•人工智能•

DOI:10.12454/j.jsuese.202400082

本刊网刊

基于知识蒸馏的民间文学文本修复

曹熊能[1,2],王笳辉[1,2*],岳　昆[1,2],段　亮[1,2],张　多[3]

(1.云南大学 云南省智能系统与计算重点实验室, 云南 昆明 650500;2.云南大学 信息学院, 云南 昆明 650500;3.云南大学 文学院, 云南 昆明 650091)

摘　要:民间文学是一种描述民众社会生活和思想文化的重要载体。由于自然、历史或人为等因素,记录民间文学的文本存在字词模糊不清、难以辨识甚至完全缺失等情况。为更好地研究、传播和传承民间文学,需要对不完整的民间文学文本作品进行修复。为补全不完整民间文学文本句子中的缺失字词信息,确保修复后句子与原不完整句子在内容和结构上保持一致,本文将民间文学文本修复视为一种可控文本生成任务。针对民间文学文本标记数据稀缺、存在特殊词汇和具有结构性特征等特点,考虑现有的可控文本生成方法面临的灾难性遗忘现象和处理民间文学领域数据时存在的泛化性不足等问题,提出了一种基于知识蒸馏的民间文学文本修复方法。首先,使用预训练语言模型和学生网络,获取民间文学文本中字符的基础特征向量,构建语义特征矩阵并对其进行知识蒸馏,保证学生网络与教师网络输出特征之间较小的分布差异,增强学生网络对句子整体语义的理解。然后,基于语义特征矩阵构建结构特征矩阵并对其进行知识蒸馏,以加强结构知识在学生网络参数更新过程中的约束。针对民间文学作品的3类典型体裁,构建了相应的数据集并对本文方法进行了实验测试,实验结果表明:修复结果句子的双语互译替代评估(BLEU)指标提升了1%～6%,困惑度(PPL)指标降低了15～300,验证了本文方法的有效性。

关键词:民间文学;文本修复;知识蒸馏;灾难性遗忘;结构知识

中图分类号:TP391　　**文献标志码:**A　　**文章编号:**2096-3246(2025)06-0119-12

民间文学是民众在日常文化生活中传承、传播、共享的口头传统和语言艺术,凝聚着民众世代相传的智慧、经验与情感,积淀着其最深层的精神追求[1]。保护、研究并传播民间文学,对激励民族文化自觉、提升民族文化自信具有重要意义。早期的民间文学作品主要依靠口耳相传,之后随着书面文本化记录手段的普及,口头文学逐渐以书面方式记录在民间抄本或搜集整理文本中。然而,由于自然、历史或人为等因素,记录民间文学的书面文本会产生破裂、腐朽及字迹褪色等各种损伤,导致文本中的某些字词变得模糊不清、难以辨识甚至完全缺失,从而使得作品出现情节不连贯、人物形象不具体、主题思想不清晰等问题,阻碍了民间文学的研究和传播。目前,不完整的民间文学文本作品主要依靠人工经验进行修复,这种方式效率低、成本高、不稳定,且无法满足数量日益增长、准确性要求不断提升的修复需求。修复任务示例如图1所示,输入带有缺失的不完整民间文学文本,预测生成不句子中的缺失字词,补全不完整民间文学文本句子,以确保文本信息的准确和完整,有助于促进民间文学的保护、传承和发展。

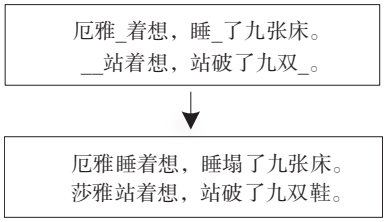


图1　民间文学文本修复任务示例

Fig. 1　Illustration of folk literature text restoration

民间文学文本作为一种垂直领域文本数据,与通用领域文本数据存在显著差异。首先,民间文学文本中广泛存在着通用领域文本中不存在或很少出现的

收稿日期:2024－01－30　　**修回日期:**2024－06－13　　**网络出版日期:**2024－07－01

基金项目:云南省基础研究计划面上项目(202201AT070394);云南省智能系统与计算重点实验室项目(202405AV340009);国家社会科学基金项目(20CZW059);云南大学研究生科研创新项目(KC－22221717)

作者简介:曹熊能(1996—),女,硕士. 研究方向:自然语言处理. E-mail: caoxiongneng@mail.ynu.edu.cn

* **通信作者:**王笳辉,助理研究员,E-mail: wjh@ynu.edu.cn

特殊词汇，如语句“厄沙苦思冥想，天天坐卧不安”中的“厄沙”。准确识别这些词汇的含义并判断其在句子中的作用，才能正确理解句子所表达的意义及语境[2]。对不完整句子所表达的意义及语境理解越充分，修复后的句子表达的语义也越准确、内容也越完整。其次，民间文学文本相较于通用领域文本结构性特征更显著、语句结构更对仗工整，如语句“盘古哪里去算命?庙中王家去算命”。准确提取不完整民间文学文本句子中的结构知识，并在修复过程中加强对结构知识的关注，是确保修复后结果符合民间文学文本句子格式与写作风格的关键。

目前，基于深度神经网络及语言模型的文本修复方法逐渐崭露头角，在不完整古铭文句子[3]及中医古籍修复[4]等方面取得了不错的进展。然而，由于民间文学其特有的语言特点的存在，使得现有方法难以直接扩展应用于民间文学文本修复领域。从文本空白错误识别并修订的角度看，民间文学文本修复可看作带有缺失的民间文学文本自动纠错任务。然而，针对空缺字符的自动纠错仍存在一定的局限性。

可控文本生成（CTG）的任务是指导语言模型生成满足特定主题、风格和情感等要求的文本内容[5]，是生成满足特定需求文本的惯用方法[6]。从缺失字符生成并补全的角度看，民间文学文本修复也可看作是在给定输入语句下的民间文学可控文本生成任务。近年来，自然语言处理领域逐渐形成“预训练+微调”的研究范式，基于预训练语言模型的文本生成技术也逐渐成为了 CTG 的主流方法[7]。预训练语言模型在预训练阶段获取了语义和语法等通用语言知识，极大提升了其对语言信息的表示能力及可控文本生成任务的效果。同时，现有的可控文本生成方法多针对通用领域数据[8–9]，难以在垂直领域数据上取得良好的泛化效果，不能直接用于民间文学文本修复。

此外，预训练语言模型多在新闻和百科等通用领域语料上进行预训练，面对垂直领域数据上的不同下游任务时，微调数据与预训练数据之间的差异性使得预训练语言模型需进行大量的参数调整，忘记学习到的通用语言知识，引起灾难性遗忘问题[10–11]。知识蒸馏（KD）[12]是一种基于“教师网络–学生网络”思想的训练方法，教师网络通过为学生网络提供正则化约束而实现对学生网络参数更新的约束，可有效防止学生网络产生灾难性遗忘[13]。

为使修复后句子符合民间文学文本的特点，并在内容和格式上与原句保持一致，从而控制生成的缺失字词满足特定内容、风格和结构等约束，同时，解决预训练语言模型在民间文学文本上面临的灾难性遗忘问题，提升模型有效性，本文在可控文本生成的基础上，研究基于知识蒸馏的民间文学文本修复方法，主要包括以下 3 个方面的工作：

1）针对预训练语言模型难以准确理解民间文学文本中特殊词汇的含义这一问题，拓展中间特征的知识蒸馏技术，利用预训练语言模型输出的字符基础特征向量既包含语义信息、也具有句法成分信息的特点，构建民间文学文本句子的语义特征矩阵。通过为学生网络提供语义特征矩阵的正则化约束，使学生网络在预训练语言模型的引导下有效更新民间文学文本句子中每个字符的基础特征向量，提升对其中特殊词汇的认知，从而增强对句子整体语义的理解，使修复后的句子语义准确、流畅且内容完整。

2）为提取并表示民间文学文本句子的结构知识，提出了一种结构知识表示方式，即基于预训练语言模型中隐藏的句子结构信息，用点积方式计算字符基础特征向量之间的结构相关性，并按句子语序拼接成句子的结构特征矩阵。为了提高学生网络对民间文学文本句子结构知识的关注，给出了一种结构特征知识蒸馏技术，使用结构特征矩阵为学生网络提供正则化约束，强化结构知识在学生网络参数更新过程中的影响，使修复结果满足民间文学文本句子格式要求。最后，利用学生网络完成民间文学文本修复任务。

3）针对 3 类典型的民间文学作品体裁，构建了相应的数据集，并在不同类型的预训练语言模型上进行了实验测试。

1 相关工作

1.1 基于自动纠错的文本修复

文本自动纠错指的是对文本内容中出现的错误进行自动识别和修正。Al–Sabahi 等[14]使用序列标记模型识别文本编辑，实现将不正确的句子修复为语法正确的句子。Pan 等[15]提出一种分层方法，实现对移动通信中的噪声信息进行检测和纠正。Lee 等[16]提出一种基于屏蔽语言模型的错字校正模型，修复医疗数据中的排版错误问题。文本自动纠错主要针对的是拼写错误、语法错误等文本错误，而对于句子中存在缺失字符的错误，尤其是句子结构或语义因缺失字符产生明显偏差时，难以进行有效的识别和修复。

1.2　基于可控文本生成的文本修复

Chan 等[17]通过改变原始预训练语言模型的体系结构，在生成式预训练变换器（GPT）[18]的基础上注入条件控制模块，用于在单词和短语级别上对生成文本实现更为精确的控制。He 等[19]通过再训练双向自回归变换器（BART）[20]以生成高质量和多样性的词汇约束文本，在编码器上添加一个标记级分类器，用于指示解码器替换和插入词汇，进而利用解码器并行预测所有令牌（token）。这类重构或再训练预训练语言模型的可控文本生成技术[21−22]需要大量的计算资源和标记数据来驱动语言模型的训练，例如，Wu 等[23]开发的 BloombergGPT 金融领域语言模型，依托于一个 3 630 亿个字符的金融标记数据集。然而，民间文学文本标记数据稀缺，语言模型无法得到充分训练，难以实现高质量的民间文学文本修复。

Dathathri 等[24]训练属性判别模型，判断预训练语言模型生成的文本是否符合约束要求。Pascual 等[25]将词汇表上的概率分布解码为语义上与约束词（主题或关键字）相似的词。Qin 等[26]使用基于能量的朗之万动力学（Langevin dynamics）约束解码，通过基于梯度的采样得到满足要求的文本。这类将预训练语言模型参数固定，仅在文本生成的解码阶段对生成的文本进行筛选的方法[27−28]，需要以预训练语言模型能够生成满足约束条件的高质量文本为前提，才能确保文本修复的效果，但现有预训练语言模型难以生成高质量民间文学文本句子，因此，无法基于这类方法实现民间文学文本修复。

Ribeiro 等[29]在预训练语言模型的编码器和解码器的前馈子层后添加结构感知适配器模块，并对图结构进行编码，减小控制属性和预训练语言模型间的差距，实现从抽象语义到文本的生成。Li 等[30]修改预训练语言模型的位置编码和注意力计算方式，使其能更好地感知文本生成控制信息，生成更流畅、更多样化和属性更相关的文本填充内容。这类仅通过增加模块和改变输入等方式微调预训练语言模型参数的方法[31]能够充分利用预训练语言模型在预训练阶段获取的通用语言知识，降低可控文本生成任务的难度，但预训练语言模型在垂直领域数据上微调时面临灾难性遗忘，需要对预训练语言模型的参数调整进行约束。

1.3　基于知识蒸馏的文本处理

目前，由于知识蒸馏在模型压缩和模型增强上显现出的优越特性，已逐渐被应用到阅读理解、命名实体识别和文本分类等文本处理任务中。例如，Hu 等[32]将知识从多个教师模型转移到单个模型，提高阅读理解任务的性能。Wu 等[33]利用标签丰富的语言来帮助其他标签贫乏的语言进行训练，实现单源或多源的跨语言命名实体识别任务。Yang 等[34]引入知识蒸馏和类感知经验重放两种策略，减轻连续文本分类任务中面临的灾难性遗忘问题。本文任务是补全不完整民间文学文本句子中的缺失字词信息，实现不完整民间文学文本作品的自动修复，与上述知识蒸馏在文本处理中的应用存在显著区别。

2　基本框架

2.1　任务定义

民间文学文本修复任务，旨在根据不完整的民间文学文本句子信息，得到对应的完整句子。给定一个字符缺失数量未知的不完整句子 $\widetilde{X}$，$\widetilde{X}=\left\{\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_d,\cdots,\hat{x}_i,\cdots,\hat{x}_j,\cdots,\tilde{x}_n\right\}$，其中，$\tilde{x}_d(1\le d\le n,d\ne i,j)$ 为未缺失字符，$\hat{x}_k(1\le k\le n,k$ 可为 i、$j)$ 为一个或多个连续缺失字符的占位符，n 为句子 $\widetilde{X}$ 的总字符数。学生网络以句子 $\widetilde{X}$ 为条件，生成在语义和格式上与句子 $\widetilde{X}$ 相似且语法合理的完整句子 $X=\left\{x_1,x_2,\cdots,x_e,\cdots,x_m\right\}$，$m$ 为句子 X 的总字符数，x_e 为句子 X 的字符（$1\le e\le m$）。值得注意的是，对于一个不完整民间文学文本句子，可使用多种修复方式。例如，可以只生成缺失字符、然后将其填充到不完整句子中的缺失位置，也可直接生成完整的句子。由于可能修复的方式并不唯一，本文将民间文学文本修复任务转换成条件概率 $P\left(X|\widetilde{X}\right)$ 的计算，用 $P\left(X|\widetilde{X}\right)$ 表示在给定缺失句子 $\widetilde{X}$ 时，生成完整句子为 X 的可能性。选择 $P\left(X|\widetilde{X}\right)$ 的值最大时对应的完整句子 X 作为最终的修复结果。

2.2　民间文学修复方法框架

知识蒸馏方法是从教师网络的输出结果中“蒸馏”出“知识”，并帮助训练学生网络。具体而言，给定学生网络和教师网络，首先使用调节参数和构建矩阵等方法表示并突出教师网络中的可用知识，显式地展示其相似性，然后根据所展现知识的形式采用不同的蒸馏损失函数训练学生网络[35−36]。本文提出一种基于联合知识蒸馏的民间文学文本修复方法，框架如图 2 所示，主要包含中间特征知识蒸馏和结构特征知识蒸馏两个模块。图 2 中，使用预训练语言模型作为教师网络，分别针对中间特征知识和结构特征知识提出两个不同的蒸馏损失函数，用于约束学生网络参数更新，提升学生网络性能，实现不完整民间文学文本句子修复。以上各模块将在第 3.1、3.2 和 3.3 节详细阐述。

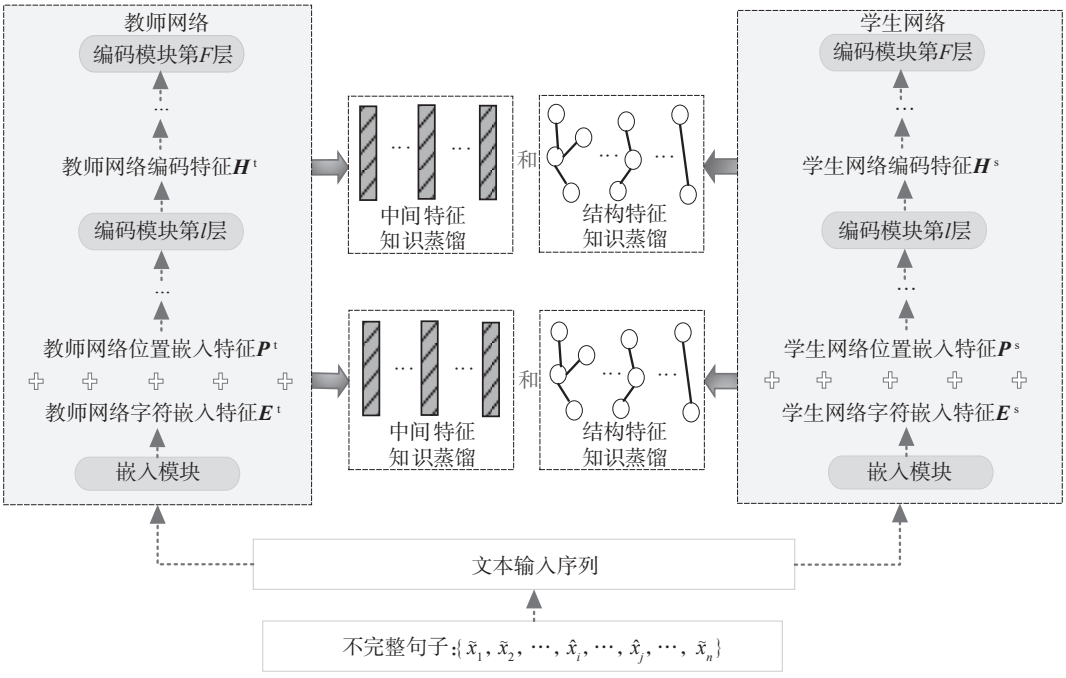


图2 基于联合知识蒸馏的民间文学文本修复模型框架

Fig. 2 Framework of folk literature text restoration model based on combined knowledge distillation

3 模型算法

3.1 中间特征知识蒸馏

给定不完整民间文学文本句子$\widetilde{X}$,首先,使用〈mask〉标记替换$\widetilde{X}$中缺失字符的占位符$\hat{x}_i$和$\hat{x}_j$,得到输入序列$\{\tilde{x}_1,\cdots,\langle\text{mask}\rangle,\cdots,\langle\text{mask}\rangle,\cdots,\tilde{x}_n\}$;然后,将其输入教师网络和学生网络的嵌入模块和编码模块,得到对应的字符嵌入特征$\boldsymbol{E}(\boldsymbol{E}=\{\boldsymbol{e}_1,\cdots,\boldsymbol{e}_i,\cdots,\boldsymbol{e}_j,\cdots,\boldsymbol{e}_n\})$、位置嵌入特征$\boldsymbol{P}(\boldsymbol{P}=\{\boldsymbol{p}_1,\cdots,\boldsymbol{p}_i,\cdots,\boldsymbol{p}_j,\cdots,\boldsymbol{p}_n\})$和编码特征集合$\boldsymbol{H}(\boldsymbol{H}=\{\boldsymbol{H}_1,\cdots,\boldsymbol{H}_l,\cdots,\boldsymbol{H}_F\})$,其中,$F$为教师网络和学生网络中编码模块的总层数,$\boldsymbol{e}_n$和$\boldsymbol{p}_n$为最后一个字符对应的字符嵌入特征向量和位置嵌入特征向量,$\boldsymbol{H}_l$为编码模块第l层输出的编码特征,$\boldsymbol{H}_l=\{\boldsymbol{h}_{1,l},\cdots,\boldsymbol{h}_{i,l},\cdots,\boldsymbol{h}_{j,l},\cdots,\boldsymbol{h}_{n,l}\}$,$\boldsymbol{h}_{n,l}$为编码模块第$l$层输出的编码特征向量。本文将嵌入模块输出的字符嵌入特征$\boldsymbol{E}$和位置嵌入特征$\boldsymbol{P}$进行对位相加,作为句子的第0层语义特征矩阵$\boldsymbol{H}_0$,计算公式如下:

$$\boldsymbol{H}_0=\boldsymbol{E}+\boldsymbol{P}=\{\boldsymbol{h}_{1,0},\cdots,\boldsymbol{h}_{i,0},\cdots,\boldsymbol{h}_{j,0},\cdots,\boldsymbol{h}_{n,0}\} \tag{1}$$

编码模块输出的编码特征表示,能够较好地捕捉民间文学文本句子中的语义与词性等语言信息,因此,直接将编码模块每一层输出的编码特征作为民间文学文本句子的每一层语义特征矩阵。为了使学生网络获得民间文学文本句子中字符的高质量基础特征向量、增强其对民间文学文本句子语义的理解,本文基于语义特征矩阵进行中间特征的知识蒸馏,计算教师网络和学生网络中每一层语义特征矩阵之间的平滑L1(SmoothL1)损失,利用教师网络所理解字符的通用知识进一步提升学生网络对民间文学文本句子中字符的认识,计算式如下:

$$\mathcal{L}_{\text{fkd}}^{l}=\begin{cases}\dfrac{1}{2}\left(\boldsymbol{H}_l^{\text{t}}-\boldsymbol{H}_l^{\text{s}}\right)^2, \left|\boldsymbol{H}_l^{\text{t}}-\boldsymbol{H}_l^{\text{s}}\right|<1;\\ \left(\boldsymbol{H}_l^{\text{t}}-\boldsymbol{H}_l^{\text{s}}\right)-\dfrac{1}{2}, \left|\boldsymbol{H}_l^{\text{t}}-\boldsymbol{H}_l^{\text{s}}\right|\geq 1\end{cases} \tag{2}$$

式中,$\mathcal{L}_{\text{fkd}}^{l}$为对应的L1损失,$\boldsymbol{H}_l^{\text{t}}$和$\boldsymbol{H}_l^{\text{s}}$分别是民间文学文本句子在教师网络和学生网络中的第$l$层语义特征矩阵。

将每一层SmoothL1损失相加,得到最终的中间特征知识蒸馏损失$\mathcal{L}_{\text{fkd}}$,计算公式如下:

$$\mathcal{L}_{\text{fkd}}=\sum_{l=0}^{F}\mathcal{L}_{\text{fkd}}^{l} \tag{3}$$

3.2 结构特征知识蒸馏

针对民间文学文本语句结构对仗工整的特点,本文基于语义特征矩阵中的字符基础特征向量,将其间的结构性关联关系作为民间文学文本句子的结构知识,并研究其表示方法。给定民间文学文本句子的第l层语义特征矩阵$\boldsymbol{H}_l$,首先,计算$\boldsymbol{H}_l$中每个字符基础特征向量之间的点积,并用向量之间的点积表达字符之间的结构相关性,其中,$\boldsymbol{H}_l$中第i个字符特征向量$\boldsymbol{h}_i^l$和第j个字符特征向量$\boldsymbol{h}_j^l$的点积ρ_{ij}^l的计算式如下:

$$\rho_{ij}^{l}=\text{sigmoid}\left(\frac{\boldsymbol{h}_i^l\boldsymbol{h}_j^{l\,\text{T}}}{\sqrt{d}}\right) \tag{4}$$

式中，d为$\boldsymbol{h}_i^l$和$\boldsymbol{h}_j^l$的维度，sigmoid(·)为激活函数。

然后，按照文本输入序列中字符的顺序，将获得的点积计算结果拼接成结构特征矩阵$\boldsymbol{F}_l$，以表示民间文学文本句子中字符之间的结构性关联关系。由于ρ_{ij}^l和ρ_{ji}^l表达的是相同字符之间的结构相关性，故$\boldsymbol{F}_l$只需为上三角矩阵，计算公式如下：

$$\boldsymbol{F}_l=\begin{bmatrix}\rho_{11}^l & \cdots & \rho_{1n}^l\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \rho_{nn}^l\end{bmatrix} \tag{5}$$

为了提高学生网络对民间文学文本句子结构知识的关注，使生成的完整句子符合民间文学文本句子的格式要求，本文对结构特征矩阵进行知识蒸馏，最小化教师网络和学生网络中每一层结构特征矩阵之间的差距，以强化结构知识在学生网络参数更新过程中的影响。考虑到字符基础特征向量之间的点积不能完全等同于字符之间的结构相关性，为了减少其间的偏差，本文不强制学生网络与教师网络的结构特征矩阵完全匹配，只要求其间的差异小于阈值$\Delta(\Delta\in[0,1])$，损失的计算式如下：

$$\mathcal{L}_{\text{skd}}^l=\max\left(\left|\boldsymbol{F}_l^{\text{t}}-\boldsymbol{F}_l^{\text{s}}\right|,\Delta\right) \tag{6}$$

式中，$\boldsymbol{F}_l^{\text{t}}$和$\boldsymbol{F}_l^{\text{s}}$分别为民间文学文本句子在教师网络和学生网络中第$l$层结构特征矩阵。

将每一层结构特征矩阵的损失相加得到如下的最终结构特征知识蒸馏损失：

$$\mathcal{L}_{\text{skd}}=\sum_{l=0}^{F}\mathcal{L}_{\text{skd}}^l \tag{7}$$

3.3　模型训练

民间文学文本句子中的缺失字符位置主要可分为非初始位置缺失、初始位置缺失和混合缺失这3类，分别对应预训练语言模型的自编码器(AE)模型[37]、自回归(AR)模型[18]和序列到序列(Seq2Seq)模型[20]，因此，本文基于这3类预训练语言模型评估联合知识蒸馏方法的修复效果。为降低学生网络对民间文学文本修复任务的敏感性，提高其收敛速度，本文基于教师网络使用的预训练任务，采用不同的修复方式，即基于不同的方式计算条件概率$P\left(X|\widetilde{X}\right)$。

AE模型的预训练任务为掩码语言模型(MLM)和下句预测(NSP)，因此，本文对AE模型采用填充式的修复方式。首先，基于不完整句子$\widetilde{X}$中未缺失字符的信息，独立地生成填充字符x_k；然后，将填充字符x_k填充到不完整句子$\widetilde{X}$相应的缺失位置，得到完整句子X，即独立地计算每个不同位置的缺失字符概率；进一步累乘作为最终完整句子X的概率$P\left(X|\widetilde{X}\right)$，计算式如下：

$$P\left(X|\widetilde{X}\right)=\prod_{k=\{\hat{x}_k\in B\}}P\left(x_k|\widetilde{X}_{\setminus B}\right) \tag{8}$$

式中：B为所有缺失字符的集合，$B=\{\hat{x}_k\}$；$\widetilde{X}_{\setminus B}$为在句子$\widetilde{X}$中但不在集合$B$中的剩余字符的集合，$\setminus$为集合减法运算。

AR模型的预训练任务是标准语言建模任务，因此，本文对AR模型采用单向生成式的修复方式。根据给定字符的上文计算后一字符的生成概率，从前往后逐步生成所有字符直至终止，得到最终生成的句子X。在给定前序生成字符序列的情况下，生成句子X的概率$P\left(X|\widetilde{X}\right)$计算式如下：

$$P\left(X|\widetilde{X}\right)=\prod_{i=1}^{n}P\left(x_i|\tilde{x}_1,\tilde{x}_2,\cdots,\tilde{x}_{i-1}\right) \tag{9}$$

Seq2Seq模型的预训练任务基于编码器–解码器框架，因此，本文对Seq2Seq模型采用编码–解码生成式的修复方式。首先，使用编码器对不完整句子$\widetilde{X}$进行编码；然后，结合编码器表示结果和当前已生成字符，使用解码器从前往后逐步解码，得到最终的完整生成句子X。在给定编码器结果和前序位置生成字符的条件下，生成句子X的概率$P\left(X|\widetilde{X}\right)$，计算式如下：

$$P\left(X|\widetilde{X}\right)=\prod_{i=1}^{n}P\left(x_i|x_1,x_2,\cdots,x_{i-1},\widetilde{X}\right) \tag{10}$$

考虑将上述3种不同类型的预训练语言模型作为模型的骨干网络，使用完整句子X中生成字符的交叉熵损失$\mathcal{L}_{\text{ce}}$、中间特征知识蒸馏损失与结构特征知识蒸馏损失共同训练学生网络，整体损失函数如下：

$$\mathcal{L}=\mathcal{L}_{\text{ce}}+\lambda_1\mathcal{L}_{\text{fkd}}+\lambda_2\mathcal{L}_{\text{skd}} \tag{11}$$

式中，$\lambda_1(0<\lambda_1<1)$和$\lambda_2(0<\lambda_2<1)$分别为中间特征蒸馏目标和结构特征知识蒸馏目标对应的损失函数权重。

保持教师网络参数固定不变，学生网络训练的步骤见算法1。

算法1　学生网络训练

输入：不完整民间文学文本句子$\widetilde{X}$

输出：完整民间文学文本句子X

1: 用教师网络参数初始化学生网络

2: For i=1 to n do

3:　基于嵌入模块得到$\widetilde{X}$的字符嵌入特征表示$\boldsymbol{E}^{\text{t}}$和$\boldsymbol{E}^{\text{s}}$

4:　基于嵌入模块得到$\widetilde{X}$的位置嵌入特征表示$\boldsymbol{P}^{\text{t}}$和$\boldsymbol{P}^{\text{s}}$

5:　使用式(1)得到第0层语义特征矩阵$\boldsymbol{H}_0^{\text{t}}$和$\boldsymbol{H}_0^{\text{s}}$

6:　得到其余层语义特征矩阵$\boldsymbol{H}_l^{\text{t}}$和$\boldsymbol{H}_l^{\text{s}}$

7:　使用式(2)～(3)计算中间特征知识蒸馏损失$\mathcal{L}_{\text{fkd}}$

8:　使用式(4)～(5)构建结构特征矩阵$\boldsymbol{F}_l^{\text{t}}$和$\boldsymbol{F}_l^{\text{s}}$

9:　使用式(6)～(7)计算结构特征知识蒸馏损失$\mathcal{L}_{\text{skd}}$

10: 基于式(8)～(10)得到完整民间文学文本句子X

11: 使用式(11)计算损失$\mathcal{L}$

12: 使用$\mathcal{L}$的梯度信息更新学生网络参数

13: End For

4 实验

4.1 实验设置

1) 数据集。民间文学从体裁上可分为韵文作品、散文作品和说唱作品3类,因此,本文分别选用《云南少数民族古典史诗》《傣族民间故事选》和《中国相声作品集》3个民间文学文本,这3个民间文学文本摘要信息如表1所示。

表1 民间文学文本摘要信息

Tab. 1 Information of folk literature text

民间文学文本	体裁	文章篇数	总句子数
云南少数民族古典史诗	韵文	46	44 939
傣族民间故事选	散文	50	4 848
中国相声作品集	说唱	30	5 432

将上述民间文学文本中的文章按照句号和感叹号等标点符号划分成句子,对句子中的字词进行随机掩码处理,构建了3个面向民间文学文本修复任务的数据集,其中,《云南少数民族古典史诗》数据集平均句子长度为24.1个字,《傣族民间故事选》数据集平均句子长度为32.2个字,《中国相声作品集》数据集平均句子长度为31.5个字。

2) 评价指标。本文使用自然语言生成(NLG)任务的常用评价指标:双语互译替代评估(BLEU)和困惑度(PPL)评估本文方法的有效性。BLEU[38]是一种基于精确度的相似度量方法,通过测量生成句子和参考句子之间n–gram字词(对连续n个字词的分析)的重叠度来推断字词使用的准确性,由此衡量生成句子的准确度,BLEU分数越高代表句子准确度越高,记为S_{BLEU};PPL是一种基于信息理论的度量方法,用于衡量生成句子的流畅度,PPL值越高代表句子流畅度越低,记为S_{PPL}。

3) 对比模型。本文选择BERT[37]、GPT[18]和BART[20]3类预训练语言模型做为代表,采用传统微调方法和本文提出的联合知识蒸馏方法对语言模型参数进行调整,以测试本文方法的有效性。

4) 模型参数。学习率(learning rate)设为1×10^{-5},批处理大小(batch size)设为16,最大序列长度设为512,使用AdamW优化器,式(6)中的边际值Δ取值为0.1,式(11)中的λ_1和λ_2取值为0.1。

5) 实验环境。使用Python 3.7编程语言,PyTorch1.8.1深度学习框架,CPU和GPU分别为Intel i9–10900X和NVIDIA GeForce RTX 3090(显存为24 GB),内存为128 GB。

4.2 对比实验

针对不同的预训练语言模型类型,在3种不同类型的民间文学文本数据集上进行了对比实验。在联合知识蒸馏方法中,AE选用bert–base–chinese为教师网络,AR选用gpt2–chinese–cluecorpussmall为教师网络,Seq2Seq选用bart–base–chinese为教师网络,实验结果如表2所示,其中,下标FT表示在预训练语言模型上使用传统的微调方法,下标KD表示使用本文提出的联合知识蒸馏方法,S_{BLEU1}和S_{BLEU4}分别为1–gram和4–gram字词的重叠值,1–gram是对单个词的分析,4–gram是对连续4个词的分析,A_{BLEU}是从1–gram到4–gram字词的平均重叠值,能够更全面、准确地衡量生成句子的准确度,综合考虑了不同n–gram级别的匹配情况,可以看出:

1) 在《云南少数民族古典史诗》《傣族民间故事选》和《中国相声作品集》构建的民间文学文本数据集上,BERT模型的A_{BLEU}值分别提升了0.12%、0.80%和0.29%,S_{PPL}值分别降低了146.07、168.80和72.52。

2) 在《云南少数民族古典史诗》《傣族民间故事选》和《中国相声作品集》构建的民间文学文本数据集上,GPT模型的A_{BLEU}值分别提升了1.00%、1.28%和0.66%,S_{PPL}值分别降低了233.25、303.39和144.96。

表2 对比实验结果

Tab. 2 Results of comparative experiments

模型	方法类型	云南少数民族古典史诗				傣族民间故事选				中国相声作品集			
		S_{PPL}	S_{BLEU1}	S_{BLEU4}	A_{BLEU}	S_{PPL}	S_{BLEU1}	S_{BLEU4}	A_{BLEU}	S_{PPL}	S_{BLEU1}	S_{BLEU4}	A_{BLEU}
AE	$\mathrm{BERT_{FT}}$	288.62	0.877 8	0.694 4	0.778 4	464.11	0.873 2	0.680 7	0.769 6	141.97	0.835 3	0.606 8	0.710 5
	$\mathrm{BERT_{KD}}$	142.55	0.879 2	0.6957	0.779 6	295.31	0.877 8	0.690 9	0.777 6	69.45	0.835 9	0.609 8	0.713 4
AR	$\mathrm{GPT_{FT}}$	269.90	0.805 7	0.592 8	0.689 9	342.71	0.743 9	0.510 4	0.615 3	160.47	0.698 1	0.453 1	0.560 1
	$\mathrm{GPT_{KD}}$	36.65	0.816 5	0.602 6	0.699 9	39.32	0.760 5	0.519 8	0.628 1	15.51	0.707 9	0.460 5	0.566 7
Seq2Seq	$\mathrm{BART_{FT}}$	41.33	0.810 5	0.546 5	0.665 4	13.02	0.827 4	0.583 8	0.695 0	18.21	0.747 9	0.492 9	0.604 3
	$\mathrm{BART_{KD}}$	2.58	0.833 6	0.635 9	0.727 3	5.54	0.852 6	0.677 1	0.759 1	3.39	0.835 1	0.626 4	0.721 0

3）BART 在《云南少数民族古典史诗》《傣族民间故事选》和《中国相声作品集》构建的民间文学文本数据集上，A_{BLEU} 值分别提升了 6.19%、6.41% 和 11.67%，S_{PPL} 值分别降低了 38.75、7.48 和 14.82。

综上所述，在 3 种不同类型的民间文学文本数据集上，本文提出的联合知识蒸馏方法在指标上都优于传统的微调方法。A_{BLEU} 指标提升了 1.00%～6.00%，S_{PPL} 指标降低了 15～300，验证了联合知识蒸馏方法的有效性。其中，在 BART 模型上的 S_{PPL} 指标降低幅度最小，A_{BLEU} 指标提升幅度最大，这是因为 Seq2Seq 模型基于不完整民间文学文本句子中所有的字符，解码生成完整民间文学文本句子，相较于 BART 模型和 GPT 模型，中间特征知识蒸馏能够捕捉的新的语义与词性等语言信息有限，进而对提升民间文学文本句子中字符的基础特征向量表示的能力有限，表现为 S_{PPL} 指标降低幅度较小，但在高质量基础特征向量表示的条件下，结构特征知识蒸馏能够表达字符之间更多的结构性关联关系，从而使得 A_{BLEU} 指标得到显著提升。在 GPT 上的 S_{PPL} 指标降低幅度最显著，A_{BLEU} 指标提升幅度超过 BERT，这是因为 AR 模型基于不完整民间文学文本句子中缺失字符的前文信息，解码生成完整民间文学文本句子，联合知识蒸馏方法让学生网络不仅获得了缺失字符的信息，同时也获得了缺失字符的后文信息，相较于 BART 模型，捕捉了更多的语义与词性等语言信息，进而能为学生网络提供更多的语义和结构知识。

4.3　消融实验

为验证中间特征知识蒸馏和结构特征知识蒸馏对民间文学文本修复任务的影响，本文进行了消融实验，结果如表 3 所示，其中，w/F 代表消除了结构特征知识蒸馏，w/S 代表消除了中间特征知识蒸馏。可以看出：

1）消除了中间特征知识蒸馏模型的 A_{BLEU} 指标比消除了结构特征知识蒸馏模型平均高出 0.30%，表明结构特征知识蒸馏相较于中间特征知识蒸馏，对提高民间文学文本句子的准确性有更好的效果。

2）消除了结构特征知识蒸馏模型的 S_{PPL} 指标比消除了中间特征知识蒸馏模型的 S_{PPL} 指标平均低 22，表明中间特征知识蒸馏相较于结构特征知识蒸馏，对提高民间文学文本句子的流畅性有更好的效果。

综上所述，消融实验结果验证了每种知识蒸馏方式对提高民间文学文本修复任务的积极效果，且联合这两种蒸馏方式能进一步提升修复效果。

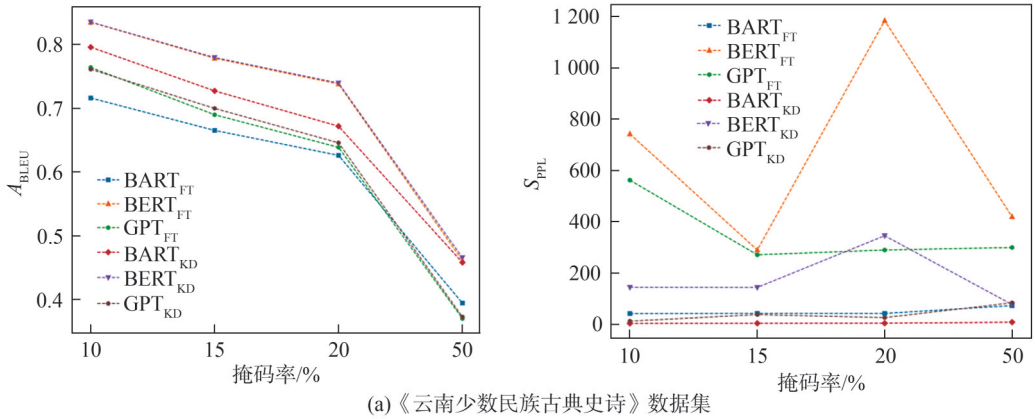
表 3　消融实验结果

Tab. 3　Results of ablation experiments

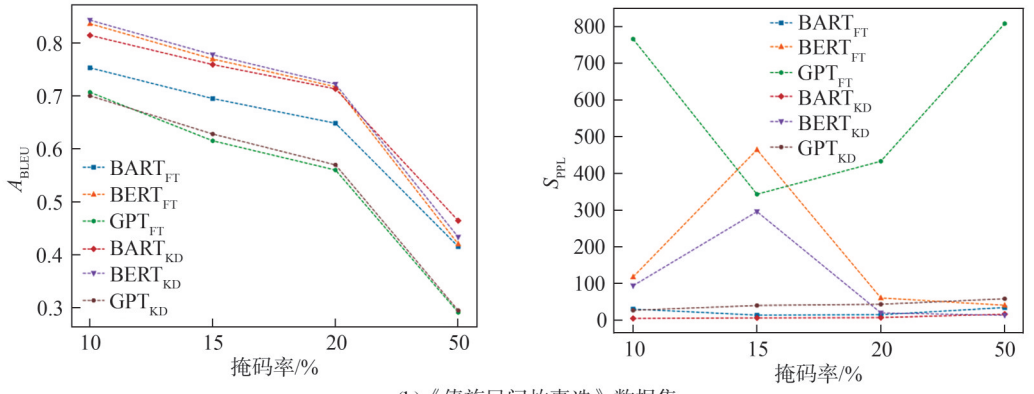
模型	方法类型	数据集为云南少数民族古典史诗				数据集为傣族民间故事选				数据集为中国相声作品集			
		S_{PPL}	S_{BLEU1}	S_{BLEU4}	A_{BLEU}	S_{PPL}	S_{BLEU1}	S_{BLEU4}	A_{BLEU}	S_{PPL}	S_{BLEU1}	S_{BLEU4}	A_{BLEU}
AE	BERT(w/F)	203.26	0.878 7	0.695 7	0.779 4	228.52	0.876 4	0.685 1	0.773 7	24.60	0.836 5	0.609 4	0.712 4
	BERT(w/S)	276.66	0.879 2	0.695 5	0.779 4	556.27	0.877 2	0.690 1	0.776 6	75.49	0.835 9	0.607 2	0.711 3
	$BERT_{KD}$	142.55	0.879 2	0.695 7	0.779 6	295.31	0.877 8	0.690 9	0.777 6	69.45	0.835 9	0.609 8	0.713 4
AR	GPT(w/F)	40.37	0.811 6	0.596 6	0.694 2	102.96	0.743 3	0.502 5	0.609 7	125.11	0.693 2	0.447 5	0.553 4
	GPT(w/S)	15.23	0.812 7	0.598 6	0.696 1	46.42	0.746 3	0.503 9	0.612 4	49.54	0.689 3	0.447 1	0.552 1
	GPT_{KD}	36.65	0.816 5	0.602 6	0.699 9	39.32	0.760 5	0.519 8	0.628 1	15.51	0.707 9	0.460 5	0.566 7
Seq2Seq	BART(w/F)	2.81	0.810 3	0.609 9	0.702 3	6.86	0.854 2	0.675 6	0.758 9	2.69	0.822 3	0.614 8	0.708 4
	BART(w/S)	3.88	0.828 8	0.628 0	0.720 7	6.32	0.849 7	0.671 3	0.754 3	3.43	0.827 5	0.624 9	0.716 8
	$BART_{KD}$	2.58	0.833 6	0.635 9	0.727 3	5.54	0.852 6	0.677 1	0.759 1	3.39	0.835 1	0.626 0	0.627 0

4.4　掩码率的影响

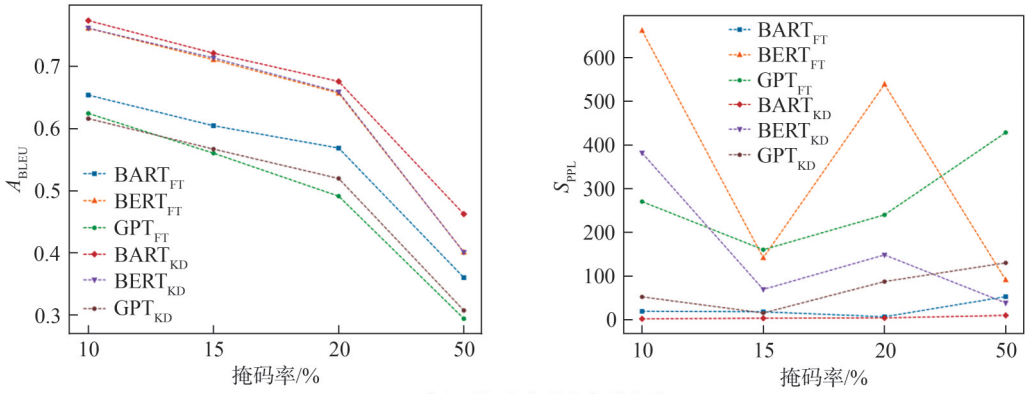
为验证联合知识蒸馏方法的鲁棒性，在 3 种不同类型的民间文学文本数据集上，对民间文学文本句子随机掩码 10%、15%、20% 和 50%，并针对不同的预训练语言模型类型进行了测试，结果如图 3 所示。由图 3 可以看出，在不同的掩码率下，联合知识蒸馏方法相对于基于传统微调方法，A_{BLEU} 指标都有所提高，S_{PPL} 指标都有所降低，说明了联合知识蒸馏方法可有效增强学生网络对句子整体语义和结构的理解，提升修复效果，具有较好的鲁棒性。同时可以看出，当掩码比例为 15% 或 20% 时，A_{BLEU} 指标和 S_{PPL} 指标通常呈现最优的提升效果，说明了民间文学文本句子中缺失字符的数量适中时，联合知识蒸馏方法能更有效地从不完整民间文学文本句子中捕捉关键的语言信息，为学生网络提供丰富且准确的语义和结构知识。



(a)《云南少数民族古典史诗》数据集



(b)《傣族民间故事选》数据集



(c)《中国相声作品集》数据集

图3 掩码率对联合知识蒸馏方法性能的影响

Fig. 3 Impacts of mask rate on the combined knowledge distillation method

4.5 案例分析

本文通过民间文学文本句子修复案例展示联合知识蒸馏方法的执行结果、并验证其有效性，具体的案例修复结果如表4所示。以掩码率为10%的情形为例，可以看出：

1）在填补句子“[MASK]来主持”时，BERT$_{KD}$模型得到的“共同”相较于BERT$_{FT}$模型得到的“杜自”，用词更准确，使修复后的句子更连贯。

2）GPT$_{FT}$模型生成的句子没有表达“来主持”的内容，而GPT$_{KD}$模型却生成了“下凡来主持”，使修复后的句子内容表达更完整。

3）BART$_{KD}$模型得到的“亲自来主持”相较于BERT$_{FT}$模型得到的“他来主持”，更符合民间文学文本句子的结构规范。

类似地，掩码率为50%时，使用不同预训练语言模型作为骨干网络，本文提出的方法均能更有效地修复民间文学文本句子。

从上述案例分析结果可以看出，本文提出的联合知识蒸馏方法可有效增强学生模型对民间文学文本句子整体语义的理解，也强化了结构知识在学生网络参数更新过程中的影响，可修复出语义连贯、内容完整和格式规范的句子。

表4　民间文学文本修复示例

Tab. 4　Examples of folk text restoration result

模型	掩码率为10%的结果	掩码率为50%的结果
缺失句子	天庭六[MASK]，其中有一位，下凡[MASK]人间，[MASK]来主持，带领沙生妹，带领都阿木，共同来献酒。	[MASK]六毕摩，其中有一位[MASK]到人间[MASK]来主持[MASK]带领[MASK]带领都阿木[MASK]来[MASK]酒。
原句子	天庭六[毕摩]，其中有一位，下凡[到]人间，[婚礼]来主持，带领沙生妹，带领都阿木，共同来献酒。	[天庭]六毕摩，其中有一位[，下凡]到人间[，婚礼]来主持[，]带领[沙生妹，]带领都阿木[，共同]来[献]酒。
$BERT_{FT}$	天庭六[姊弟]，其中有一位，下凡[到]人间，[杜自]来主持，带领沙生妹，带领都阿木，共同来献酒。	[圣时]六毕摩，其中有一位[，灵天]到人间[派阿玛]来主持[，]带领[众阿帮]，带领都阿木[，摆生]来[酒]酒。
$BERT_{KD}$	天庭六[兄弟]，其中有一位，下凡[到]人间，[共同]来主持，带领沙生妹，带领都阿木，共同来献酒。	[这位]六毕摩，其中有一位[，领人]到人间[，当天]来主持[，]带领[众阿木]，带领都阿木[，敬敬]来[酒]酒。
GPT_{FT}	天庭六[位神]，其中有一位，下凡[到]人间，带领沙生妹，带领都阿木，共同来献酒。	[捧腊哈纳罗神]，其中有一位[，名叫纳里本腊，他派]到人间[，]来主持[人间的事情，带领百姓送回酒。]
GPT_{KD}	天庭六[位神]，其中有一位，下凡[来主持]，带领沙生妹，带领都阿木，共同来献酒。	[神叫]六毕摩，其中有一位[，]到人间[去主持，]带领[人马来主持，]带领都阿木，来[吃]酒。
$BART_{FT}$	天庭六[大神]，其中有一位，下凡[到]人间，[他]来主持，带领沙生妹，[和]领都阿木，共同来献酒。	[有]六毕摩，其中有一位[是]到人间[来]来主持[并]带领[他]带领都阿木[来]来[喝]酒。
$BART_{KD}$	天庭六[神仙]，其中有一位，下凡[到]人间，[亲自]来主持，带领沙生妹，带[着]都阿木，共同来献酒。	[南古蒂提拉和]六毕摩，其中有一位[，]到人间[去，]来主持[祭祀，]带领都阿木[，]来[敬]酒。

5　结　论

本文提出了一种基于联合知识蒸馏的民间文学文本修复方法，通过拓展中间特征知识蒸馏，引入结构特征知识蒸馏，有效缓解了预训练语言模型在民间文学文本修复任务中面临的灾难性遗忘，实现了不完整民间文学文本句子自动修复。本文的方法既能使修复结果句子语义准确、流畅且内容完整，又能让其满足民间文学文本句子的格式要求，实验结果验证了本文方法的有效性。本文提出的民间文学文本修复方法，为民间文学研究、推动民间文学的保护和传承提供了技术支撑。未来将进一步研究民间文学文本缺失文档修复方法，考虑整篇文档中的全局信息，以得到更高质量的民间文学文本修复结果。

参考文献：

[1] Zhang Yixin.Excavating the times value of folk literature[N]. Social Sciences in China,2022–06–06(2).[张译心．挖掘民间文学的时代价值[N]. 中国社会科学报,2022–06–06(2).]

[2] Tang Xianlun,Lin Wenxing,Du Yiming,et al.Short text feature extraction and classification based on serial-parallel convolutional gated recurrent neural network[J].Advanced Engineering Sciences,2019,51(4):125–132.[唐贤伦,林文星,杜一铭,等．基于串并行卷积门阀循环神经网络的短文本特征提取与分类[J]. 工程科学与技术,2019,51(4):125–132.]

[3] Assael Y, Sommerschield T, Prag J. Restoring ancient text using deep learning:A case study on Greek epigraphy[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing Stroudsburg:ACL,2019:6367–6374.

[4] Sheng Wei, Lu Yanjie, Liu Wei, et al. Research on restoration of missing texts in ancient Chinese medicine books based on deep learning[J].Chinese Journal of Medical Library and Information Science,2022,31(8):1–7. [盛威,卢彦杰,刘伟,等．基于深度学习的中医古籍缺失文本修复研究[J]. 中华医学图书情报杂志,2022,31(8):1–7.]

[5] Prabhumoye S,Black A W,Salakhutdinov R.Exploring controllable text generation techniques[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: International Committee on Computational Linguistics,2020:1–14.

[6] Zhang Xuliang.Research and Application of Content-controlled Text Generation[D].Chengdu:University of Electronic Science and Technology of China,2023.[张绪亮．内容可控的文本生成研究与应用[D]. 成都:电子科技大学,2023.]

[7] Zhang Hanqing, Song Haolin, Li Shaoyu, et al. A survey of controllable text generation using transformer-based pre-trained language models[J].ACM Computing Surveys,2023,56(3):1–37.

[8] Li Jinpeng, Zhang Chuang, Chen Xiaojun, et al. Survey on automatic text summarization[J]. Journal of Computer Research and Development,2021,58(1):1–21.[李金鹏,张闯,陈小军,等．自动文本摘要研究综述[J]. 计算机研究与发展,2021,58(1):1–21.]

[9] Shen T, Quach V, Barzilay R, et al.Blank language model [C]//Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing.Stroudsburg:

ACL,2020:5186–5198.

[10] Goodfellow I J,Mirza M,Xiao Da,et al.An empirical investigation of catastrophic forgetting in gradient-based neural networks[EB/OL].(2015–03–04)[2024–01–01].https://arxiv.org/abs/1312.6211.

[11] Yang Jiacheng, Wang Mingxuan, Zhou Hao, et al. Towards making the most of BERT in neural machine translation [J]. Proceedings of the AAAI Conference on Artificial Intelligence,2020,34(5):9378–9385.

[12] Huang Zhenhua, Yang Shunzhi, Lin Wei, et al. Knowledge distillation:A survey[J].Chinese Journal of Computers,2022, 45(3):624–653.[黄震华,杨顺志,林威,等. 知识蒸馏研究综述[J]. 计算机学报,2022,45(3):624–653.]

[13] Shmelkov K,Schmid C,Alahari K.Incremental learning of object detectors without catastrophic forgetting[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision.Venice:IEEE,2017:3420–3429.

[14] Al–Sabahi K,Yang Kang,Liu Wangwang,et al.Multi-head sequence tagging model for Grammatical Error Correction [J].Engineering Applications of Artificial Intelligence, 2024,133:108314.

[15] Pan Fayu, Cao Bin, Fan Jing. A multi-task learning framework for efficient grammatical error correction of textual messages in mobile communications[J]. EURASIP Journal on Wireless Communications and Networking,2022, 2022(1):99.

[16] Lee E B,Heo G E,Choi C M,et al.MLM-based typographical error correction of unstructured medical texts for named entity recognition[J].BMC Bioinformatics,2022,23 (1):486.

[17] Chan A,Ong Y S,Pung B,et al.CoCon:A self-supervised approach for controlled text generation[EB/OL].(2022–06–10) [2024–01–01].https://arxiv.org/abs/2006.03535

[18] Radford A,Wu J,Child R,et al.Language models are unsupervised multitask learners[J].OpenAI Blog,2019(8):1–9.

[19] He Xingwei. Parallel refinements for lexically constrained text generation with BART[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.Stroudsburg:ACL,2021:8653–8666.

[20] Lewis M, Liu Yinhan, Goyal N, et al.BART:Denoising sequence-to-sequence pre-training for natural language generation,translation,and comprehension[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.Stroudsburg:ACL,2020:7871–7880.

[21] Ma Tinghuai, Yu Xin, Rong Huan, et al. Cross-domain text generation method based on semantic conduction of intermediate domains[J].Journal of Computer Research and Development,2023,60(12):2844–2863.[马廷淮,于信,荣欢,等. 基于中间域语义传导的跨领域文本生成方法[J]. 计算机研究与发展,2023,60(12):2844–2863.]

[22] Hu Renfen,Li Shen,Zhu Yuchen,et al.Knowledge representation and sentence segmentation of ancient Chinese based on deep language models[J]. Journal of Chinese Information Processing,2021,35(4):8–15.[胡韧奋,李绅,诸雨辰,等. 基于深层语言模型的古汉语知识表示及自动断句研究[J]. 中文信息学报,2021,35(4):8–15.]

[23] Wu Shijie,Irsoy O,Lu S,et al.BloombergGPT:A large language model for finance[EB/OL].(2023–12–21)[2024–01–01].https://arxiv.org/abs/2303.17564.

[24] Dathathri S, Madotto A, Lan J, et al. Plug and play language models: A simple approach to controlled text generation[EB/OL]. (2020 – 03 – 03) [2024 – 01 – 01]. https://arxiv. org/abs/1912.02164.

[25] Pascual D,Egressy B,Meister C,et al.A plug-and-play method for controlled text generation[C]//Proceedings of the Findings of the Association for Computational Linguistics:EMNLP 2021.Stroudsburg:ACL,2021:3973–3997.

[26] Qin Lianhui, Welleck S, Khashabi D, et al. COLD decoding: Energy-based constrained text generation with Langevin dynamics[EB/OL].(2022–10–13)[2024–01–01].https://arxiv.org/abs/2202.11705.

[27] Anderson P,Fernando B,Johnson M,et al.Guided open vocabulary image captioning with constrained beam search [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.Stroudsburg:ACL, 2017:936–945.

[28] Holtzman A,Buys J,Du Li,et al.The curious case of neural text degeneration[C]//Proceedings of the 8th International Conference on Learning Representations.Washington DC: ICLR,2020.

[29] Ribeiro L F R,Zhang Yue,Gurevych I.Structural adapters in pretrained language models for AMR-to-text generation[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.Stroudsburg:ACL,2021:4269–4282.

[30] Li Dongyuan,You Jingyi,Funakoshi K,et al.A–TIP:Attribute-aware text infilling *via* pre-trained language model[C]//Proceedings of the International Conference on Computational Linguistics.Gyeongju:ICCL,2022:5657–5869.

[31] Yang Jinfeng, Liang Xiangui, Wang Liuan, et al. Controlled medical dialogue generation based on prompt[J]. Journal of Chinese Information Processing, 2023, 37(4): 118 – 125. [杨锦锋,梁先桂,王刘安,等. 基于 Prompt 策略的医疗对话生成[J]. 中文信息学报,2023,37(4):118–125.]

[32] Hu Minghao,Peng Yuxing,Wei Furu,et al.Attention-guided answer distillation for machine reading comprehension[C]//

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018:2077–2086.

[33] Wu Qianhui,Lin Zijia,Karlsson B,et al.Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg:ACL,2020:6505–6514.

[34] Yang Fengqin,Che Yinshu,Kang Mei,et al.Continual text classification based on knowledge distillation and class-aware experience replay[J].Knowledge and Information Systems,2023,65(10):3923–3944.

[35] Shao Renrong, Liu Yuang, Zhang Wei, et al. A survey of knowledge distillation in deep learning[J].Chinese Journal of Computers,2022,45(8):1638–1673.[邵仁荣,刘宇昂,张伟,等.深度学习中知识蒸馏研究综述[J].计算机学报,2022,45(8):1638–1673.

[36] Fang Weiwei,Chen Aifang,Meng Na,et al.Incremental deep learning method for object detection model based on knowledge distillation[J].Advanced Engineering Sciences,2022,54(6):59–66.[方维维,陈爱方,孟娜,等.基于知识蒸馏的目标检测模型增量深度学习方法[J].工程科学与技术,2022,54(6):59–66.]

[37] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics. Minnesota:ACL,2019:4171–4186.

[38] Papineni K,Roukos S,Ward T,et al.BLEU:A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics–ACL'02.Morristown:ACL,2001:311–318.

Text Restoration of Folk Literature Based on Knowledge Distillation

CAO Xiongneng[1,2], *WANG Jiahui*[1,2*], *YUE Kun*[1,2], *DUAN Liang*[1,2], *ZHANG Duo*[3]

(1.Yunnan Key Laboratory of Intelligent Systems and Computing, Yunnan University, Kunming 650500, China;
2.School of Information Science and Engineering, Yunnan University, Kunming 650500, China;
3.School of Chinese Language and Literature, Yunnan University, Kunming 650091, China)

Abstract:

Objective Folk literature serves as an important carrier for depicting the social life and cultural perspectives of the general public. Due to natural, historical, or human factors, the words in folk literature texts are often ambiguous, difficult to identify, or even completely missing. For effective research and dissemination, it is necessary to repair incomplete folk literature texts. A significant difference exists between folk literary text data and the pre-training data used during the pre-training phase of pre-trained language models. For example, differences occur in the form of specialized vocabularies and structural features. These differences lead to catastrophic forgetting when directly fine-tuning pre-trained language models, as the model must perform extensive parameter adjustments and can forget previously learned universal language knowledge. Avoiding catastrophic forgetting in pre-trained language models for this repair task and ensuring that the restored sentences align with the linguistic characteristics of folk literature are the two main challenges. A knowledge-distillation-based method for folk literature text restoration is proposed to address these issues.

Methods Considering the characteristics of limited annotated data, the presence of specialized vocabularies, and the structural nature of folk literary texts, this study adopted a pre-trained language model to expand knowledge distillation and train the student network, enabling the automatic restoration of incomplete folk literary sentences. First, the pre-trained language models and student networks were utilized to extract the basic feature vectors of characters from folk literary texts. These basic feature vectors were then utilized to construct semantic feature matrices, which underwent intermediate feature knowledge distillation. This process involved computing the SmoothL1 loss between the semantic feature matrices of each layer in the pre-trained language model and the student network, ensuring minimal distribution differences between the output features of the student network and the teacher network. The student network's comprehension of the overall semantic meaning of sentences was enhanced by leveraging the teacher network's understanding of character-level general knowledge. Then, the structural relationships among the basic feature vectors in the semantic feature matrix were treated as the structural knowledge of folk literary text sentences. A structural feature matrix was constructed and subjected to structural feature knowledge distillation to reinforce the constraints of structural knowledge during the parameter update process of the student network, enhancing the structural regularity of the repaired sentences.

Results and Discussions For the three typical genres of folk literature, the corresponding datasets were constructed, and experimental studies were conducted. In the comparative experiments, BERT applied to the constructed folk literary text datasets showed improvements in average bilingual evaluation understudy (BLEU) values by 0.12%, 0.80%, and 0.29%, and reductions in PPL (perplexity) values by 146.07, 168.80, and 72.52, respectively. GPT applied to the constructed folk literary text datasets showed improvements in average BLEU values by 1.00%, 1.28%, and 0.66%, and reductions in PPL values by 233.25, 303.39, and 144.96, respectively. BART applied to the constructed folk literary text datasets

demonstrated improvements in average BLEU values by 6.19%, 6.41%, and 11.67%, and reductions in PPL values by 38.75%, 7.48%, and 14.82%, proving the effectiveness of the proposed method. In the ablation experiments, the average BLEU of the w/S model was, on average, 0.3% higher than that of the w/F model, indicating that structural feature knowledge distillation has a better effect on improving the accuracy of folk literary text sentences compared to intermediate feature knowledge distillation. The PPL index was, on average, 22 higher, indicating that intermediate feature knowledge distillation has a better effect on improving the fluency of folk literary text sentences. The results of the ablation experiments also indicated that combining these two distillation methods further improved the average BLEU index and reduced the PPL index compared to the w/S model and w/F model. In the mask rate experiment, the combined knowledge distillation method showed improvements in average BLEU indices and reductions in PPL indices relative to traditional fine-tuning methods, demonstrating the robustness of the combined knowledge distillation method. In addition, when the mask ratio was set to 15% or 20%, the average BLEU and the PPL metrics typically demonstrated the most optimal improvement, indicating that the combined knowledge distillation method was more effective in capturing crucial linguistic information from incomplete folk literary text sentences when the number of missing characters was moderate, providing the student network with rich and accurate semantic and structural knowledge. The case study intuitively demonstrated the execution results of the combined knowledge distillation method, indicating that the method generated coherent, complete, and well-formatted sentences.

Conclusions Considering the specific vocabulary and structural features of folk literature, the catastrophic forgetting phenomenon faced by existing controllable text generation methods, and the insufficient generalization when handling data from the vertical domain of folk literature, a combined knowledge distillation method is proposed. This method constructs semantic and structural feature matrices and conducts knowledge distillation on both. Experimental results demonstrate that the method effectively prevents catastrophic forgetting in pre-trained language models and generates sentences with more accurate semantics, comprehensive content, and improved alignment with the formatting requirements of folk literature texts.

Key words: folk literature; text restoration; knowledge distillation; catastrophic forgetting; structural knowledge

（编辑　吴芝明）

引用格式: Cao Xiongneng,Wang Jiahui,Yue Kun,et al.Text restoration of folk literature based on knowledge distillation[J]. Advanced Engineering Sciences,2025,57(6):119–130.[曹熊能,王笳辉,岳昆,等.基于知识蒸馏的民间文学文本修复[J].工程科学与技术,2025,57(6):119–130.]