使用艺人信息生成音乐播放列表标题

Haven Kim, [1] **Seungheon Doh,** [1] **Junwon Lee,** [2] **Juhan Nam** [1]

[1]文化技术研究生院，KAIST [2]电气工程系，KAIST khaven@kaist.ac.kr, seungheondoh@kaist.ac.kr, james39@kaist.ac.kr, juhan.nam@kaist.ac.kr

arXiv:2301.08145v1 [cs.IR] 14 Jan 2023

摘要

给定一组曲目，自动生成音乐播放列表标题或为其添加说明，在音乐流媒体服务中具有重要意义，因为定制播放列表广泛应用于个性化音乐推荐，而精心撰写的标题能够吸引用户并帮助他们发现音乐。我们提出了一种编码器-解码器模型，可根据一系列音乐曲目生成播放列表标题。以往的音乐播放列表标题生成工作将曲目 ID 作为词元化输入，而我们使用与曲目对应的艺术家 ID，以缓解播放列表数据集中曲目长尾分布带来的问题。此外，我们引入了一种按时间顺序划分数据的方法，以应对真实场景中新发行的曲目。通过比较以曲目 ID 和艺术家 ID 作为输入序列的方式，我们证明，基于艺人的方法在词语重叠度、语义相关性和多样性方面显著提升了性能。

1 引言

随着音乐流媒体服务的普及，音乐播放列表已成为从海量音乐中发现音乐的重要资源（Dias、Gonc.alves 和 Fonseca，2017年）。由于浏览大量音乐播放列表预计会有一定难度，人们通常会为播放列表提供一个能体现其内容信息的音乐播放列表标题，以吸引用户并帮助他们发现音乐。例如，标题可以描述音乐流派（如“把这个播放列表分享给那个不知道什么是独立摇滚的朋友”）、场景（如“最适合激励你慢跑的音乐”）、独特性（如“让你感觉自己像阿拉伯公主的播放列表”）等。然而，并非每个人都能轻松写出既有信息量又吸引人的标题；我们见过许多未包含有音乐意义信息的标题（如“我的最爱”），也有些标题仅由几个相关词语组成（如“冥想、禅修、瑜伽、睡眠、学习、专注”）。因此，如果能够自动生成描述性播放列表标题，就能让播放列表更吸引人，也更容易被发现。

已有多种方法被提出，用于自动生成播放列表的标题或描述。这些方法主要基于神经机器翻译模型，该模型将整段音乐信息编码为定长潜在向量，再将其解码为文本（Bahdanau、Cho 和 Bengio，2014年；Vaswani 等人，2017年）。具体而言，研究者使用基于 RNN 的序列到序列模型，从音频和补充文本数据（例如音乐流派）中学习（Choi 等人，2016年）；或使用基于 Transformer 的模型，将一系列曲目 ID（分配给每首曲目的唯一 ID）作为输入（Doh、Lee 和 Nam（2021）），为音乐播放列表生成标题。通过以往的研究，我们发现了两个主要问题。其一，大量数据很少出现，这使模型难以捕捉语义。因此，模型往往会反复生成在音乐上毫无意义的标题，例如“我最喜欢的歌曲”。另一个问题是，模型无法反映现实场景，即从包含新发行曲目的播放列表中进行推断。理想的音乐播放列表标题生成系统应能很好地泛化至不同类型的音乐播放列表；在根据包含罕见曲目或新发行曲目的播放列表进行推断时，其表现不应出现偏差。

为了解决这些局限性，我们提出了一种音乐播放列表标题生成模型，以艺术家 ID 而非曲目 ID 作为输入（第 2 节）。由于同一位艺术家的曲目都使用相同的输入嵌入向量，因此无论歌曲出现在播放列表中的频率如何，该模型都能够捕捉歌曲语义。为了更好地模拟真实世界的情境，我们提出了一种按时间顺序划分数据的方法，以模拟根据新发布曲目进行推断的情形（第 3 节）。此外，我们针对词语重叠度、语义相关性和多样性进行了系统评估（第 4 节）。第 5 节表明，艺术家 ID 嵌入方法提升了模型处理冷门曲目或艺术家的能力，同时不影响其处理热门曲目或艺术家的能力。最后，第 6 节讨论了本文的局限性和未来工作，并以此作结。可复现的代码已在线发布 [1]。

2 音乐播放列表标题生成

编码器-解码器模型

与先前的研究（Doh、Lee 和 Nam，2021年）类似，我们的音乐播放列表标题生成模型采用编码器—解码器框架：编码器网络用于获取音乐曲目的语义，解码器用于生成

版权所有 © 2023，人工智能促进协会（www.aaai.org）。保留所有权利。

[1]github.com/havenpersona/标题生成

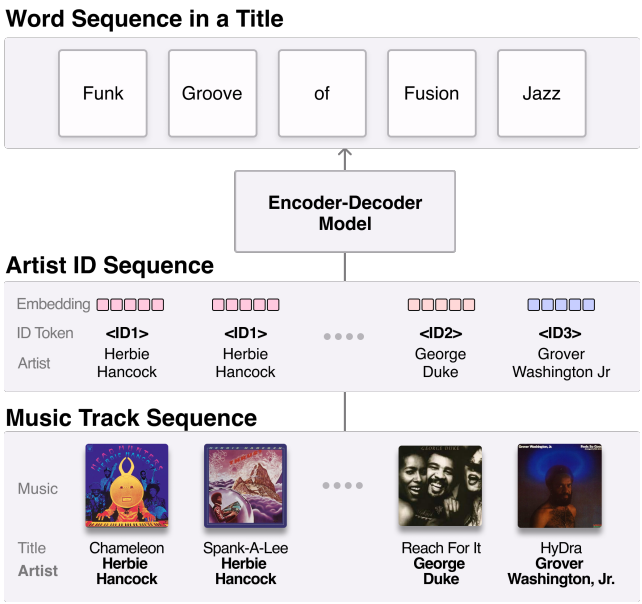


图1：模型架构。

自然语言序列。我们采用 Transformer 架构（Vaswani 等人，2017），该架构包含两到三个子层：多头注意力子层、全连接子层，以及解码器中的掩码多头注意力层。如图1所示，该架构以音乐曲目序列中的唯一词元 ID 作为输入。在所提方法中，音乐曲目序列由曲目或艺人信息 x 构成，其中每个曲目或艺人都表示为一个唯一词元（x_1、……、x_n）。词元 ID 由 d_{model} 维随机向量初始化，并由编码器网络 f：$\mathcal{X} \mapsto \mathcal{Z}$ 进行处理。这些向量被转换为一系列潜在表示$z = (z_1, \ldots\ldots, z_n)$，并输入解码器 g：$\mathcal{Z} \mapsto \mathcal{Y}$，由其生成输出序列 $\hat{y} = (\hat{y}_1, \ldots\ldots, \hat{y}_m)$，其中 $\hat{y}$ 表示生成的音乐播放列表标题，而 $(\hat{y}_1, \ldots\ldots, \hat{y}_m)$ 中的每个元素都表示标题中的一个词元。我们使用 softmax 交叉熵损失来比较真实值与预测结果。

音乐曲目表示

我们将音乐曲目序列作为编码器的输入。曲目序列中的每个元素都由元数据表示，例如曲目 ID 和艺术家 ID。与艺术家相比，曲目能提供更具体的信息。如此高的具体性或许能捕捉音乐语义的细节，但如果某首曲目在训练阶段出现次数很少，系统往往无法提取有意义的语义。例如，在我们实验所用的播放列表数据集中，超过 17% 的曲目在训练阶段出现的次数不超过两次。另一方面，出现次数超过两次的艺术家 ID 不到 5.0%。因此，使用艺术家 ID 或许能让系统从同一艺术家此前歌曲的信息中提取语义。

数据集	统计数据	原始	Doh 等人	所提方法
Melon	播放列表数量	148,826	51,404	80,160
	唯一曲目数量	649,902	429,266	516,220
	独立艺人数量	115,457	77,145	92,020
	独立标题数量	115,318	59,209	77,905
	唯一词数量	88,524	49,978	62,834
	平均字符长度	2.8	3.6	3.1
	平均标题长度	3.6	4.7	5.2
	平均曲目长度	39.7	46.2	44.6
MPD	播放列表数量	1,000,000	50,083	34,115
	独立曲目数量	2,262,292	402,523	303,779
	独立艺人数量	286,787	69,641	55,103
	独立标题数量	17,381	1,859	1,229
	唯一词数量	11,146	1,886	1,258
	平均字符长度	5.2	4.2	4.1
	平均标题长度	1.4	3.4	3.4
	平均曲目长度	66.3	66.3	64.2

表 1：数据集统计。

数据集	实体	<UNK> 比例
Melon	曲目 ID	14.70%
	艺术家 ID	3.53%
MPD	曲目 ID	9.86%
	艺术家 ID	2.91%

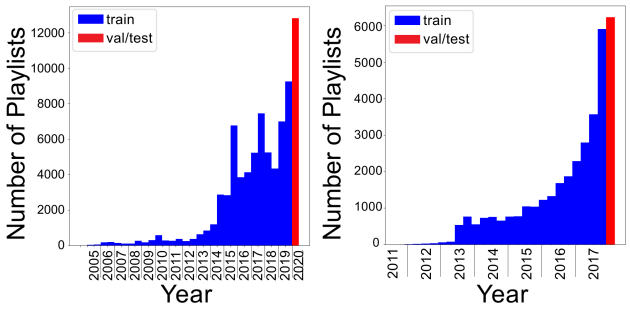
表 2：验证集和测试集中 <UNK> 词元的比例

3 数据集与实验

我们使用了两个不同的数据集：Melon 播放列表数据集（Melon）（Ferraro 等，2021）和 Spotify MillionPlaylist 数据集（MPD）（Chen 等，2018）。这两个数据集的主要语言分别为韩语和英语。它们最初是为自动播放列表续接（APC）任务而创建的，包含播放列表标题、音乐曲目序列、音乐元数据（曲目、艺术家名称）以及最后修改日期。利用这两个数据集，先前的研究（Doh、Lee 和 Nam（2021））提出了一种随机数据划分方案和一种用于生成播放列表标题的噪声清理策略。然而，这种方法有两个局限：第一，在随机划分的数据上进行验证和测试，无法模拟根据新发布的曲目和艺术家进行推断的情形。第二，过滤策略只考虑了标题的结构质量，却忽略了语义质量。这会降低系统在实际应用中的可用性。为解决这一问题，我们提出了一种新的数据划分和噪声过滤策略。

按时间顺序划分数据

在现实世界中，由于每天都有大量音乐曲目上传到音乐流媒体平台，模型需要对包含新发行曲目的播放列表进行预测。在随机划分的数据集上进行验证和测试，无法反映这一现实。为克服这一局限，我们按照时间顺序划分数据进行训练、验证和测试，如图 2 所示，并用一个新发行曲目的



(a) Melon 播放列表 (b) Million 播放列表

图 2：按时间顺序划分的数据。训练数据以蓝色标示，验证/测试数据以红色标示。横轴按年份和季度划分。

特殊标记 <UNK> 替代该曲目。对于 Melon 播放列表数据集，我们将 2020年创建的播放列表用作验证集和测试集。在此之前创建的播放列表均用作训练数据。对于 Million 播放列表数据集，我们将 2017年最后一个季度创建的播放列表用作验证集和测试集。在此之前创建的播放列表均用作训练数据。训练集与验证集/测试集的近似比例分别为 83:17 和 82:18。如表 2 所示，以曲目 ID 而非艺术家 ID 作为输入时，<UNK> 词元的比例显著更高，这表明以曲目 ID 作为输入时，模型更有可能处理未见数据。

过滤噪声数据

正如先前的研究（Doh、Lee 和 Nam（2021））所指出的，数据集中有些播放列表的标题不理想，或曲目数量很少（包括没有曲目的播放列表）。因此，他们只使用符合以下条件的播放列表：i）标题经空白字符分词后至少包含三个词元。ii）词元的平均字符长度至少为三个。iii）播放列表至少包含十首曲目。我们通过放宽第二和第三项条件，并新增一项条件，对这些标准进行了修改。新增条件旨在避免生成缺乏音乐意义信息的标题，例如 “适合聆听的好音乐” 或 “史上最佳音乐”。因此，我们只使用至少包含一个预定义音乐标签词语的播放列表。我们通过以下步骤获取音乐标签列表。

韩语：韩语音乐标签列表基于 Melon 播放列表数据集中提供的 30,652 个唯一标签。我们仅考虑了在整个数据集中出现至少 50 次的标签。我们排除了诸如“推荐” 和“合辑”等在音乐上毫无意义的标签。韩语标签总数为 731 个。

英语：Million 播放列表数据集未提供音乐标签。我们从两个公开可用的音乐自动标注研究数据集中获取了英语标签列表（MTG-Jamendo 数据集（Bogdanov 等（2019））和 Million Song 数据集（Bertin-Mahieux 等（2011））。由于这些数据集不包含近期流行起来的许多音乐流派（例如，“陷阱音乐”），我们还收集了其他标签，这些标签来自 YouTube Music 上频繁出现的词语。同样，我们移除了不适用于音乐播放列表描述的标签。最终，我们共获得 1,013 个英语标签。

总而言之，我们筛除了不符合以下标准的播放列表：i）经过空白字符分词后，标题至少包含三个词元。ii）词元的平均字符长度至少为两个。iii）播放列表至少包含两首曲目。iv）标题至少包含一个在音乐上有意义的词。通过调整筛选标准，我们收集到了包含有意义信息的标题；所有标题都至少包含一个与音乐相关的词，同时没有降低标题的平均长度（见表 1）。

训练细节

模型使用 Adam优化器进行优化，学习率为 0.005，并采用 10^{-4} 学习率衰减（Kingma 和 Ba，2014年）。学习率通过余弦退火调度器进行调整，最低学习率为 10^{-6}（Loshchilov 和 Hutter，2017年）。批量大小设为 64，并在每个训练轮次结束时进行验证。遵循先前的研究，我们忽略了输入序列的位置，因为在大多数情况下，生成标题时，音乐播放列表中曲目或艺术家的顺序并不提供关键信息。

4 评估指标

在评估中，我们使用文本生成评估指标（Celikyilmaz、Clark 和 Gao，2020年），并将其分为三类：(i) 基于 n-gram 的重叠指标 (ii) 基于 BERT 的神经网络指标 (iii) 基于 n-gram 的多样性指标。我们利用这些指标评估整体表现。此外，我们还衡量鲁棒性，其定义为模型在处理较少出现的数据与频繁出现的数据时，性能变化的程度。

n-gram 重叠指标

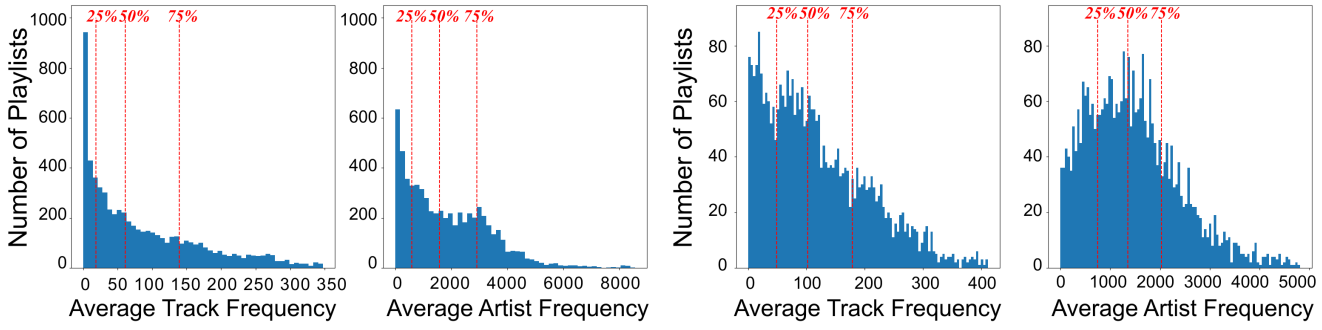
统计参考文本 $y = (y_1, \ldots\ldots, y_l)$ 与候选文本 $\hat{y} = (\hat{y}_1, \ldots\ldots, \hat{y}_m)$ 之间重叠的 n-gram 数量，是评估文本生成模型最常见的方法之一。本文使用三种流行的 n-gram 重叠指标来评估模型性能——BLEU、ROUGE 和 METEOR。

双语评估替补（BLEU）被提出用于评估机器翻译性能，其中包括修正后的 n-gram 精确率（Papineni 等人，2002年）。对于一个参考句子和一个候选句子，修正精确率分数的计算方法是：统计匹配的 n-gram 数量，再除以候选句子中 n-gram 的数量。本文针对两个不同的 n 值（$n = 1, 2$）计算了 BLEU。

面向召回率的摘要评估替补（ROUGE）最初用于评估文本摘要，后来也用于评估短文本生成（Lin，2004年）。BLEU侧重于精确率，而ROUGE

数据集	输入	Loss	n 元语法重叠度指标				基于 BERT 的指标		
		负对数似然	BLEU-1	BLEU-2	ROUGE-1	ROUGE-2	METEOR	BERT 分数	SentBERT
Melon	曲目 ID	6.40	0.048	0.01	0.061	0.01	0.041	0.417	0.775
	艺术家 ID	6.23	0.056	0.01	0.074	0.01	0.049	0.430	0.778
MPD	曲目 ID	1.35	0.232	0.180	0.237	0.174	0.212	0.889	0.408
	艺术家 ID	1.17	0.288	0.229	0.293	0.223	0.265	0.897	0.460

表 3：曲目 ID 嵌入模型与艺术家 ID 嵌入模型的比较。结果表明，使用艺术家 ID 嵌入时，词语重叠度和语义相关性均有所提升。



(a) Melon 播放列表

(b) Million 播放列表

图 3：平均曲目频率（F_t）和平均艺术家频率（F_a）的分布。我们移除了高于第 99 百分位数的项目，以排除统计异常值。添加红线是为了直观展示将数据分为四部分的数值。

同时使用召回率和精确率。本文采用 ROUGE-N 的 F1 分数（ROUGE 的一种变体）来评估我们的模型。该分数通过计算预测（匹配的 n-gram 数量除以候选文本 n-gram 的数量）和召回率（匹配的 n-gram 数量除以参考文本 n-gram 的数量）的调和平均数得出。n 的取值为 1 和 2。

显式排序翻译评估指标（METEOR）旨在解决 BLEU 在机器翻译评估中的一些局限性（Banerjee 和 Lavie，2005年）。与 ROUGE 类似，该指标同时考虑召回率和精确率。其计算方法是求单元语法精确率分数与单元语法召回率分数的加权调和平均值，其中召回率的权重是精确率的九倍。最终得分结合了这一结果与块惩罚。块数（连续单元语法）越多，惩罚就越大。如果没有二元语法匹配，则给予最高惩罚（0.5）。

基于 BERT 的神经网络指标

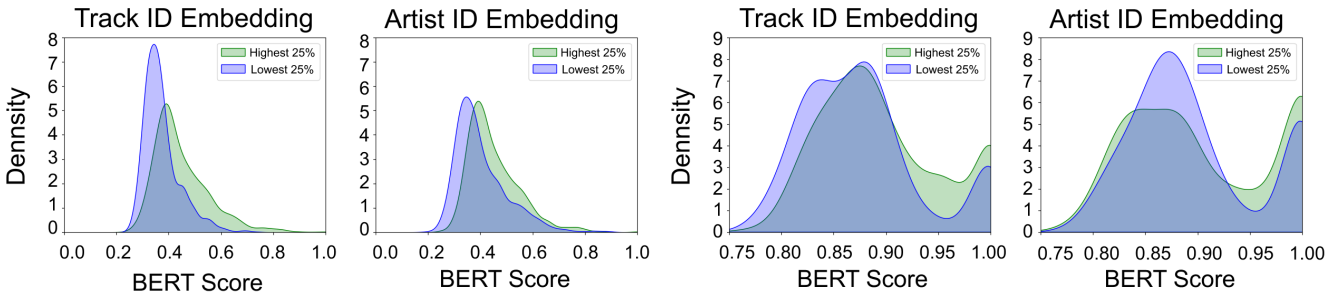
基于 BERT 的指标不再统计精确匹配的数量，而是使用模型生成的预训练上下文嵌入，捕捉参考句子 $y = (y_1, \ldots \ldots y_l)$ 与候选句子 $\hat{y} = (\hat{y}_1, \ldots\ldots \hat{y}_m)$ 之间的语义相似度。

BERT 分数通过计算参考句子中的每个词元与候选句子中的每个词元之间的余弦相似度，并使用贪心匹配使分数最大化（Zhang 等，2019）。本文使用预训练的 KcBert-Base 模型（Lee，2020）评估 Melon 播放列表数据集，并使用预训练的 RoBERTa-large 模型（Zhuang 等，2021）评估 Million Playlist 数据集，以获取上下文嵌入。

Sentence BERT（SentBERT）最初是为语义文本相似度评估任务提出的（Reimers 和 Gurevych，2019）。SentBERT 使用孪生预训练 BERT 网络，生成具有语义意义的句子嵌入。每个嵌入通过对每个词元的向量表示取平均值计算得出，并计算它们之间的余弦相似度。本文采用 KLEU RoberTa small（Park 等，2021）的嵌入来评估 Melon 播放列表数据集，并采用 all-MiniLM-L6-v2（Wang 等，2020）的嵌入来评估 Million 播放列表数据集。

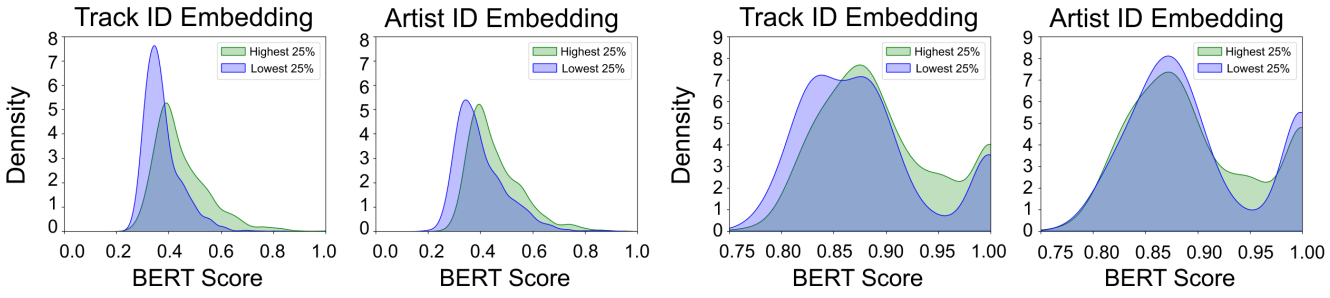
n-gram 基于的多样性指标

成功的音乐播放列表标题生成模型应避免生成重复的标题，而应生成多样化的输出。研究人员指出，基于 n-gram 的多样性指标与词汇多样性高度相关（Tevet 和 Berant，2021年）。用于评估多样性的 n-gram 指标之一是 *distinct-n*，即不同 n-gram 的数量除以 n gram 的总数（Li 等，2016年）。在本文中，我们使用以下指标评估词汇变化



(a) Melon 播放列表　　(b) Million 播放列表

图 4：最高 25% F_t 组和最低 25% F_t 组的 BERT 分数分布。使用艺术家 ID 嵌入后，最低 25% 组的分数有所提升，同时未影响最高 25% 组的表现。



(a) Melon 播放列表　　(b) Million 播放列表

图 5：最高 25% F_a 组和最低 25% F_a 组的 BERT 分数分布。使用艺术家 ID 嵌入后，最低 25% 组的分数有所提升，同时未影响最高 25% 组的表现。

distinct-n，并采用 n 的三个取值（$n = 1$，2，3）。

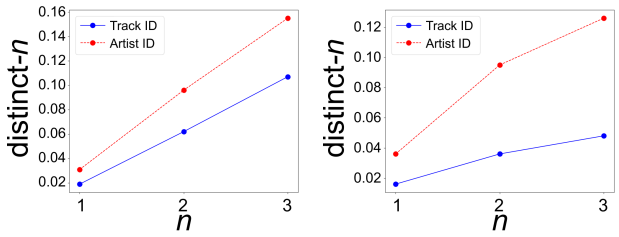
5 结果与讨论

词语重叠度与语义相关性

表 3比较了不同输入模型（曲目 ID、艺术家 ID）的所有 n-gram 重叠度表现（BLEU、ROUGE 和 METEOR）。与曲目 ID 嵌入模型相比，艺术家 ID 嵌入模型的性能有明显提升。对于 Melon 播放列表数据集，由于韩语的语法特点（后置助词紧跟名词或代词，且中间没有空格），BLEU-2 的变化并不明显。在负对数似然（NLL）损失和基于 BERT 的神经网络指标（BERT 分数和 Sentence BERT）上，我们也观察到了类似趋势。这表明，利用艺术家 ID 可以缓解处理低频出现曲目的问题，使系统能够提取更具音乐意义的信息。

多样性

图6显示，无论 n 的值如何，使用艺术家 ID 嵌入时 *distinct-n* 都会增加。尽管曲目 ID 的特异性更高，艺术家 ID 嵌入模型生成的输出在词汇上更加多样。这表明，曲目 ID（表 2）中较高的 <UNK> 词元比例不仅会降低语义相关性，也会影响多样性表现。



(a) Melon 播放列表　　(b) Million 播放列表

图6：*distinct-n* (n = 1,2,3) 表示艺术家 ID 嵌入会生成词汇多样性更高的标题。

鲁棒性

我们通过比较模型处理低频出现数据与高频出现数据时的表现，评估其鲁棒性。我们从两个角度定义出现频率：平均曲目频率和平均艺术家频率。基于这两个角度，我们将整个播放列表平均分成四组，并比较模型在最高的 25% 和最低的 25% 数据上的表现。在本文中，我们分别将它们记为 F_t和 F_a。平均曲目频率 F_t 和平均艺术家频率 F_a 按以下方式计算。当一个播放列表由曲目（t_1，……，t_m）和艺术家（a_1，……，a_n）组成时，平均曲目频率 F_t 是通

PID	143802	75870
F_t	355.39	0.75
真实值	充满咖啡馆的情感歌曲 （触动人心的音乐氛围弥漫在咖啡馆中。）	适合慵懒午后的轻柔器乐演奏 （在昏昏欲睡的下午，聆听令人平静的旋律）
曲目 ID 嵌入	伴着一杯热咖啡聆听抒情曲 (手捧一杯咖啡，聆听抒情曲))	电影原声音乐精选 (电影原声带合辑)
艺术家 ID 嵌入	感伤夜晚聆听的情歌 （伴着动人音乐度过的感伤之夜）	聆听舒缓的钢琴旋律演奏的古典音乐 （古典音乐中舒缓的钢琴旋律）

表4：I nference 最高25% F_t 组的示例（**左侧**）和最低25% F_t 组的示例（**右侧**） .

PID	117646	116152
F_a	3161.39	540.40
真实值	月光下传来的甜美嗓音 （月光下的一把甜美嗓音）	跟上韩国嘻哈的潮流 （跟上韩国嘻哈圈的潮流）
曲目 ID 嵌入	伴着一杯热咖啡聆听抒情曲 （手捧一杯咖啡，聆听抒情曲）	电影中的背景摇滚音乐，电影背景音乐合集 （电影中的背景摇滚音乐，电影背景音乐合辑）
艺术家 ID 嵌入	在静谧夜晚聆听的动人歌曲 （适合在宁静夜晚聆听的动人歌曲）	韩国嘻哈的起步与成熟 （韩国嘻哈：从起步到成熟）

表 5：得分最高的 25% F_a 组（**左侧**）和得分最低的 25% F_a 组（**右侧**）的推理示例。

过取曲目频率（f_1，……， f_m）的平均值获得的。类似地，我们可以通过对训练集 ^ ^（f_1，……， f_n）中每位艺术家的频率取平均值，计算平均艺术家频率得分F_a （ m 和 n 的数量可能不同，因为一首曲目可能有多位艺术家）。

$$F_t = \frac{1}{m}\Sigma_{x=1}^{m} f_x \tag{1}$$

$$F_a = \frac{1}{n}\Sigma_{x=1}^{n} \hat{f}_x \tag{2}$$

如图 3 所示，低 F_t 的播放列表数量远多于低 F_a的播放列表。

如图 4 所示，在语义相关性方面，艺术家 ID 嵌入模型在得分最低的 F_t 组（得分为 0% 至 25%）上的表现明显优于曲目 ID 嵌入。与此同时，该模型并未损害其在得分最高的 F_t 组（得分为 75% 至 100%）上的表现。事实上，与 Million 播放列表数据集中的情况一样，使用艺术家 ID 嵌入时，得分最高的 F_t 组的表现也有所提升。因此，各子组之间的性能差距得以缩小。我们还观察到，艺术家 ID 嵌入提升了得分最低的 F_a 组（得分为 0% 至 25%）的表现，并缓解了性能差距，如表 5 所示。

定性结果

我们在表4和表5中提供了 Melon 播放列表数据集中的四个播放列表推理示例及其 PID（分配给每个播放列表的唯一识别码），并附有英语翻译。每个播放列表分别来自 F_t 最高的 25%、 F_t 最低的 25%、 F_a 最高的 25% 和 F_a 最低的 25% 组。无论输入如何，模型都成功生成了一个短语式标题，用于描述来自 F_t 最高的 25% 组和 F_a 最高的 25% 组的播放列表的情绪（例如，动人）或地点（例如，咖啡馆）。使用艺术家 ID 嵌入训练的模型生成了内容丰富的标题，体现了来自 F_t 最低的 25% 组和 F_a 最低的 25% 组的播放列表的情绪（舒缓）、原创性（韩语）和音乐流派（嘻哈）。然而，使用曲目 ID 训练时，模型无法为来自 F_t 最低的 25% 组和 F_a 最低的 25% 组的播放列表生成信息正确的标题（PID：75870）。真实值将每个播放列表分别定义为舒缓旋律 和韩国嘻哈，而模型则分别将其预测为电影原声带 和电影中的摇滚音乐。

6 结论

本文提出了一种音乐播放列表标题生成模型，以一系列艺术家 ID 作为输入。在整个评估过程中，我们表明，艺术家 ID 嵌入能够提升模型在词语重叠度、语义相似度和多样性方面的表现。我们还证明，艺术家 ID 嵌入增强了模型的鲁棒性，因为模型较不容易处理罕见或从未见过的数据。我们的研究结果得到了两种不同语言的数据集的支持。然而，我们的研究存在两个局限。首先，在现实世界中，单个艺术家可能会随着时间推移涉足多种音乐流派和风格。固定的艺术家 ID 嵌入无法体现这种多样性。其次，艺术家 ID 嵌入模型仍需处理未见数据（新艺术家）。解决这一问题的一种方法是使用基于内容的音频嵌入作为编码器输入，并观察模型的表现。未来，我们计划尝试不同的输入，例如音频、音乐流派标签，或多种输入的混合。

参考文本

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014年。通过联合学习对齐和翻译实现神经机器翻译。载于国际学习表征会议（*ICLR*）论文集。

Banerjee, S.; and Lavie, A. 2005. METEOR：一种与人工判断相关性更高的机器翻译评估自动指标。载于机器翻译和/或摘要的内在与外在评估指标*ACL*研讨会论文集。Bertin-Mahieux, T.; Ellis, D.; Whitman, B.; and Lamere, P. 2011. 百万歌曲数据集。载于第*12*届国际音乐信息检索会议（*ISMIR*）论文集，591–596。Bogdanov, D.; Won, M.; Tovstogan, P.; Porter, A.; and Serra, X. 2019. 用于自动音乐标注的MTG-Jamendo 数据集。载于音乐发现机器学习研讨会，国际机器学习会议（*ICML*）。美国加利福尼亚州长滩。Celikyilmaz, A.; Clark, E.; and Gao, J. 2020. 文本生成评估：综述。*arXiv*预印本 *arXiv:2006.14799*。Chen, C.-W.; Lamere, P.; Schedl, M.; and Zamani, H. 2018. RecSys挑战赛2018：自动续播音乐播放列表。载于第*12*届*ACM*推荐系统会议论文集，527–528。Choi, K.; Fazekas, G.; McFee, B.; Cho, K.; and Sandler, M. 2016. 探索音乐描述生成：生成音乐播放列表描述。第*22*届国际音乐信息检索学会会议（*ISMIR*）最新研究成果演示。Dias, R.; Gonc,alves, D.; and Fonseca, M. J. 2017. 从手动到辅助播放列表创建：综述。载于《多媒体工具与应用》。Doh, S.; Lee, J.; and Nam, J. 2021. 音乐播放列表标题生成：一种机器翻译方法。第二届音乐与语音音频自然语言处理研讨会、第*22*届国际音乐信息检索学会会议（*ISMIR*）论文集。Ferraro, A.; Kim, Y.; Lee, S.; Kim, B.; Jo, N.; Lim, S.; Lim, S.; Jang, J.; Kim, S.; Serra, X.; et al. 2021. Melon 播放列表数据集：用于基于音频的播放列表生成和音乐标注的公开数据集。载于国际声学、语音与信号处理会议（*ICASSP*）论文集。Kingma, D. P.; and Ba, J. 2014. Adam：一种随机优化方法。载于国际学习表征会议（*ICLR*）论文集。Lee, J. 2020. KcBERT：韩语评论BERT。载于第*32*届人类与认知语言技术年会论文集，437–440。Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, W. B. 2016. 面向神经对话模型的多样性促进目标函数。载于*2016*年计算语言学协会北美分会年会：人类语言技术论文集，110–119。

Lin, C.-Y. 2004. ROUGE：一种用于自动评估摘要的软件包。载于文本摘要的拓展，74–81。Loshchilov, I.; and Hutter, F. 2017. 随机梯度下降与热重启（SGDR）。载于国际学习表征会议（*ICLR*）。Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU：一种机器翻译自动评估方法。载于第*40*届计算语言学协会年会论文集，311–318。Park, S.; Moon, J.; Kim, S.; Cho, W. I.; Han, J.; Park, J.; Song, C.; Kim, J.; Song, Y.; Oh, T.; Lee, J.; Oh, J.; Lyu, S.; Jeong, Y.; Lee, I.; Seo, S.; Lee, D.; Kim, H.; Lee, M.; Jang, S.; Do, S.; Kim, S.; Lim, K.; Lee, J.; Park, K.; Shin, J.; Kim, S.; Park, L.; Oh, A.; Ha, J.; and Cho, K. 2021. 韩国语言理解评测：韩语语言理解评测。arXiv:2105.09680。Reimers, N.; and Gurevych, I. 2019. 句子BERT：使用孪生BERT网络的句子嵌入。载于*2019*年自然语言处理实证方法会议暨第*9*届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，3982–3992。Tevet, G.; and Berant, J. 2021. 自然语言生成多样性评估的评估。载于欧洲计算语言学协会分会第*16*届会议论文集：主卷，326–346。Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. 注意力就是你所需要的一切。载于神经信息处理系统进展会议论文集（*NeurIPS*）。Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; and Zhou, M. 2020. MiniLM：面向任务无关压缩的深度自注意力蒸馏预训练Transformer。神经信息处理系统进展，33：5776–5788。Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. BERT 分数：使用BERT评估文本生成。载于国际学习表征会议（*ICLR*）。Zhuang, L.; Wayne, L.; Ya, S.; and Jun, Z. 2021. 一种稳健优化的BERT预训练与后训练方法。载于第*20*届中国计算语言学大会论文集，1218–1227。