基于预训练及控制码法的藏文律诗自动生成方法

色差甲[1,2]、慈祯嘉措[1,2]、才让加[1,2(✉)]、华果才让[1,2]
1. 省部共建藏语智能信息处理及应用国家重点实验室;
2. 青海省藏文信息处理工程研究中心，青海 西宁 810008。
sechajia@126.com czjcaiyaogun@hotmail.com
(✉)zwxxzx@163.com 65332395@qq.com

摘要

诗歌自动写作研究是自然语言生成的一个重要研究领域，被认为是极具挑战且有趣的任务之一。本文提出一种基于预训练及控制码法的藏文律诗生成方法。在藏文预训练语言模型上进行微调后生成质量显著提升，然而引入控制码法后在很大程度上确保了扣题程度，即关键词在生成诗作中的平均覆盖率居高。此外，在生成诗作中不仅提高词汇的丰富性，而且生成结果的多样性也明显提升。经测试表明，基于预训练及控制码法的生成方法显著优于基线方法。

关键词： 藏文律诗自动生成 ； 藏文预训练模型 ； 控制码法

Automatic Generation of Tibetan Poems based on Pre-training and Control Code Method

Secha Jia, [1,2] Cizhen Jiacuo [1,2] , Cairang Jia [1,2(✉)] and Huaguo Cairang [1,2]
1. The State Key Laboratory of Tibetan Intelligent Information Processing and Application;
2. Tibetan Information Processing Engineering Technology and Research Center of Qinghai Province，Qinghai Xining 810008.
sechajia@126.com czjcaiyaogun@hotmail.com
(✉)zwxxzx@163.com 65332395@qq.com

Abstract

The study of automatic poetry writing is an important area of study in natural language generation and is considered one of the most challenging and interesting tasks. In this paper, a method for generating Tibetan poems based on pre-training and control code methods is proposed. The quality of the generation was significantly improved after fine-tuning on the Tibetan pre-trained language model. However, the introduction of the control code method has largely ensured the degree of deduction, that is, the average coverage of keywords in the generated poems is high. In addition, the richness of vocabulary is not only improved in generative poetry, but also the diversity of generative results is significantly improved. Tests have shown that the generation method based on pre-training and control code methods is significantly better than the baseline method.

Keywords: Tibetan poems automatically generated , Tibetan pre training model , Control code method

基金项目： 青海省重点研发与转化计划项目 (2022-GX-104)、青海省科技厅项目 (2020-ZJ-704)、国家自然科学基金 (62166034)

1 引言

近年来，如新闻、作文、诗歌等的自动写作，日益得到人工智能学界的逐渐兴起。其中，诗歌是人类文化的瑰宝，其短小精悍的语言却能表达出极其丰富的含义和主题，从古至今吸引了无数爱好者的欣赏。诗歌自动写作研究是自然语言生成的一个重要研究领域，被认为是极具挑战且有趣的任务之一。尤其是中文古诗的自动写作，是自然语言生成中最引人注目的研究课题之一 (孙茂松, 2020; 矣晓沅, 2021)。

从生成技术方面，自 2014 年基于端到端框架 (Sutskever and Vinyals et al., 2014; Cho and Merrienboer et al., 2014) 提出之后，文本生成迅速成为研究热点。其中 Transformer 已成为文本生成领域很受青睐的模型之一，同时近期在预训练加微调的新兴训练范式中也得到了广泛的应用，如 BERT (Devlin and Chang et al., 2019)、GPT (Alec and Karthik et al., 2018)、T5 (Raffel and Shazeer et al., 2020) 模型等。然而对于藏语自然语言生成任务而言，除了汉藏机器翻译 (桑杰端珠, 2019; 慈祯嘉措等, 2019; 头旦才让, 2021)、复述生成 (柔特, 2019)、摘要生成 (李亮, 2020) 的技术相对成熟之外，其他生成任务的技术研究尚处于初步探索阶段。色差甲等 (2018; 2019) 人实现了基于端到端的藏文律诗生成模型，并通过实验发现，该方法虽然能够提升生成质量和诗行之间的语义连贯性，但是对于关键词的扣题程度有所欠缺，进而会出现一些主题漂移问题，同时无法生成具有多样性的藏文律诗。因此，有必要在生成多样性和扣题程度等方面进行进一步的改进和完善。

针对以上问题，本文将提出一种预训练语言模型和控制码法相结合的藏文律诗生成方法。本方法特点为：其一，模型结构简洁: 只是使用了一个结构简单且高效的 Transformer 模型。与基线模型相比，主要区别在于本文模型提前进行了预训练。其二，扣题程度更好: 用藏文律诗语料进行预处理时引入了控制码法，即每个关键词、诗行之间以及生成任务中存在特定的分割标记，有助于模型引导生成，从而很大程度上能够确保扣题程度，防止出现主题漂移问题。其三，生成结果多样化：在解码过程中采用新的采样方法，从而在相同的形式和主题下能够生成多样化的藏文律诗，其结果显著优于基线模型。其四，能够生成藏头诗：本文方法还能够生成藏文藏头诗，即给定每个诗行的首位的藏文音节（四个音节），便可以在相应的位置生成给定的音节，并且保证在生成结果中形式和质量的要求。

本文方法所生成的藏文律诗比较接近于人类创作的诗歌。如图 1给出了四首藏文律诗，其中一首是人类创作的，其余三首是由本文方法所生成的。

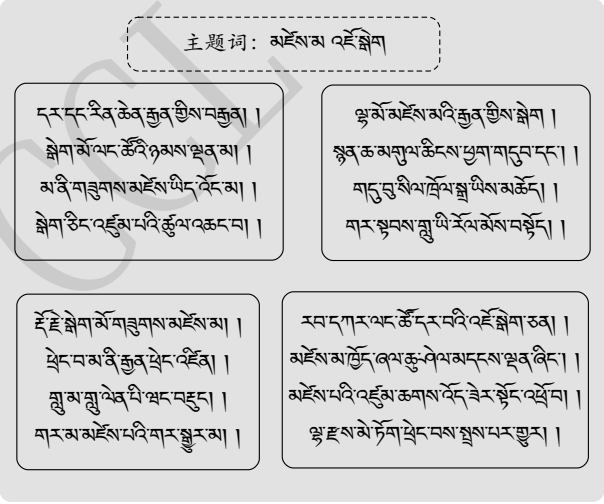

图 1: 一首真实和三首生成的藏文律诗

2 模型及数据预处理方式

为了能够生成句子更加流畅、语义更加连贯的一首藏文律诗，我们利用大规模的藏文文本语料预训练了语言模型。预训练语言模型将从综合性大规模文本语料中学到的词法信息、语法信息、以及语义信息等可直接迁移于藏文律诗生成模型。因此，该生成模型不是从零开始学习，而是在具备一定先验知识的基础上进行更进一步的学习。

第二十一届中国计算语言学大会论文集，第366页-第373页，南昌，中国，2022年10月14日至16日。

2.1 预训练模型及其数据处理方式

2.1.1 藏文预训练模型

T5 (Raffel and Shazeer et al., 2020) 和 BART (Lewis and Liu et al., 2019) 模型都是序列到序列的预训练降噪自编码器，与 BERT (Devlin and Chang et al., 2019) 相比有两个改变的点：第一种是在 BERT 的双向编码器架构中增添了因果解码器，即架构中包含编码器和解码器；另一种是用更复杂的预训练任务代替 BERT 的掩码语言模型任务。本文借鉴 BART 模型及其文本预处理方法，实现一个基于 Transformer 的藏文预训练语言模型，本文称之为 TiPLMT(Tibetan Pre-training Language Model based on Transformer)。TiPLMT 的数据处理及读取方式见图 2。从图 2中可知，TiPLMT 的源端输入是利用文本增强方法增强之后的文本（对原始文本进行增

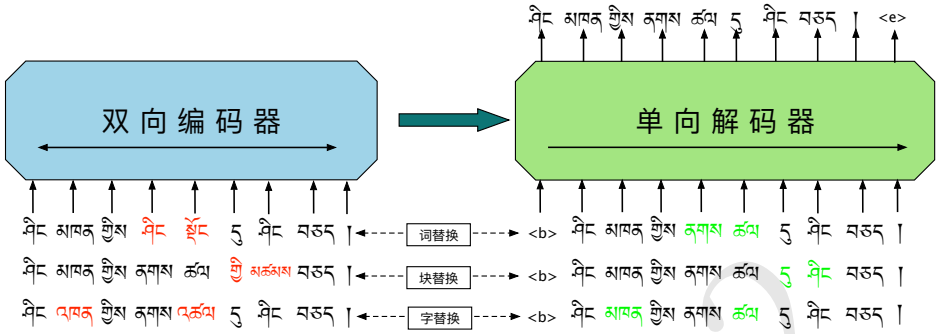

图 2: TiPLMT 模型及其数据读取格式

强的结果)，而目标端的输入便是原始文本。其目的便是通过语义信息不完备、缺失或含噪的文本，重构成语义完整且流畅的文本，使得模型不仅具有语言的表征能力，同时还具有文本错误纠正的能力。另外，TiPLMT 模型中弃用了特殊符号 [MASK] 的遮蔽机制，而是利用了藏文文本数据增强方法，也就是基于音节混淆子集和基于上下文的增强方法。从字（音节）、块（连续的片段，即 n-gram）、词等三个不同的层面对文本进行了增强并训练。

2.1.2 针对 TiPLMT 的文本预处理

预训练语言模型之前，从大规模无标注文本语料中通常以字、词或块为基本单位进行掩码(或替换) 处理来自动构建监督学习式信息，以便充分利用无标注文本语料。对于不同任务或不同语种而言，句子切分的颗粒度也不一样，预训练语言模型中把英语句子会切成词元 (Devlin and Chang et al., 2019)，汉文句子会切成字 (Cui and Che et al., 2021)，而本文将藏文句子切成音节。TiPLMT 从藏文文本语料中自动构建监督信息的方式进行了以下四种方法:

(1) 弃用特殊符号“[MASK]”的掩码方法，而是使用基于音节混淆子集和基于上下文的增强方法，从而可以把含加噪的句子和原句分别作为源句和目标句进行训练。

(2) 每个藏文句子中只处理 15% 的音节，其中不包含音节分隔符、垂直符、数字、特殊符号以及非藏文字符等，可又分为三种情况并分别为:

- 块层面的处理方式：先从原始文本中随机选取多个连续的音节（需要保证含有 15% 的音节），再随利用基于上下文的增强方法进行加噪处理其中单音节、双音节以及三音节所占比率分别为 20%、30% 和 50%;
- 音节层面的处理方式：先从原始文本中随机选取 15% 的音节，再利用基于音节混淆子集的增强方法进行处理，藏文虚词和实词所占比例分别为 40% 和 60%。
- 词层面的处理方式：先从已分词的原始文本中随机选取多个词（需要保证含有 15% 的音节），再随利用基于上下文的增强方法进行加噪处理，其中单音节、双音节以及多音节所占比率分别为 30%、50% 和 20%;

其中，音节混淆集合摘译藏文正字法，即由相互容易混淆的音节组成，共整理了 3912 个藏文音节，而且均是正式语料中出现频率相对较高的音节。基于上下文的增强方法的具体处理步

骤为：先对藏文原句中随机选取某个词或块，并用特殊符号 [MASK] 进行替换；然后利用基于上下文的增强方法进行重构；最后筛选出未能完全重构正确的句子，并与原句作为训练句对。藏文文本数据增强结果表 1 所示。

表 1: 藏文文本数据增强结果的示例

类型	藏文原句	重构的句子
词掩码	གང་ལ་[MASK]མེམས་དང་ལྡན་པ་དེ་ནི་མི་བཟང་བོ་འོ།། འདི་ལྟར་སྨོན་སྡུང་ལ་མི་དགའ་བ་ཞེག་ཡིན་ནང་[MASK] ཡོངས་ལ་ཤིན་ཏུ་དགའ་བ་ཞེག་ཏུ་སྐྱེ།	གང་ལ་[MASK]དང་ལྡན་པ་དེ་ནི་མི་བཟང་བོ་འོ།། འདི་ལྟར་སྨོན་སྡུང་ལ་མི་དགའ་བ་ཞེག་ཡིན་ནང་[MASK] ཡོངས་ལ་མི་དགའ་བ་ཞེག་ཏུ་སྐྱེ།
块掩码	ང་ལ་གནེར་སྡང་བརྒྱ་ཡོད་ན་ཅི་མ་རུང་། [MASK] ཡོངས་ལ་ཤིན་ཏུ་དགའ་བ་ཞེག་ཏུ་སྐྱེ།	ང་ལ་གནེར་ཁྲི་ཞེག་ཡོད་ན་ཅི་མ་རུང་། [MASK] ཡོངས་ལ་དགའ་བ་མང་བ་ཞེག་ཏུ་སྐྱེ།
字替换	རྟོམ་ཁ་འི་ཚོ་ན་ཤེས་བྱ་འི་རྣམ་གྲངས་ཟད།། འདི་ལྟར་སྨོན་ལ་མི་དགའ་བ་ཞེག་ཡིན་ནང་[MASK] ཡོངས་ལ་ཤིན་ཏུ་དགའ་བ་ཞེག་ཏུ་སྐྱེ།	རྟོམ་བ་འི་ཚོ་ན་ཤེས་བྱ་འི་རྣམ་གྲངས་ཟད།། འདི་ལྟར་སྨོན་ལ་མི་དགའ་བ་ཞེག་ཡིན་ནང་[MASK] ཡོངས་ཤིན་ཏུ་དགའ་ཁ་ཞེག་ཏུ་སྐྱེ།

从表 1 中可知，字替换是由混淆子集完成的，不仅能快速重构，而且根据藏文正字法能仿造含有真字拼写错误的句子，与正式文本中出现拼写错误现象一样，具有逼真的效果，但只能仿造音节级别的噪声；词或块替换是由提前预训练好的小模型完成的，因此重构速度相对较慢，但能仿造出含有语法或语义错误的句子，同样具有逼真的效果。原句中蓝色标记的音节是待替换的词（块），加噪句子中红色标记的音节是模型预测的音节，或者是用混淆子集随机替换的音节。

2.2 模型微调及控制码法

2.2.1 模型微调方式

我们的主要任务是在预训练模型 TiPLMT 的基础上，利用藏文律诗的语料进行进一步的微调。在微调过程中，将主题信息和藏文律诗以控制码法的方式嵌入到生成模型中，从而有效增强了模型的扣题程度，提高了生成结果的语义连贯性。模型微调方式及藏文律诗的输入格式如图 3所示。

图 3中，橙色部分表示编码器，蓝色部分表示解码器。关键词序列 S 作为编码器的输入（即源序列），藏文律诗序列 T 作为作为解码器的输入（即目标序列）。此处的目的便是微调一个条件自回归语言模型。

2.2.2 模控制码法的应用

控制码法 (Control Code)：是一种简单且有效的控制方法，即将所需要的控制指令，以字符的方式输入模型并作为生成的条件，该方法广泛应用于 BART、GPT-3 等预训练模型中。在汉文古诗生成方面，张家瑞等人 (2021) 把诗词序列化转化为由格式、主题和诗体等组成统一格式化的文本序列，作为训练数据并在 GPT 模型上进行微调，在绝句、律诗、藏头诗、词以及对联等生成任务上表现突出；Liao 和 Wang 等人 (2019) 在 GPT-2 模型中融入隐狄利克雷分配 (Blei and Ng et al., 2001) 模型的方法来实现了主题可控的诗歌生成，同时检验了主题模型 LDA 的有效性。

基于上述的研究成果，我们将结合藏文律诗的特征，实现一个预训练语言模型的基础上融入控制码法的藏文律诗生成方法。语料的主要处理格式见图 3所示，每首藏文律诗都按统一格式进行处理，是以“[Top] 主题词 1 [Top1] ... 主题词 n [Topn] [BOS] 藏文律诗 [EOS]”的格式进行预处理。由于 TiPLMT 模型进行预训练时，在藏文音节的粒度上为完成训练的，所有微调时同样把藏文律诗的切分粒度选为藏文音节，即每首藏文律诗切分成藏文音节。在具体处理过

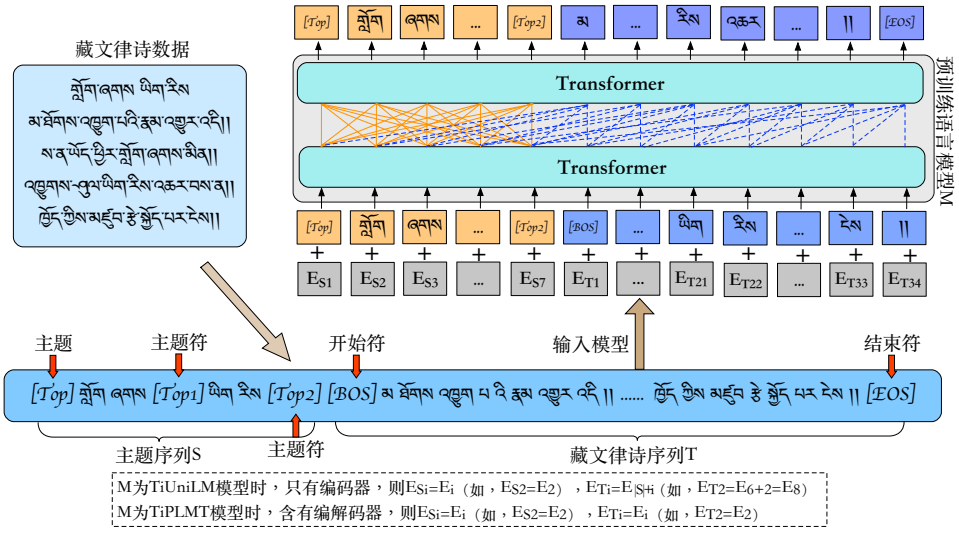

图 3: 模型微调方式及藏文律诗的输入格式

程中，我们将每个藏文律诗及其主题词用标识符进行标记和分割，并注入生成模型中作为生成的条件文本。对每首藏文律诗可以取 n 个主题词，则相对应的主题标识符也有 n 个（如 [Top1]，[Top2]，...，[Topn]）。本文主题词的取值范围设定为 $1 \leq n \leq 4$，其中含有一个、两个、三个和四个主题词的所占比例都一样，各占 25%（是人为设定的）。

3 实验

3.1 数据来源及规模

本文从藏文电子书籍及网页中共获取了含有 46.55 亿字符（4.53 亿藏文音节）的藏文文本语料，其主题包括文学、自传、诗歌、格言、散文和新闻等。对该语料进行了清洗、音节切分、音节拼写检错和纠正、以及句子切分等等预处理工作，最后共获得了 7.6 千万余藏文句子。其中，通过藏文律诗中垂直符的使用规律和诗行长度一致性等特征从中抽取去 131.3 万余首藏文律诗。据统计分析得出，藏文律诗中大部分为七言和九言律诗，共占 94.2%。首先从每首藏文律诗中通过关键词抽取算法 TextRank (Mihalcea and Tarau, 2004) 来抽取若干个关键词; 然后预处理成上述的格式要求; 最后从中抽取 2.5 千首藏文律诗对作为测试集，剩余部分作为训练集。

3.2 参数设置

本文的实验是在开源代码 Transforms1 的基础上完成的，藏文预训练语言模型 TiPLMT 的具体参数设置详见表 2 所示。

表 2: 模型参数设置情况

层数	词嵌入	全连接维度	注意力头数	优化器	学习率	最长序列	词表大小	参数规模
10	512	1024	10	Adam	3.8e-05	100	8663	4.7 千万

表 2 中，词表摘自《藏文规范音节频率词典》（多拉和扎西加, 2015）一书中，是从大规模的藏文文本中统计和整理的，只含有具有实际意义的音节，以及部分梵文。

3.3 基线方法与评测指标

由于目前面向藏文文本生成领域的相关研究较少。所以无法直接与前人的工作进行对比来验证本文所提出方法可行性和有效性，只能与重新复现的多个生成模型相比较。基于上述原因，本文将选用神经网络中常用于生成任务的几个经典模型作为基线模型，并分别为基于完全注意

力机制的 Transformer 模型 (Vaswani and Shazeer et al., 2017) 和基于生成式的预训练语言模型 GPT (Alec and Karthik et al., 2018)。

评价指标方面：将采用自动评测方法，从生成质量和生成多样性两方面进行评测。其中，注重生成质量方面的自动评价指标采用 PPL 和 BLEU 值，而注重的多样性的自动评价指标将采用 JS 值和 Distinct 值。

3.4 实验结果

由于以观察不同生成模型的有效性为目的，分别考查了 TiPLMT 以及基线模型等在藏文律诗的生成效果，是从语言建模能力和生成结果多样性方面进行评测分析，其对比实验结果如表 3 所示。

表 3: 不同模型在生成质量及多样性方面的对比实验结果

模型	PPL↓	BLEU↑(%)	Distinct↑(%)	JS↓(%)
Transformer	15.05	41.09	54.94	2.28
GPT	13.73	47.19	83.07	1.97
未用控制码法	15.82	46.58	79.96	2.02
未进行预训练	17.73	43.09	57.39	2.41
TiPLMT	9.28	51.02	94.46	1.15
未用控制码法	10.93	48.72	91.03	1.23
未进行预训练	15.98	46.05	63.74	2.31

从表 3 中可以看出，基线模型 Transformer 但在其他指标上则不然，这表明极限模型的生成结果中还是缺乏词的汇多样化使用。与基线模型 GPT 相比，TiPLMT 模型得益于语言模型预训练时融引入了文本数据增强技术和基于端到端的框架天然生成的优势，从而本文方法在生成藏文律诗的整体效果上仍获得更佳效果。TiPLMT 模型在 BLEU 值和 Distinct 值最高分别提升了 9.93 和 40.02 个百分点，JS 值降低了 1.13 个百分点，同时 PPL 取得了最低分数。这足以表明本文模型 TiPLMT 在生成质量、语句通顺度、以及词汇使用率等方面相对突出。此外，对于 TiPLMT 和 TiUniLM 模型而言，控制码法的使用和未使用之间存在显著差异，在 BLEU 值上分别能够提高 0.55 和 0.55 个点。同样，预训练语言模型的使用和未使用之间也存在显著差异，在 JS 值上分别降低 0.55 和 0.55 个点，在 PPL 值上分别降低 0.55 和 0.55 个点。

3.4.1 关键词数目对模型性能的影响

为了进一步观测各个模型的扣题能力，并统计不同数目的关键词在生成结果中的涵盖率，其统计结果如图 4所示。

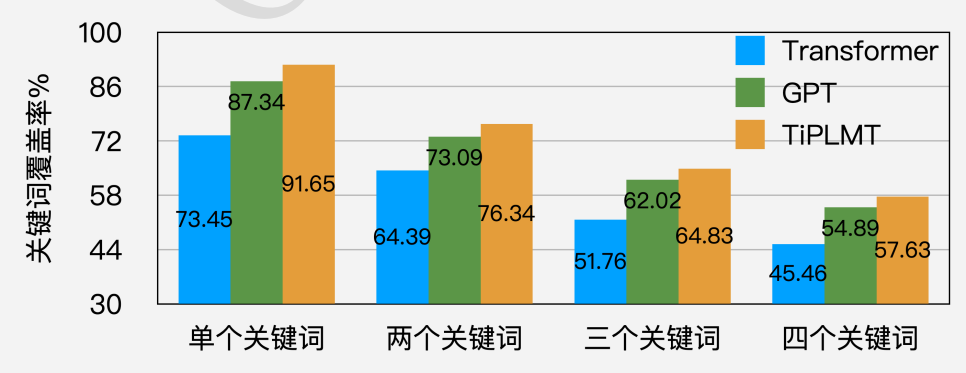

图 4: 生成结果中关键词的完全覆盖率

从图 4中可以看出，基线模型 Transformer 和 GPT 的关键词的覆盖率均不如 TiPLMT。对于 TiPLMT 模型而言，输入单个、两个、三个和四个关键词时，该模型的关键词完全包含率分

别为 91.65%、76.34%、64.83% 和 57.63%，则平均覆盖率高达 72.61%。显然，大部分关键词都能以某种形式在生成的藏文律诗中得到体现。

3.4.2 实例分析

本文方法有效提升了藏文律诗生成结果的多样性和扣题程度。下面我们将给出 TiPLMT 模型具体生成的诗作并进行分析，其部分生成实例如图 5 所示:

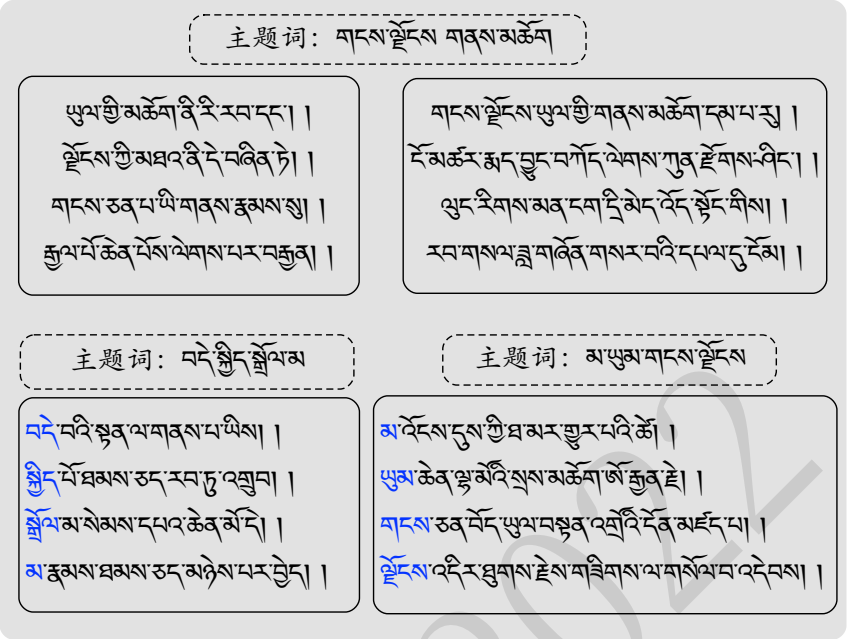

图 5: TiPLMT 模型的部分生成实例

从图 5-10 中可以看出，本文模型 TiPLMT 的生成质量不仅有所提升，而且能够生成多样化的藏文律诗。如给定主题词 “གངས་ལྗོངས་” 和”གནས་མཆོག” 时，该模型分别生成了七言和九言不同风格的藏文律诗，其中前者的音律节奏是 “ཚ+ཚ+ཚ+ཡ” 的类型，后者的音律节奏为 “ཚ+ཚ+ཚ+ཚ+ཡ” 的类型，这两个类型都是比较常用的音律节奏类型。以词为单位计算时，主题词在第一首实例中未出现，但以音节为单位时，所有主题词的音节均涵盖于该生成结果中，而且该音节在生成结果中的语义表现与主题词是一致的。主题词在第二首中是完全涵盖于生成结果中。此外，TiPLMT 模型还能够生成藏文藏头诗，如给定了四个音节的主题词“བདེ་སྐྱིད་སྐོལ་མ” 和 “མ་ཡུམ་གངས་ལྗོངས” 时，该模型分别生成了七言和九言的藏文藏头诗，其律诗节奏分别为 “ཚ+ཚ+ཚ+ཡ” 和 “ཚ+ཚ+ཚ+ཚ+ཡ”，可是，后者的第三行中出现了 “ཚ+ཚ+ཚ+ཡ+ཚ” 的律节奏类型，从而对整体朗读时稍微会影响顺口悦耳。

4 总结

为了提升扣题程度，我们提出了一种基于预训练语言模型和控制砝码相结合的生成方法。与基线模型相比，使用预训练语言模型后对于生成质量有显著提升。然而在藏文律诗语料进行预处理时引入了控制码法，即每个关键词、诗行之间以及生成任务中存在特定的分割标记，这有助于模型引导生成，从而在很大程度上确保了扣题程度，关键词的平均覆盖率高达 72.61%。

生成结果多样化方面也显著提升，一是提升了词汇使用的多样化，由于生成模型的初始参数源自预训练语言模型，从而降低了高频音节的重复使用率，更接近人类书写中高频音节的使用分布，最高频率的前 50 个藏文音节占了所生成内容的 30.5%（基线模型占了 41.3%，人类的占了 16.9%）；二是生成结果的整体多样化，我们在具体解码时采用了新的采样方法，从而在相同的形式和主题下能够生成高质且多样化的藏文律诗，其结果显著优于基线模型。

参考文献

Alec R, Karthik N, and Tim S, et al. 2018. *Improving Language Understanding by Generative Pre-Training.* https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language_unsupervised/language_understanding_paper.pdf.

Blei D M, Ng A Y, and Jordan M I. C. 2001. *Latent Dirichlet Allocation.* Advances in Neural Information Processing Systems(NIPS2001).

Cho K, Merrienboer B V, and Gulcehre C, et al. 2014. *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation.* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1724–1734.

Cui Y, Che W, and Liu T, et al. 2021. *Pre-Training With Whole Word Masking for Chinese BERT.* IEEE/ACM Transactions on Audio, Speech, and Language Processing, 3504-3514.

Devlin J, Chang M-W, and Lee K, et al. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL), 4171–4186.

Lewis M, Liu Y, and Goyal N, et al. 2019. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.* arXiv preprint arXiv:1910.13461v1.

Liao Y, Wang Y, and Liu Q, et al. 2019. *GPT-based Generation for Classical Chinese Poetry.* arXiv preprint arXiv:1907.00151v5.

Mihalcea R, Tarau P. 2004. *TextRank: Bringing order into text.* Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), 404-411.

Raffel C, Shazeer N, and Roberts A, et al. 2020. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* Journal of Machine Learning Research, 21(140): 1-67.

Sutskever I, Vinyals O, and Le Q V. 2014. *Sequence to Sequence Learning with Neural Networks.* Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), 3104–3112.

Vaswani A, Shazeer N, and Parmar N, et al. 2017. *Attention is All You Nee.* Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS2017), 6000-6010.

慈祯嘉措, 桑杰端珠, 孙茂松, 等. 2019. 融合单语语言模型的藏汉机器翻译方法研究. 中文信息学报, 33(12): 61-66.

多拉, 扎西加. 2015. 《藏文规范音节频率词典》. 北京: 中国社会科学出版社.

李亮. 2020. 基于 *ALBERT* 的藏文预训练模型及其应用. 兰州大学.

柔特. 2019. 藏文陈述句复述生成研究. 青海师范大学.

桑杰端珠. 2019. 稀疏资源条件下的藏汉机器翻译研究. 青海师范大学.

色差甲. 2018. 基于神经网络的藏文律诗生成研究. 青海师范大学.

色差甲, 华果才让, 才让加, 等. 2019. 注意力的端到端模型生成藏文律诗. 中文信息学报, 33(04): 68-74.

孙茂松. 2020. 诗歌自动写作刍议. 数字人文, (00): 32-38.

头旦才让. 2021. 汉藏神经机器翻译关键技术研究. 西藏大学, 61-73.

矣晓沅. 2021. 具有文学表现力的中文古典诗歌自动写作方法研究. 清华大学.

张家瑞, 李文浩, 孙茂松. 2021. 基于 *BPE* 分词的中国古诗主题模型及主题可控的诗歌生成. 第二十届中国计算语言学大会论文集 (CCL2021), 862-873.