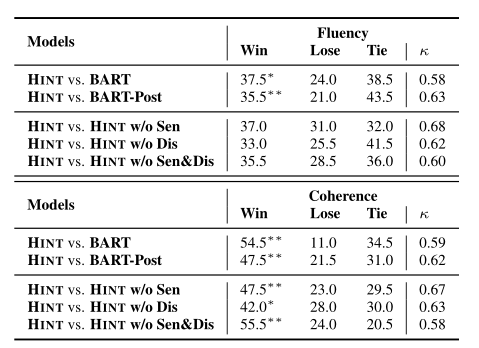
我们在表中给出了WP的结果。我们使用了大量的数据来计算词汇/语义重复，因为我们发现当生成包含数百个单词的文本时，所有模型都倾向于容易地重复相似的文本。我们没有提供相关性和顺序分数，因为很难训练出令人满意的分类器来很好地区分人类书面文本和否定样本。表中显示Hint超过了除词汇重复外表现基线，这与ROC的结果一致。因此，高级表示对于生成具有不同长度和域的长文本是有效的。



人工评估：

Pair-wise比较BART BART-Post HINT(及三个消融版本)。

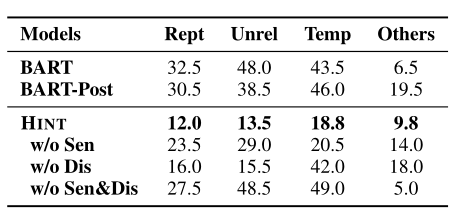
随机从ROC测试集抽样200个文本，因此数据量为1200个。



表格的K表示Flesiss’ Kappa指数，说明一致性（以前论文提到过）。\*和\*\*是显著水平<0.05和<0.01。所有结果都显示出中度的注释间一致性（0.46<κ<60.6）或实质一致性（0.6<κ<60.8）。我们可以看到，Hint通过捕捉高级特征，在连贯性方面明显优于基线，并且与基线具有相当的流利性。

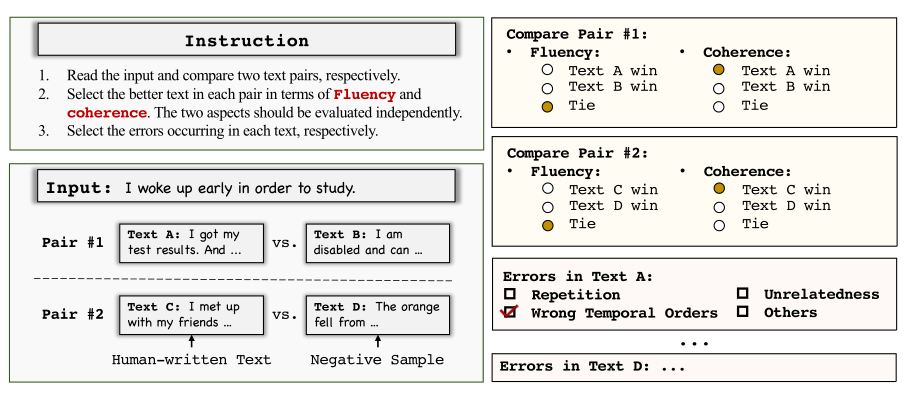
对于每一对文本（一个由我们的模型，另一个由基线以及输入），三名注释员分别在流利性和连贯性方面给出偏好（lose、win、tie）。我们采用多数票表决，在三位注释者之间作出最后决定。我们求助于Amazon Mechanical Turk（AMT）进行注释，将模糊定义为句内语言质量和语法正确性的衡量标准，而连贯是句间关联性、因果关系和时间依赖性的衡量标准。请注意，这两个方面是独立评估的。此外，我们通过过滤掉注释者在比较人类书写文本和否定样本时无法做出合理判断的注释来控制注释质量。此外，我们还要求工作人员对生成的文本中的特定错误进行注释。

以下，我们给出了注释说明和不同模型的错误分析。我们总结了表中所有注释文本（每个模型200个）中有错误的文本的百分比。当三个注释器中至少有两个为文本注释错误时，我们确定文本包含一些错误。请注意，与其他五个模型相比，HINT的每个文本都注释了五次（每次三个注释者）。因此，我们取五个注释结果的平均值。我们可以看到，提示比基线具有更少的重复性、更好的上下文关联性和时间顺序。然而，研究结果表明，生成连贯的长文本仍然具有挑战性。



所有注释文本中注释有错误的文本百分比（%）。错误类型包括重复（Rept）、未延迟（Unrel）、错误时序（Temp）等。每行中的百分比总和不是100%，因为每个文本可能包含多个错误。每种错误类型的最佳性能都在左侧突出显示。

附：注释界面

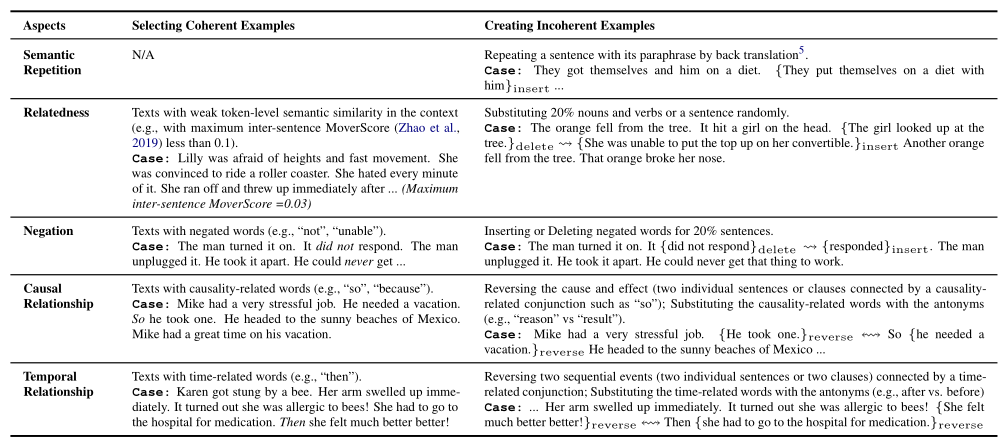


我们在图中显示了手动注释界面。在AMT的每个HIT（人类智能任务）中，我们向工作人员展示一个输入和两个文本对，包括（a）一对生成的文本（一个由Hint生成，另一个由基线生成），和（b）一对人类书写的文本和通过干扰文本（例如，重复、替换）构建的否定样本从数据中随机抽样。请注意，这两对是以随机顺序显示的。然后，我们要求工作人员分别从流利性和连贯性两个方面选择更好的文本。此外，我们还要求工作人员对每个文本中的错误进行注释，包括重复（重复相同或相似的单词）、不相关（输入或其自身上下文中的无关实体或事件）、错误的时间顺序等。如果工作人员不认为人类书写的文本比阴性样本具有更好的连贯性，或者工作人员不为阴性样本注释任何错误，则我们拒绝接受HIT。这样，我们拒绝了21.09%。最后，我们确保每对生成的文本都有三个有效且独立的比较结果。

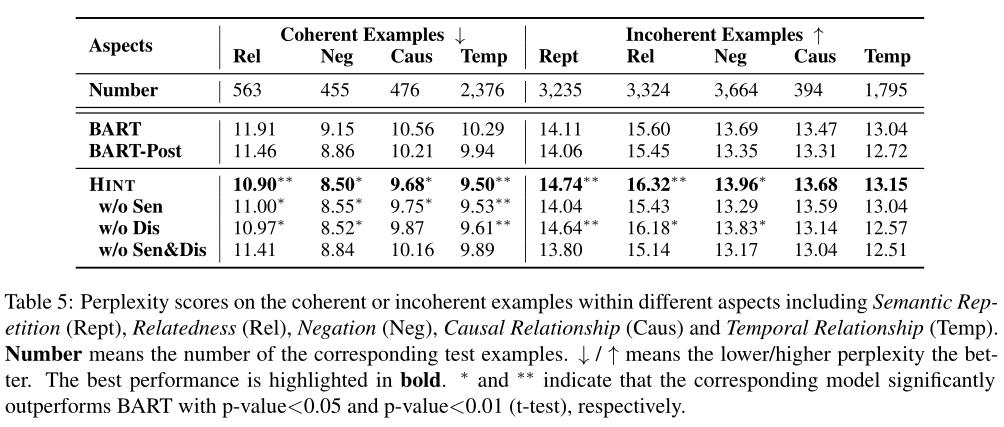
语言建模：

还需要进一步研究学习到的表征是否有助于HINT更好地捕获实现高水平的连贯性。因此，我们在不同方面使用单独的语言建模测试来评估模型。为此，我们基于ROC测试集构造了连贯和非连贯的例子，并对不同方面的例子进行了困惑度计算。具体来说，我们关注以下几个方面：语义重复、关联、否定、因果关系和时间关系。我们选择人类书写的文本作为连贯的例子，通过干扰人类书写的文本来构建非连贯的例子。例如，我们选择那些带有时间相关词（例如，“then”）的文本作为连贯的示例，用于测试时间关系。我们交换两个由人类书面文本的“then”连接的连续事件，或者将“before”替换为“after”，作为相位的不连贯示例。

注：自动构词法不可避免地会带来意想不到的语法错误，这也会影响语篇的连贯性。为了缓解这个问题，我们在CoLA语料库上训练了一个二元分类器，学习判断语法性，然后过滤掉那些被归类为非语法的例子（分类器得分小于0.5）。为简单起见，我们直接使用TextAttack的公共模型作为分类器，其精度达到82。在可乐的测试套件中占90%。最后，我们过滤掉大约15.51%的测试示例。



我们在下表中给出了结果。HINT可以在上述方面比基线模型更好地模拟上下文连贯性（连贯示例的困惑度更低），并且更有效地识别非连贯错误（非连贯示例的困惑度更高）。相比之下，与BART相比，BART Post和HINT（w/o Sen&Dis）即使在负面示例上也实现了总体上的困惑度下降，这表明他们可能仍然专注于捕获令牌级特征。在消融研究中，我们可以看到句子层面的表征提高了HINT捕捉关联性、否定性和语义重复的能力，而语篇层面的表征主要作用于因果关系和时间关系。然而，我们也注意到，与BART相比，提示在识别不合理的因果关系和时间关系方面没有显著的改进，这可能需要在学习句子顺序的同时注入显性推理知识。



案例分析：

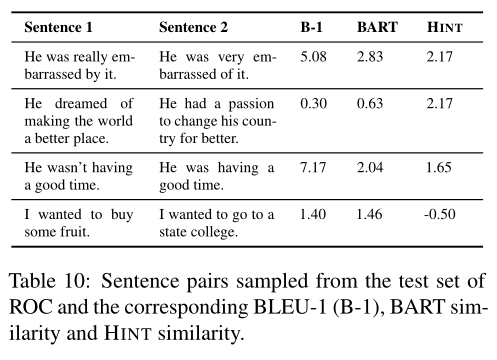
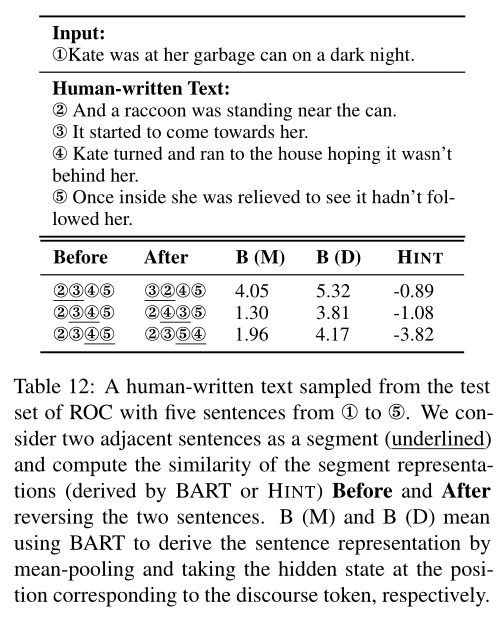


表10展示了ROC测试集的一些案例，以证明学习到的提示句子级表征的有效性。我们计算不同句子对的BLEU-1、BART相似度和HINT相似性，其中BART/HINT相似度是指两个句子的BART/HINT表示之间的余弦距离。为了获得句子的BART表示，我们将其输入BART解码器（及其上下文），并在最后一层的隐藏状态上应用均值池。HINT表示指解码句子后相应的句子级表示。我们将所有结果标准化为标准高斯分布。我们可以看到，HINT可以推导出有意义的句子级表示，并对语义相似的句子对（前两对）给出高分，但不同配对（最后两对）的得分较低。相比之下，BART更多地关注标记级的相似性，从而得出与BLEU一致的相似性。

语篇相似性：

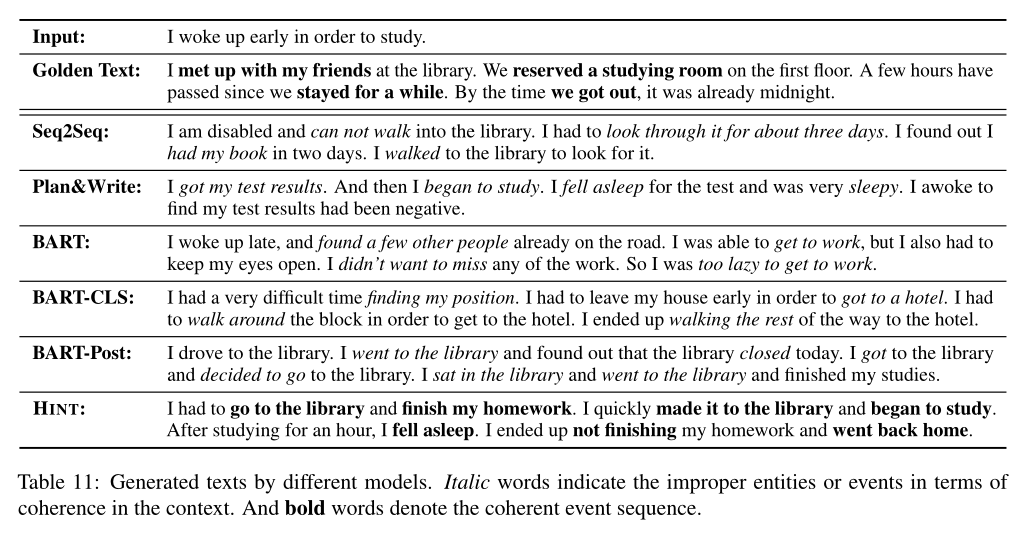
我们还在表12中展示了一个案例，以说明学习到的语篇层面暗示表征的有效性。我们在表12的文本中考虑了一个段，它由两个相邻的句子组成（例如，34在2345之中）。然后，我们可以通过连接两个句子的上下文化表示来导出段表示。此外，如果我们将两个句子（34替换为43中的其他句子不变）颠倒过来。注意，在这种情况下，我们仍然按照正常顺序连接两个句子表示（即，首先是3的表示，然后是4）。如果句子表示包含语篇层面的信息，我们期望复归前后的段表示在向量空间中是遥远的。否则，由于复归前后段具有相同的标记，因此段表示将是相似的。



对于BART，我们推导了句子复述通过将整个文本输入BART，意思是将隐藏状态汇集到句子中标记的位置。对于HUGO，我们将每个句子对应的语篇级表示作为句子表示。我们还使用BART在B(D)位置的隐藏状态显示结果，以供参考。表12显示了句子回复前后的段表示之间的相似性。所有结果均标准化为标准高斯分布。结果表明，无论是使用平均池还是隐藏状态，BART都为回复前后的段推导出了相似的表示对比而言，尽管复归没有改变，但是Hint推导的句子语义、片段表示非常不同，这表明HINT可以推导出有意义的语篇级表示。

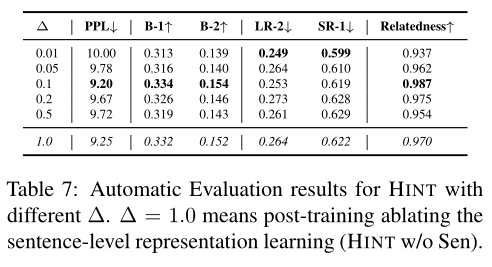
文本生成

我们在下表中展示了一些生成的案例。HINT可以生成比基线更连贯的故事。具体来说，基线可以很容易地预测一些与输入相关的单词（例如，“sleepy”、“library”）或其自身上下文（例如，“test results”、“hotel”）。然而，这些词的使用是不连贯的。例如，Plan&Write生成的文本在句子中的时间顺序错误（首先是“got test results”，然后是“fell asleep for the test”）。Seq2Seq、BART和BART-CLS生成的文本在语义和语篇结构上都是混乱的。BART Post生成的文本存在重复情节（“went to the library”）和冲突逻辑（“sat in the libaray”vs “the library closed”）。相比之下，Hint生成的文本是一个连贯的事件序列，具有相关的内容和适当的时间顺序。结果表明，在长文本生成中建立高水平连贯模型是有效的。



∆的选择:

∆用于平衡Hint预测的相似性得分与SentenceBert预测的相似性得分之间的差值的影响。我们给出了一些不同类型的自动评估结果∆。在下表中，注意到我们将使用∆ = 0.1主要论文中的实验。



可以看到，更小∆（例如，0.01）会导致较少的词汇和语义重复，但更糟糕的流利性（表现为更高的困惑度）和上下文关联性，这可能是由于教师模型的过度拟合偏差造成的。另一方面，一个更大的∆（例如0.5）将导致几乎所有指标的性能都比∆ = 1.0（没有相似性预测任务）差。结果表明，一个大的∆使模型不能有效地从教师模型中学习，并影响模型本身的表征。相比之下∆ = 0.1将带来更好的整体性能。