文章编号：1003-0077(2023)08-0150-09

基于片段预测的词汇约束文本生成

聂锦燃[1,2]，杨麟儿[1,2]，杨尔弘[1]

(1. 北京语言大学 国家语言资源监测与研究平面媒体中心，北京 100083；
2. 北京语言大学 信息科学学院，北京 100083)

摘　要：词汇约束文本生成是自然语言处理领域的重要研究任务之一，旨在给定一组有序词汇，生成包含这些词汇的流畅文本，在语言教学、文本生成、信息检索等领域有广泛应用。现有的生成方法存在生成速度慢、无法包含所有约束词等问题，难以满足实际应用需求。该文提出一种基于片段预测的端到端词汇受限文本生成方法，将词汇约束文本生成视为对约束词之间的文本片段的预测，利用基于二维位置编码的预训练语言模型预测所有片段，再将其填充回约束词的对应位置，从而保证了生成速度和词汇约束；利用词性标注方式构造多参考数据进行数据增强，进一步提升了文本生成质量。为验证方法的有效性，该文在公开的英文数据集，以及基于国际中文教材构建的中文数据集上进行了实验，结果表明，该文提出的 LCTG-SP 方法可以满足所有词汇约束、具有较快生成速度，生成文本的流利度和多样性表现更好。本文中的模型代码和数据开源在 GitHub 上①。

关键词：词汇约束；片段预测；文本生成；数据增强

中图分类号：TP391　　**文献标识码**：A

Lexically Constrained Text Generation Based on Segments Prediction

NIE Jinran[1,2], YANG Lin'er[1,2], YANG Erhong[1]

(1. National Language Resources Monitoring and Research Center for Print Media,
Beijing Language and Culture University, Beijing 100083, China;
2.School of Information Science, Beijing Language and Culture University, Beijing 100083, China)

Abstract: Lexically constrained text generation aims to generate fluent text containing these words given a set of ordered words, which is widely used in language teaching, text generation, information retrieval, and other fields. This paper proposes an end-to-end lexically constrained text generation method based on fragment prediction, which considers the lexically constrained text generation task as an end-to-end prediction of text fragments between constrained words. It uses two-dimensional position encoding to learn semantic relationships between segments and within segments, thereby speeding up text generation while ensuring generation quality and lexical constraints. In addition, the part-of-speech tagging method is used to construct multi-reference data for data augmentation. Experiments are conducted on the English dataset publicly available and a Chinese dataset of international Chinese textbooks constructed by this paper. The experimental results show that the method proposed in this paper has significantly improved generation speed, fluency, and diversity(code and data available at https://github.com/blcuicall/LCTG-SP).

Keywords: lexical constraints; segment prediction; text generation; data augmentation

0　引言

自然语言生成是自然语言处理的一个重要领域，实现文本的自动生成也是人工智能走向成熟的一个重要标志。为了适应不同场景的需求，文本生成通常是受约束的，可控文本生成技术便是解决约束条件下的文本生成。可控文本生成也称为受控文本生成，根据控制条件和目标的不同可以分成多种任务。从控制条件来看，分为属性(风格)控制[1]、话

收稿日期：2022-03-01　　**定稿日期**：2022-03-26

① https://github.com/blcuicall/LCTG-SP

题控制[2]、词汇控制[3]、格式控制[4-5]、结构化数据控制[6]等。

本文关注的是词汇控制，称之为词汇约束文本生成[3,7-8]。词汇约束文本生成是指在文本生成过程中，控制某些词汇必须出现在生成的文本中。这一任务的输入是给定的一个或多个词汇，输出是包含这些给定约束词汇的通顺文本。

词汇约束文本生成的应用非常广泛，可应用于机器翻译[9]、信息检索[10]、文本生成[3]等自然语言处理任务。在机器翻译中，可以将已有翻译知识库中专业词汇作为机器翻译过程中受约束的词汇，使得对于专业术语的翻译更准确。查询重写[11-12]是信息检索领域中非常重要的研究任务，而查询重写过程中往往需要约束用户输入文本中的关键词仍然包含在重写之后的文本中。在文本生成领域，故事生成[13]通常要求按照某些关键词生成，这些词汇需要出现在生成的故事中；会议记录自动生成旨在由多个关键词生成完整的会议记录；广告语生成则是要求品牌名或产品名必须出现在生成的广告语中。这些都是词汇受限文本生成所适用的场景。

除此之外，词汇约束文本生成在语言学习和教学领域也有着广阔的应用场景。一方面，可以利用词汇约束文本生成帮助语言学习者进行词汇的学习。例如，词汇运用能力不足的学习者想表达某句话时在脑海中浮现几个词，但难以用其表达成一句话，词汇约束文本生成则可以帮助其学习这些词汇的使用。另一方面，对于语言教学者来说，词汇教学无疑是重要且基础的一环，教学中也会设计连词成句的题型来加强学生对于词汇的理解与运用[14]。可以通过词汇约束文本生成辅助教师例句编写。

词汇约束文本生成的方法可以分为三类：①改进的柱搜索[15]；②随机搜索[3,8]；③直接端到端的生成[7]。改进的柱搜索是一类解码方法，在模型的解码阶段即插即用，通过控制柱搜索过程，实现约束词出现在生成的文本中。这类方法在源端和目标端信息对等的机器翻译任务上表现良好，但在搜索空间较大的文本生成任务中需要花费大量时间搜索候选句，生成质量也较差[16]。随机搜索的方法将约束词顺序拼接作为初始状态的文本，利用随机采样的替换、插入和删除操作不断迭代修改当前状态，直到得到满足要求的文本。随机采样会造成很多重复和冗余的操作[7]，生成过程需要消耗大量时间，难以在实际应用中使用。端到端的方法以深度神经网络作为基本模型，将多个约束词拼接起来直接作为模型输入，训练模型生成完整的句子。端到端的生成速度较快，但是无法保证所有约束词都包含在生成的句子中，其原因在于神经网络的生成过程是基于网络参数所拟合的概率分布的，解码时通过每一步的概率分布来预测下一个词，因此无法保证一定生成出输入的约束词[7]。

为了解决上述问题，本文提出片段预测的端到端方法用于词汇约束文本生成，称之为 LCTG-SP (Lexically Constrained Text Generation Based on Segments Prediction)方法。将该任务视为预测约束词之间的片段，完成片段预测后将约束词填充到对应位置，从而保证了生成的文本百分之百包含所有约束词，同时端到端生成片段也具有较快的生成速度。为了有效预测所有片段，我们利用二维位置编码来捕获片段之间和片段内部的位置关系，从而更好地学习到片段的语义。我们构建了基于对外汉语教材的数据集作为词汇约束文本生成的基础资源，并在该数据集和已经公开的英文数据集上进行了实验。实验结果表明，本文提出的 LCTG-SP 方法在保证百分百包含约束词的同时，有效提升了生成速度和质量。本文贡献如下：

(1) 提出了基于片段预测的词汇约束文本生成方法 LCTG-SP，解决了主流方法存在的生成速度慢或无法百分百包含约束词的问题；

(2) 提出了多参考数据增强的方法来提升文本生成质量；

(3) 构建了用于词汇约束文本生成的中文数据集。在中英文数据集上进行了实验，实验结果表明，本文所提出的方法在生成速度和生成质量方面都具有较大优势。

1　相关工作

1.1　改进的柱搜索方法

改进的柱搜索方法是一类解码方法，在模型解码阶段即插即用，通过控制柱搜索过程，使约束词包含在生成的文本中。最典型的工作是 Hokamp 等人提出的网格柱搜索(Grid Beam Search，GBS)，该方法增加包含约束词的候选搜索维度，使柱搜索过程中保留包含约束词的候选。动态柱分配(Dynamic

Beam Allocation，DBA）方法是网格柱搜索的扩展[15]，它将满足相同数量约束的可选对象进行分组，动态分配候选数量，以加速推理过程。我们的方法不是只在解码阶段使用，而是在训练阶段就考虑了约束词，保持了训练和预测的一致性，具有更快的生成速度和更好的生成质量。

1.2 随机搜索的方法

随机搜索的方法将约束词顺序拼接作为初始状态的句子，随机采样替换、插入和删除操作不断迭代修改当前状态，直到得到满足要求的句子。这一类方法有一系列工作，例如，Berglund 等人首先使用吉布斯采样从句子空间直接生成句子[17]。Wang 等人扩展了吉布斯采样方法[18]，从预训练模型 BERT[19] 生成文本。Miao 等人提出了 CGMH 方法[20]，可以在给定的词汇约束下通过替换、插入和删除操作生成文本。

随机采样的方法的缺点是可能会造成较多重复和冗余的操作，导致生成过程缓慢。例如，会将某个词插入和删除多次。因此 He 等人提出 X-MCMC-C 方法[3]，利用预先训练的分类器来预测需要在句子的哪个位置做何种操作，由此来避免一些无效的重复操作。其模型训练和预测需要分成多步进行，首先训练分类器和语言模型，分类器用来判断操作和操作执行的位置，语言模型用来计算迭代修改过程中的状态转移接受率。同样，为了缓解随机搜索过程中出现的重复和冗余操作问题，Sha 等人[16]提出了一个可微的目标函数，并利用梯度来帮助确定序列中哪个位置的词被改变。这些方法有效缓解了迭代修改过程中重复和冗余操作的问题，但是每一步的迭代修改只能对一个词进行，因此为了提升搜索的速度，Zhang 等人[21]提出一种类似于非自回归的并行预测方式，允许同时在当前状态的句子中每两个词之间最多插入一个词，使得每步的搜索迭代能同时插入多个词，从而提升搜索速度。

这些改进都是基于迭代搜索的，仍然都是插入式方法，需要将给定的关键词作为初始状态的文本，不断进行多次增、删、改迭代操作，训练和生成的过程依然繁琐，非常耗时，难以在实际应用中使用。我们的方法则是端到端生成文本，无须多次迭代，生成速度比随机搜索方法快很多。

1.3 端到端的方法

端到端的方法以深度神经网络作为基本模型，将多个约束词拼接起来直接作为模型输入，训练模型生成完整的文本，其生成速度比随机搜索和改进的柱搜索方法都要快很多。但是基于深度神经网络的端到端生成方法无法保证输入的约束词都包含在生成的句子中，其原因在于神经网络的生成过程是基于网络参数所拟合的概率分布，解码时利用每一步的概率分布来预测下一个词，因此无法保证一定生成出给定的约束词。为了缓解这一问题，Wang 等人[7]提出在注意力机制中引入约束词的标记，即标记当前候选的输出是否为约束词，如果约束词出现则修改对应约束词的标记，以此引导输入中尽可能包含更多的约束词，从而提升约束词出现的比例。Qin 等人提出基于能量的郎之万动力学约束解码[22]，通过基于梯度的采样对约束进行有效的可微分推理。该方法可直接用于从左到右的端到端模型，提升词汇约束满足的比例。这些方法在端到端的训练和预测中同时考虑了词的约束，但仍然无法保证百分百包含所有约束词。本文提出的基于片段预测的方法可以在端到端生成的同时保证包含所有约束词。

2 方法

2.1 任务定义

词汇约束文本生成任务要求给定一组约束词，生成一个文本包含所有约束词。假设给定 $c_1,c_2,\cdots,c_k$ 这 k 个约束词，则该任务所建模的公式如式(1)所示。

$$X^* = \arg\max_X P(X \mid c_1,c_2,\cdots,c_k) \quad (1)$$

其中，X 是包含了所有约束词的通顺文本。

2.2 片段预测

我们将词汇约束文本生成任务视为预测约束词之间的片段，即约束词组成的序列是一段不完整的文本，模型需要预测每两个约束词之间空缺的片段，将预测结果按片段对应位置填充到约束词组成的模板中，得到完整的包含所有约束词的通顺的句子。图 1 展示了本文提出的基于片段预测的词汇约束文本生成的基本思想。假设约束词为“喜欢”和“篮球”，期望模型生成完整的包含这两个约束词的文本，例如，“我喜欢在公园打篮球。”就是一句符合生成要求的文本。片段预测的思想是不预测约束词本身，而是预测约束词之间的片段，在这个例子中，需要预测“我”“在 公园 打”“。”三个片段。

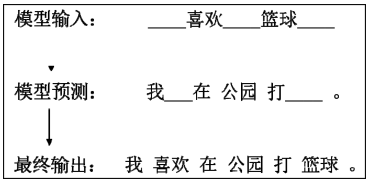

图 1　基于片段预测的词汇约束文本生成

基于片段预测的词汇约束文本生成方法在形式上与预训练语言模型的遮蔽式的训练目标有相似之处，但是前者的挑战性和难度更大。具体来说，预训练语言模型的遮蔽式训练目标就是文本填空任务，是给定一个不完整的文本，即文本中存在一些词语或片段空缺，要求预测这些空缺位置的词语或片段。预训练语言模型的文本填空任务通常是随机遮蔽文本中 15％的词语或片段[19]，而在词汇约束文本生成中需要预测部分远远多于已知部分，被遮蔽的词语或片段达到 80％甚至更高。由于给定的约束词很少，需要预测的词语或片段很多，因此挑战性较大，无法直接使用现有的预训练语言模型来预测词汇约束文本生成中的片段。

2.3　模型

本文设计了适用于词汇约束文本生成的片段预测（LCTG-SP）方法，同时设计了可顺序预测多个片段的模型架构，以便更好地学习预测片段与约束词的语义关系，并且构建了相应的训练数据。

GLM 预训练语言模型[22]的训练目标是多任务的，主要考虑两个目标：一个是只有单个的片段覆盖 50％的 Tokens，另一个是多个片段整体覆盖 15％的 Tokens。我们的训练基于词汇约束文本生成任务构造的数据，遮蔽的比例取决于原始句长和所选取的关键词的数量。我们使用 GPT-2 模型，利用所构造的数据对其进行训练，并借鉴了 GLM 的二维位置编码方式。与 GLM 模型不同的是，我们按照约束词位置进行遮蔽，且不打乱片段的顺序。下面详细介绍具体的模型结构和数据构造。

图 2 是本文生成模型结构图，采用的是 GPT-2 的模型架构。假设模型的输入源自给定的约束词“喜欢”和“篮球”（约束词可以有任意多个），在约束词之间添加［MASK］标签，期望模型预测出［MASK］位置的片段。为了更好地学习片段和约束词的语义关系，我们引入二维位置编码。位置编码 1 用于编码片段间的位置关系，每个［MASK］位置对应片段的 Token 编码相同，例如第一个［MASK］位置对应的 Token 是“［SEP］我”，则二者在位置编码 1 中均为“1”。位置编码 2 用于编码片段内部的位置关系，位置“0”代表带有［MASK］的输入序列，其他位置则为需要被预测的序列。被预测的序列是由多个片段组成的，每个片段单独编码位置，例如第二个片段是“［SEP］在-公园-打”，该片段有 4 个 Token，从位置“1”开始编码，［SEP］位置为 1，直到最后一个 Token“打”的位置为“4”。模型训练和预测是按 Teacher-forcing 的方式，输入真实的 Token 来预测下一个 Token，而预测阶段则是由上一个预测结果来预测下一个 Token。

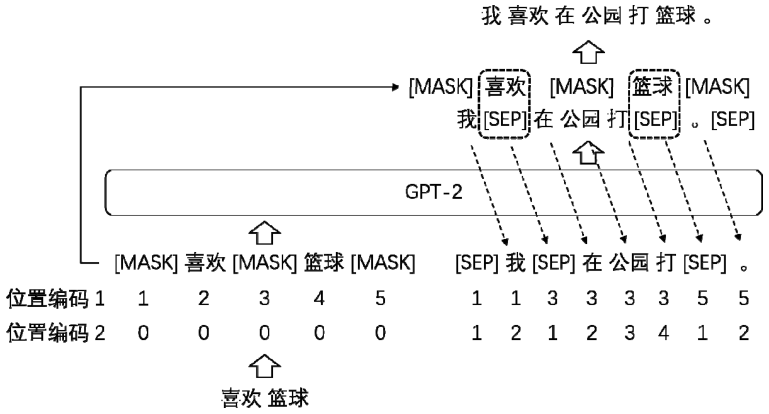

图 2　模型结构

模型建模过程描述如下：训练数据的构造是从输入文本 $X=[x_1,x_2,\cdots,x_n]$ 中通过一定策略选定约束词，截取约束词之间的多个文本的片段 $\{s_1,s_2,\cdots,s_m\}$，其中每一个片段 s_i 是 X 中的一串连续的 Token，在片段间添加［SEP］标签作为约束词位置的标记，得到模型输出序列 $S=\{s_1,[\text{SEP}],s_2,[\text{SEP}],\cdots,s_m,[\text{SEP}]\}$。在原文本中将每一个片段用单个的［MASK］代替，从而得到输入序列 X_c。模

型利用自回归的方式预测 X_c 中被遮蔽的多个片段。则需要建模从输入序列 X_c 到输出的文本片段的映射关系，即建模条件概率，如式(2)所示。

$$S=\arg\max_{S} P(S \mid X_c) \tag{2}$$

多个片段是从左到右进行自回归预测的，每个片段内部的多个 Token 也是从左到右进行自回归预测的。即预测当前片段中的某一个 Token 时，模型是基于之前所有片段和当前片段中的当前 Token 之前的所有 Token。则上述建模公式的训练目标，如式(3)所示。

$$J=\prod_{j=1}^{l_i} p_\theta(s_{i,j} \mid X_c, s_{<i}, s_{i,<j}) \tag{3}$$

其中，l_i 表示第 i 个片段 s_i 的长度，$s_{<i}$ 表示第 i 个片段之前所有的片段，$s_{i,<j}$ 表示第 i 个片段中第 j 个 Token 之前的所有 Token。

2.4 数据构造方法

图 3 是数据构造方法的具体过程，以一条数据的构建过程为例。首先利用关键词提取工具从句子中提取出关键词作为约束词。约束词的数量根据句长变化，并设定最大值和最小值，并将句子中的约束词之间的片段替换为[MASK]，然后在每个片段后加入[SEP]标签作为模型需要预测的输出。最后将约束词按顺序填入[SEP]位置(除去最后一个[SEP])即可得到包含所有约束词的完整的句子。

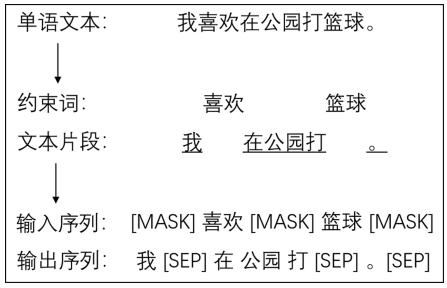

图 3 训练数据构造流程

前人工作在英文数据集 One-billion-words① 上进行实验[3]，为了推动该任务在中文上的应用，我们利用对外汉语教材构建了用于词汇约束文本生成任务的数据集。对 500 本对外汉语教材进行语料清洗和分句，按照上述形式构造数据集。

2.5 数据增强方法

前人工作是使用关键词提取工具来得到约束词，而通常来说约束词可以是任意词，从实际应用角度出发，约束词为实词即可。因此，为了尽可能充分地使用已有的单语数据，本文提出使用词性标注的方式提取约束词，并构造多参考来进行数据增强。

我们利用关键词提取工具和词性标注结果来增强获取约束词。首先将文本进行词性标注，筛选出其中的动词、名词(除专有名词)、形容词、副词等具有实际意义的词，然后将其与关键词提取工具得到的词一同作为约束词集合，并进行去重。将约束词按其在文本中的顺序进行排列，并根据文本的长度决定所选取的约束词数量。设置文本长度与约束词比例为 10∶1，由此确定约束词的数量。依据前人工作[3]，设置约束词数量最少为 1 个，最多为 4 个。

从约束词集合中多次随机采样确定约束的数量词，对于同一组约束词，利用检索方式在训练集中找出包含所有约束词的多个文本作为多参考，构成一对多的平行语料。下面以图 3 的例子说明数据增强方法。对于文本“我喜欢在公园打篮球”，利用关键词提取和词性标注工具可以得到约束词集合“喜欢，公园，打，篮球”，将约束词两两组对并到已有数据中检索，可以得到一对多的平行数据：“喜欢 篮球→我喜欢在公园打篮球。”和“喜欢 篮球→我非常喜欢篮球明星科比。”“公园 篮球→我喜欢在公园打篮球。”和“公园 篮球→在公园里有一个很大的篮球场。”

表 1 给出了经过数据增强方法构造的数据集规模，相比原有训练数据，中文和英文数据集都增加了一倍的训练数据。其中，中文数据的训练集在 210K 句基础上增加了 281K 句，英文数据的训练集在 1 000K句的基础上增加了 1 100K 句。

表 1 数据集规模 (单位：句)

数据集	训练集	验证集	测试集
中文	210K	10K	2K
＋数据增强	＋281K	—	—
英文	1 000K	50K	10K
＋数据增强	＋1 100K	—	—

我们通过消融实验，验证了上述数据增强方法的有效性。

3 实验

本文在上述中英文数据集上进行实验，从生成

① https://www.statmt.org/lm-benchmark/

质量、包含约束词的比例和生成速度三方面与基线模型进行对比。生成质量的评测包括自动评测和人工评测。

3.1　基线方法

此实验与前人工作的三类方法(包括四个模型)进行对比：X-MCMC-C[3]，DBA[15]，Transformer[7]和 T5[7](中文上为 mT5)。X-MCMC-C 是随机搜索方法中较为先进的方法，DBA 是增强的柱搜索中的典型方法，Transformer 和 T5 则是端到端方法中的常用模型。我们在测试集上对比生成文本的质量、包含约束词的比例、生成速度三方面。

3.2　实验设置

对于 X-MCMC-C 方法，我们使用原论文中开源的参数配置①，并进行了调优取最佳结果。DBA 方法是一类解码算法，我们选用典型的 Tansformer 模型，并在其之上使用 DBA 方法，以达到该方法的最佳性能。由于 T5 模型只针对英文，因此中文上选用 mT5 模型。Tansformer 模型则使用原始论文中的经典架构。模型训练时经过调优后的学习率为 3e-4，beam-size 为 5。基于端到端的模型训练轮次为 20 次，损失函数均收敛到稳定状态。

3.3　评测指标

我们利用自动评测和人工评测的方法分别对所有基线模型和本文提出的模型的生成结果进行评价，对生成的文本进行自动评测，其指标分为三方面：生成的文本与参考句之间的 n-gram 重合度，生成的文本中 n-gram 的多样性，以及生成的文本的流利度。生成的文本与参考句之间的 n-gram 重合度使用 BLEU 指标进行评测。文本的 n-gram 多样性则是利用 Distinct 指标来评估生成文本的词汇多样性的，我们选择该指标的 2-gram 进行评测，称之为 Dist-2。其原理是统计生成文本中不重复的词汇个数，并将其与总词数相比较。举个例子，如果一篇文本中 2-gram 总数量为1 000 个，不重复的 2-gram 为 500 个，那么这篇文本的 Dist-2 值为 0.5。Dist-2 值越高，说明生成文本中使用的词汇越丰富，文本的多样性越好。流利度可以通过困惑度(Perplexity，PPL)进行衡量[3]，使用预训练的 GPT-2 模型计算 PPL，PPL 越低则文本越流利通顺。此外，还需要评测约束词包含在生成文本中的比例，称之为 Const 指标。

3.4　实验结果与分析

我们的实验在中英文两个数据集上进行，利用训练集训练我们的模型和基线模型，在测试集上进行评测。利用上一节介绍的自动评测指标对生成文本与参考句的重合度、流利度、多样性进行评测。自动评测结果如表 2 和表 3 所示，我们在各项评测指标上进行对比。

表 2　中文数据集自动评测结果

方法	Const/%	BLEU2/4	Dist-2	PPL
X-MCMC-C	100	7.2/1.7	**58.2**	89.9
DBA	100	8.6/1.9	36.0	147.4
Transformer	88.5	9.2/2.3	34.6	73.2
mT5	92.3	10.2/2.9	43.7	69.1
LCTG-SP	100	8.9/2.0	42.9	63.3
＋数据增强	100	**10.8**/**3.1**	48.3	**52.0**

表 3　英文数据集自动评测结果

方法	Const/%	BLEU2/4	Dist-2	PPL
X-MCMC-C	100.0	10.6/3.1	**73.9**	163.6
DBA	100.0	12.7/3.7	45.8	191.9
Transformer	97.4	13.0/5.5	45.4	99.8
T5	98.5	**15.0**/**7.0**	52.2	85.3
LCTG-SP	100	11.9/3.6	57.8	99.7
＋数据增强	100	13.4/5.8	59.9	**82.5**

从自动评测结果可以看出，在测试集上，LCTG-SP 方法生成的文本相比基线方法生成的文本的 PPL 指标更低，具有更好的文本流畅度。一方面，由于随机搜索的方式(X-MCMC-C)在生成文本时利用随机策略搜索文本序列，随机性很大，因此影响了文本的流利度。增强的柱搜索方法(DBA)是在解码过程中硬约束生成包含约束词的文本，也对文本流利度造成了较大影响。另一方面，由于我们的模型是基于预训练的语言模型，具有海量数据的语言知识，比随机搜索的 X-MCMC-C 方法、增强的柱搜索 DBA 方法以及重头训练的 Transformer 模型具有更好的文本流畅性。相比预训练模型 T5，我们的模型在英文表现上流畅性略差，中文上 LCTG-SP 方法表现更好。

在文本多样性方面，LCTG-SP 比 DBA、Transformer 生成的文本多样性更好。随机搜索方法由

① https://github.com/NLPCode/MCMCXLNet

于其固有的随机性，在文本多样性方面表现最好，甚至超过了人类真实参考文本的多样性。在BLEU指标上的表现与PPL类似，这一指标是计算与真实参考句之间的重合度，只能从一定程度上反映文本生成的质量。LCTG-SP方法是端到端方法中可以保证百分百包含约束词的，而基线方法中只有DBA和X-MCMC-C方法可以保证包含所有约束词，基线方法中的Transformer和T5无法保证生成文本中包含所有约束词。

我们进一步进行了多参考的数据增强实验（+数据增强），该方法可以进一步提升文本多样性和流利度，我们的模型结合了多参考数据增强后，在流利度方面达到了最佳性能，并且在其他指标上也具有较好性能。

3.5 人工评测与实例

为了进一步评估和验证方法的性能，我们进行了人工评测。人工评测数据是从测试集中随机抽取300条，并请三位人工标注员进行评测。人工评测是对生成文本的三个方面方面进行评价：①一致性；②多样性；③流利度。具体来说，一致性用于评测生成文本与约束词之间的语义相关性，即生成的文本的整体语义与约束词之间的相关程度。多样性指生成的文本的词汇和语义丰富度。流利度是指生成文本的通顺程度。三个指标均由打分方式评测，分数范围为1～5分，在所有测试样本上取平均值。由于人工评测需要耗费较多人力成本，因此对比的基线模型从随机搜索（X-MCMC-C）、增强柱搜索（DBA）和端到端方法（T5或mT5）各选取一种，本文的方法（LCTG-SP）是加入了数据增强方法之后的结果。人工评测最终结果如表4和表5所示。

表4 中文数据集的人工评测

方法	一致性	多样性	流利度
X-MCMC-C	3.1	**4.2**	2.9
DBA	3.3	3.2	2.3
mT5	3.7	3.5	3.6
LCTG-SP	**4.1**	3.9	**4.5**

表5 英文数据集的人工评测

方法	一致性	多样性	流利度
X-MCMC-C	2.8	**4.4**	3.1
DBA	3.3	2.9	2.4
T5	**3.8**	3.2	3.5
LCTG-SP	3.6	3.7	**4.1**

表6和表7给出了一些实例，可以对比同一组约束词在不同方法上生成的文本。分析观察表中生成的文本，可以看出增强的柱搜索方法（DBA）在文本流利度方面较差，而LCTG-SP的流利度较好，在进行了数据增强之后在流利度和多样性方面也有更好的表现。

表6 中文数据集的实例

约束词	校园，社团
X-MCMC-C	校园里的社团是一位来自中外的社团
DBA	可以在校园里进行社团，同时还可以参加社团活动
Transformer	我们在校园里参加了社团活动
mT5	校园里有各种各样的社团，也有自己的社团
LCTG-SP	我们经常到校园里做学生社团的调查
+数据增强	我们的校园里有各种各样的社团组织的活动
约束词	时间，空间，限制
X-MCMC-C	时间一长，你就会感到空间的限制，你可以选择
DBA	时间不能限制生活空间，但由于考试成绩的限制，人们的生活水平改变了
Transformer	随着时间的发展，空间的限制也变得越来越少了
mT5	时间和空间的限制逐渐淡化，人们的生活习惯也越来越简单了
LCTG-SP	由于时间和空间的限制，社会生活也变得不太方便
+数据增强	由于口头语言受时间和空间的限制，人们就产生了把自己的话记录下来的想法

表7 英文数据集的实例

约束词	energy, contracts, billion, year
X-MCMC-C	New York City energy supply contracts are only worth about $ 1 billion in a four year period.
DBA	energy contracts billion year on year are expected to be completed by 2023, the report said.
Transformer	The energy sector is expected to be hit by a series of new contracts, which will cost about $ 1 billion a year.
T5	Most energy companies will sign strategic contracts and will receive $ 10 billions in investments over the next year.
LCTG-SP	The energy giant said it would be investing in new contracts, which will cost it up to $ 1 billion a year.

续表

约束词	energy，contracts，billion，year
＋数据增强	There are also energy contracts between the two countries contributing to a volume of trade of $ 12 billion last year.
约束词	footballers，dream，career，row
X-MCMC-C	Meanwhile，Scottish footballers can celebrate with their dream of a professional career when their back row ends up at Wembley
DBA	footballers dream career row ：" I'm not going to be able to do it，but I'm going to do it . "
Transformer	These football players have the great career dream to become top football players and win the World Cup.
T5	The footballers dream of a career in the top flight and the best of the best in the world in a row.
LCTG-SP	For footballers，it's a dream career，and it's a dream to be back in the top row.
＋数据增强	Most footballers would dream of winning six league titles in their career，let alone doing it six times in a row.

3.6　生成速度对比

生成速度的测试是在中文的测试集上进行的，计算测试集所有数据完成生成任务的总时间。表 8 给出了各个方法的生成速度对比。本文的方法 LCTG-SP 是一种端到端方法，相比 DBA 和 X-MC-MC-C 方法的生成速度具有明显优势。这是由于 DBA 方法在柱搜索的解码过程中增加维护，包含了约束词的所有候选，增加了柱搜索的计算开销。而 X-MCMC-C 需要迭代多次修改文本，因此生成速度最慢。在三种端到端生成模型中，LCTG-SP 相比 mT5 的生成速度略快，比 Transformer 模型略慢。三者的解码方式相同，主要原因在于模型参数量的影响。由于 mT5 的参数量较大，因此速度最慢。

表 8　生成速度对比

方法	生成耗时
X-MCMC-C	9.6 h
DBA	1.5 h
Transformer	1.5 min
mT5	1.9 min
LCTG-SP	1.7 min

4　总结与展望

本文提出基于片段预测的端到端词汇约束文本生成方法，该方法在满足所有词汇约束的同时，保证了生成速度和生成质量。自动评测和人工评测表明，本文提出的 LCTG-SP 方法所生成的文本在流利性方面表现更好，多样性方面比增强的柱搜索和直接端到端生成的方法更好。此外，本文提出的数据增强方法能有效提升多样性，并且改善了流利度。未来我们将探索该任务在语言教学领域中的应用，结合语言和词汇教学的需求调整词汇约束文本生成任务。语言教学通常要求对文本的词汇难度进行控制，这对词汇约束文本生成赋予了更高的要求。

参考文献

[1] HU Z，YANG Z，LIANG X，et al. Toward controlled generation of text[C]//Proceedings of the International Conference on Machine Learning，2017：1587-1596.

[2] TANG H，LI M，JIN B. A topic augmented text generation model：joint learning of semantics and structural features[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing，2019：5090-5099.

[3] HE X，LI V O K. Show me how to revise：Improving lexically constrained sentence generation with xlnet[C]//Proceedings of the AAAI Conference on Artificial Intelligence，2021，35(14)：12989-12997.

[4] CHEN H，YI X，SUN M，et al. Sentiment-controllable Chinese poetry generation[C]//Proceedings of the International Joint Conference on Artificial Intelligence，2019：4925-4931.

[5] SHAO Y，SHAO T，WANG M，et al. A sentiment and style controllable approach for Chinese poetry generation[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management，2021：4784-4788.

[6] ZHAO C，WALKER M，Chaturvedi S. Bridging the structural gap between encoding and decoding for data-to-text generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics，2020：2481-2491.

[7] WANG Y，WOOD I，WAN S，et al. Mention flags (MF)：Constraining transformer-based text generators[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language

Processing, 2021: 103-113.
[8] MIAO N, ZHOU H, MOU L, et al. Cgmh: Constrained sentence generation by metropolis-hastings sampling[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 6834-6842.
[9] HOKAMP C, LIU Q. Lexically constrained decoding for sequence generation using grid beam search[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1535-1546.
[10] LIU X, PAN S, ZHANG Q, et al. Generating keyword queries for natural language queries to alleviate lexical chasm problem[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018: 1163-1172.
[11] YU S, LIU J, YANG J, et al. Few-shot generative conversational query rewriting[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020: 1933-1936.
[12] LIU X, HU J, SHEN Q, et al. Geo-BERT pre-training model for query rewriting in POI search[C]//Proceedings of the Association for Computational Linguistics: EMNLP, 2021: 2209-2214.
[13] YAO L, PENG N, WEISCHEDEL R, et al. Plan-and-write: Towards better automatic storytelling [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 7378-7385.
[14] CARO K, MENDINUETA N R. Lexis, lexical competence and lexical knowledge: A review[J]. Journal of Language Teaching & Research, 2017, 8(2): 125-138.
[15] POST M, VILAR D. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation [C]//Proceedings of NAACL, 2018: 1314-1324.
[16] SHA L. Gradient-guided unsupervised lexically constrained text generation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 8692-8703.
[17] BERGLUND M, RAIKO T, HONKALA M, et al. Bidirectional recurrent neural networks as generative models [J]//Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015: 856-864.
[18] WANG A, CHO K. BERT has a mouth, and it must speak: BERT as a Markov random field language model[J]//Proceedings of NAACL, 2019: 30-36.
[19] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL, 2019: 4171-4186.
[20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000-6010.
[21] ZHANG Y, WANG G, LI C, et al. POINTER: Constrained progressive text generation via insertion-based generative pre-training [C]//Proceedings of EMNLP, 2020: 8649-8670.
[22] DU Z, QIAN Y, LIU X, et al. GLM: General language model pretraining with autoregressive blank infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 320-335.
[23] QIN L, WELLECK S, KHASHABI D, et al. Cold decoding: Energy-based constrained text generation with langevin dynamics [C]//Proceedings of NeurLPS, 2022: 1-14.

聂锦燃(1994—),博士研究生,主要研究领域为自然语言处理和智能语言学习。
E-mail: jrnie@foxmail.com

杨尔弘(1965—),博士,教授,主要研究领域为语言信息处理、语言监测、语言资源建设。
E-mail: yerhong@blcu.edu.cn

杨麟儿(1983—),通信作者,博士,副教授,主要研究领域为自然语言处理和智能语言学习。
E-mail: yangtianlin@blcu.edu.cn