

文章编号: 1003-0077(2025)02-0131-12

基于词性对齐与依存关系的中文排比句生成方法

钟茂生, 刘 蕾, 吴如萍, 甘家其, 周新宇

(江西师范大学 计算机信息工程学院, 江西 南昌 330022)

摘 要: 排比句是一种常用的修辞手法, 其使用具有增强气势、强调突出、层次清晰的效果。排比句生成对于文本生成具有重要意义, 能够丰富文本生成的风格和形式, 提升教育、广告和文学创作的质量, 但目前暂无生成模型和公开的排比句语料库。为此, 该文在收集和构建排比句数据集的基础上, 结合排比句具有的语言学特征, 提出了一种基于词性对齐与依存关系的中文排比句生成模型, 称为 CPG-PosDep。模型从语言学出发, 首先用设定的分词、句内词序和句间位置三种特殊符号及随机采样策略对排比句进行全局标识, 并结合改进的 Transformer 注意力机制学习排比句的词性对齐特征, 然后使用 BERT 和注意模块将给定分句的依存关系信息融入模型中, 融合生成排比句。在排比句数据集上的实验表明, 模型能够生成与给定分句在对应位置词性一致、依存关系相同的通顺分句, 相比使用现有对联或诗歌生成模型生成的排比句, 该文所提模型生成的排比句质量更具优势。

关键词: 中文排比句生成; 词性对齐; 随机采样策略; 依存关系

中图分类号: TP391

文献标识码: A

Chinese Parallelism Generation Based on Part-of-Speech Alignment and Dependency Relation

ZHONG Maosheng, LIU Lei, WU Ruping, GAN Jiaqi, ZHOU Xinyu

(School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi 330022, China)

Abstract: Parallelism is a popular rhetorical device featured by its ability to enhance momentum, emphasize key points, and provide clear structure, and there is now no dedicated generation model or publicly available corpus for parallelism. This paper proposes a Chinese parallelism generation model named CPG-PosDep, developed based on the collection and construction of a parallelism dataset and considering the linguistic characteristics of parallelism. Following a linguistic approach, the model employs predefined segmentation markers, intra-sentence word order indicators, and inter-sentence position symbols, combined with a random sampling strategy to globally annotate parallelism. It integrates an enhanced Transformer attention mechanism to learn part-of-speech alignment features, and adopts BERT and attention mechanisms to incorporate dependency relation information from sentences into the model. Experiments indicate that the model can generate coherent sentences that maintain consistent part-of-speech and dependency relations with the given clauses, much better than those by existing couplet or poetry models.

Keywords: Chinese parallelism generation; part-of-speech alignment; random sampling strategy; dependency relation

收稿日期: 2024-07-05 定稿日期: 2024-08-23

基金项目: 国家自然科学基金(32460214, 62366022)

0 引言

排比句是一种常用的修辞手法,指把三个或三个以上意义相关或相近、结构相同或相似、语气相同的词组或分句并排在一起组成的句子^[1]。排比句因其节奏感强、强调突出、层次清晰的特点被广泛运用在散文、诗歌、政论和演讲等各类文体文本中。通过使用排比句,可以将多个相关或相对的事物展现在面前^[2],如“春天里,鲜花怒放,绿草如茵;夏日里,炎炎烈日,百鸟纷飞;秋季里,金黄稻田,丰收的喜悦;冬天里,银装素裹,大地寂寞”。

本文研究的排比句生成是指给定一个句子,生成与给定分句长度相同、结构相似、语义相关,能够与之构成排比关系的后续分句,是自然语言生成(Natural Language Generation, NLG)领域的一个重要分支。如给定分句“奋斗有目标”,我们期望机器能够生成“实干有引领;发力有方向。”等能够与给定分句构成排比关系的其他分句。排比句生成研究不仅有助于推动自然语言处理技术的发展,提高系统对语言的理解和生成能力,还能够丰富自然语言生成的风格和形式,为文学创作等下游应用提供更多的可能性和创新思路,具有深入探讨和研究的价值。

目前关于排比句生成任务的研究相对空缺,缺乏公开发表的排比句训练数据集。为此,本文收集并构建了一个含有 30 452 组排比句的数据集^①,称为 Parallelisms Datasets。同时,文本生成领域暂无专门针对排比句生成模型的研究,与排比句生成相似的自然语言生成任务包括对联生成^[3]、诗歌生成^[4]等。排比句虽然在形式上与对联和诗歌相似,相比之下,排比句强调的是使用相同或相似的结构来突出一系列情感或观点,而诗歌和对联更加注重韵律格调。因此,诗歌和对联生成模型难以迁移到排比句任务中。

排比句不同于常规文本,它通常由一系列平行结构相似的词、词组或句子组成,以达到强调和表达的效果。常规文本则是指以表达信息、描述事件或讲述故事为主要目的的一般句子。在常规文本中,语句通常较为简洁直接,注重内容的清晰表达和逻辑的连贯性,不像排比句那样强调结构和修辞效果。从语言学角度分析,排比句的各分句具有相同或相似的格式,包括句子长度、标点符号和句法结构等。因此,相比于常规的文本生成任务,如何确保生成的

分句与给定分句结构相同、语义相关、风格一致是排比句生成的重点。常规的文本生成为保证流畅性,通常会忽略对生成文本在句子长度和标点符号上的控制,从而无法满足排比句的基本特征。此外,分句的句法结构相同是排比句的另一特征,因而各分句对应位置词语词性相同和词语之间的依存关系一致等也成为排比句生成的另一难点,这也要求排比句生成模型需考虑词性和依存关系等特征。

本文重点考虑给定分句与待生成分句的格式一致性,从排比句各分句对应位置词语词性对齐和词语之间依存关系相同两个方面着手,提出了一个基于词性对齐与依存关系的中文排比句生成模型(CPG-PosDep)。该模型可以分为两个部分,第一个部分主要对格式信息进行处理,包含词性对齐特征编码模块和词性对齐特征学习模块。词性对齐特征编码模块采用特殊标识的方法设定生成句的格式,分别用分词、句内词序和句间位置三种符号及随机采样策略对排比句进行全局标识并编码;词性对齐特征学习模块在已编码好的词性对齐特征的基础上,利用改进的 Transformer^[5]注意力机制学习排比句词性对齐特征,以保证生成句与给定分句在句子长度、词语数量、词语位置关系和词语词性上的一致性。第二部分是学习给定分句中依存关系特征学习模块,该模块使用 BERT^[6]获取给定分句的语义信息,并对处理好的依存关系输入进行编码,结合注意模块获得依存关系特征表示向量。最终,模型将词性对齐和依存关系这两个特征进行融合处理以完成排比句生成。本文的主要贡献如下:

(1) 收集了 30 452 组排比句,构建排比句生成数据集,解决了这一研究方向语料库缺乏的问题。

(2) 提出了一种学习排比句词性对齐特征的方法,通过设定特殊符号及采用随机采样策略标识排比句和改进 Transformer 的注意力机制,使模型在预测时可以重点考虑词性、词序和词语长度等全局信息。

(3) 实现了本文模型(CPG-PosDep),并通过实验测试了模型在 BLEU 值和 Distinct 值上的性能。实验表明,相比现有的对联或诗歌生成模型,本文模型生成的排比句质量更具优势。

^① 排比句数据集 <https://github.com/puppy210/Parallelisms-Datasets>

1 相关工作

从实现的角度来看,目前关于排比句的研究主要分为两个任务:排比句识别和排比句生成。本节将分别从这两个任务出发介绍相关的研究工作。

1.1 排比句识别研究

早期的排比句识别方法是根据排比句的语法特征,制定了一系列规则对排比句进行识别。例如,熊李艳等人^[7]提出了一种对主题词进行提取,并对标点符号、共现词和工整性进行匹配的中文排比句抽取方法。然而,由于汉语的句法和章法的逻辑结构并不一定完全严格^[8],单纯依赖固定规则而忽视语义信息的方法在自动提取排比句时效果往往不理想。

随着深度学习在自然语言处理领域的发展,通过深度学习提取文本的语意特征对排比句进行自动识别的研究已成为目前主流的研究方法^[9]。在基于深度学习的排比句识别研究中,穆婉清等人^[10]根据排比句结构相似、内容相关的特点,设计了一种融合卷积神经网络和结构相似度计算的排比句识别方法;Dai 等人^[11]提出了一种基于循环神经网络的排比句识别方法,将句子的语义引入模型中,这个过程不需要人为设计特征,避免了因特征选择不当而带来不利影响;朱晓亮等人^[12]通过研究中文排比句的语法特征,设计了一种融合预处理算法,并提出了基于 BERT 架构的排比句自动识别方法;刘广辉^[13]提出基于句子交互的卷积神经网络的排比句自动识别方法,该方法以排比分句间的分词、句法结构和情感极性三个基本特征作为多通道输入,经过多层卷积操作提取局部特征,最终输出是否为排比的概率。综合上述研究,无论是早期的规则制定方法还是现如今基于深度学习的特征提取技术,排比句识别方法主要都从其独特的结构特征出发进行探索和改进。

1.2 排比句生成研究

排比句生成任务是自然语言生成领域的一个重要分支。自 20 世纪 50 年代起,自然语言生成就已经出现在机器翻译领域,此时的研究主要关注构建一种语言到语言之间的映射关系,比如利用规则和字典来进行翻译^[14];在 20 世纪 80 至 90 年代,开始转向统计和机器学习方法,以此来自动生成自然语

言文本,在这个时期,已经能够实现短文本的生成^[15];而 2000 年至今,则逐渐从短文本生成转向复杂的任务和领域,随着深度学习和神经网络技术的发展,自然语言生成的研究开始更加注重语言的进一步理解和语言的创造性^[16]。此外,自然语言生成的研究也逐渐向更多应用领域拓展,比如对话系统、智能问答系统、人机写作等。

目前专门针对排比句生成的研究相对空缺,但排比句生成任务与至今已经相对成熟的诗歌、对联生成任务有着异曲同工之处。排比句的三个分句通常具有相同或相似的结构,而诗歌和对联也各有一定的格式要求。因此,排比句生成任务可以被视为分句数量不定、分句长度不定、分句格律不定的特殊诗歌或对联的生成^[17]。基于此,下面将对诗歌生成以及对联生成任务的相关工作做系统的梳理。

1.2.1 诗歌生成研究

最早的诗歌生成方法是利用预先定义的规则或结构模板来生成。其中,Wu 等人^[18]提出了一种从大规模语料库中提取诗歌的语法规则,并将这些规则应用于诗歌生成的方法。这类方法生成的诗歌语义正确性和丰富程度较差,连贯性、主题和意境的整体性也都难以保证。He 等人^[19]将统计机器翻译(Statistical Machine Translation, SMT)应用于绝句生成。然而这类基于 SMT 的方法主要依据已有数据的统计特征,生成的诗歌可能缺乏新颖性。

为了解决上述问题,研究者们开始引入神经网络模型来学习诗歌的语言模式和结构特征,然后通过生成模型生成新的诗歌。RNNPG 模型^[20]率先将循环神经网络应用于中文古典诗歌生成;为提高生成诗句上下文的连贯性和通顺度,Yi 等人^[21]提出了一种基于工作记忆模型的诗歌生成方法,该方法将关键词和信息有限的历史存储在神经记忆内存中;Tian 等人^[22]提出了一种新的分层框架来生成十四行诗,在解码前对诗歌草图进行规划;Yan 等人^[23]提出了一种基于标签嵌入的图神经网络诗歌主题表征模型,构建了面向主题的中国古代诗歌生成模型 TLPG。

在基于神经网络的诗歌生成方法中,将 Transformer 作为架构的诗歌生成模型取得了更为显著的成果。Zou 等人^[24]提出了基于反向提示的预训练语言模型的诗歌生成方法;Ormazabal 等人^[25]提出了一种无监督的方法,可以根据任意给定的韵律生成诗歌;TPoet^[26]将主题模型集成到基于 Transformer 的自回归文本生成模型中,通过设置分句、

押韵和音调等多个标识符,模型可以显式学习生成诗歌的格式约束条件。上述基于神经网络的诗歌生成方法具有强大的灵活性和表达能力,能够学习复杂的语言模式和结构特征,生成的诗歌通常能够满足格式要求并保持一定的语言风格和创造性。

1.2.2 对联生成研究

通常地,对联生成与诗歌生成都是序列到序列的文本生成任务,但它们的不同之处在于,对联生成要求更为严格的格式和语义规则。对联生成方法主要包括基于语法的方法^[27-29]和遵循序列到序列范式的神经网络方法^[30]。为了提高模型性能,先前的神经网络方法试图对先行词和后续分句之间的字符-字符级别对应进行建模,其中包括注意力机制^[31]和字符嵌入预训练^[32]等方法。Song^[33]提出了一种基于句法信息的对联生成方法 ChiCoupletGen,不仅可以更好地理解对联,而且可以通过更多的句法结构约束来增强对联生成的过程。

受排比句识别、诗歌和对联生成任务研究的启发,本文基于 Transformer 的自回归文本生成框架,通过融合词性对齐特征及给定分句的依存关系特征生成高质量中文排比句。

2 方法

2.1 任务定义

如图 1 所示,中文排比句生成任务指给定长度

为 l 的第一个分句 $\mathbf{x}=(x_1, x_2, \dots, x_l)$,通过模型 M 生成流畅通顺、句式相同、句法结构一致的后续分句 $Y=\{y^1, y^2, \dots, y^n\}$,其中, $y^n=(y_1^n, y_2^n, \dots, y_l^n)$ 。

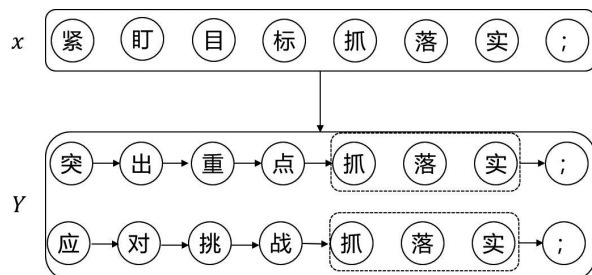


图 1 中文排比句生成示例

2.2 CPG-PosDep 模型

为了实现这项任务,本文设计一种基于词性对齐与依存关系的中文排比句生成模型(CPG-PosDep),通过学习词性对齐特征与给定分句的依存关系生成高质量的排比句。模型的整体框架如图 2 所示,主要包含词性对齐特征编码模块、词性对齐特征学习模块和依存关系特征学习模块。

本文用排比句示例“奋斗有目标;实干有引领;发力有方向。”为实例详细描述我们的模型框架。由于该模型是基于 Transformer 的自回归语言模型,因此在训练期间,输入是给定的排比句句“奋斗有目标”,相应的输出为排比句的后续两个分句,“实干有引领;发力有方向。”

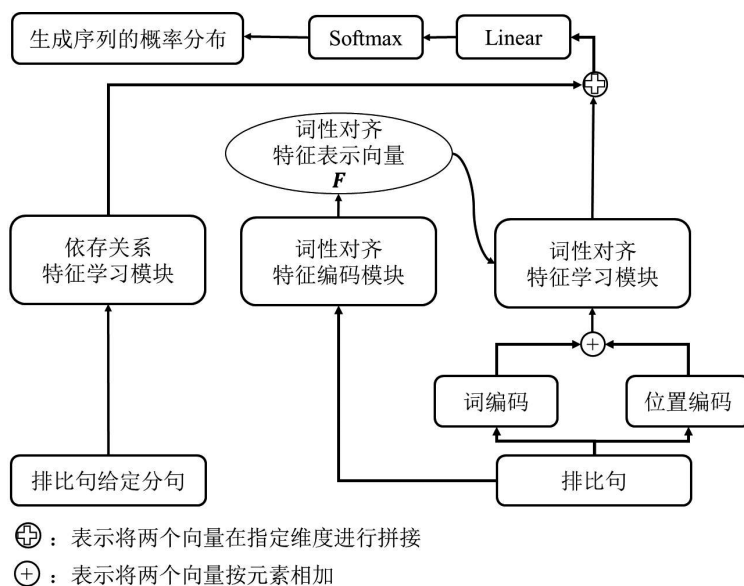


图 2 CPG-PosDep 模型总体框架

2.2.1 词性对齐特征编码模块

本文采用现有的自然语言处理工具 Stanford-CoreNLP^① 对排比句进行分词和词性标注,将训练语料中所有的排比句划分为 31 种词性。如上述实例中的输入分句被分为“奋斗”“有”“目标”三个词以及一个标点符号“;”,对应的词性分别为其他动词“VV”主要动词“VE”其他名词“NN”和标点符号“PU”。第二、三个分句的分词结果与第一个分句一致并对应位置词性对齐,如输入分句中的“奋斗”一词与第二、三个分句中的“实干”“发力”对应,均为两个字的动词。通过实例分析,我们证实排比句具有各分句对应位置词语词性和词语长度相同这一词性对齐特征。

因此,如图 3 所示,为确保排比句中各分句长度相同、对应位置词语词性一致和语义相关,本节设计了词性对齐特征编码模块对排比句的三类信息进行全局表示:分词标识 C 、句内词序标识 P 和句间位置标识 S 。其中,分词标识 C 确保各分句分词一致及对应位置词语长度相同,句内词序标识 P 和句间位置标识 S 确保句子长度相同。其次,基于排比句分句间对应位置的词语不仅词性相同,并可能为排比词(对应位置相同的词语),因此本文对分词标识的设定采用随机采样的策略。最后,通过将三种标识对应编码相加的方式得到词性对齐特征表示向量 F 。

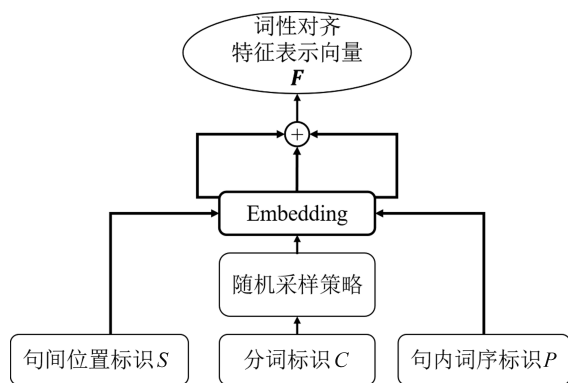


图 3 词性对齐特征编码模块结构

分词标识的设定 在得到排比句的分词结果后,本文使用符号 c_i 表示一个字,用相同的符号下标表示同一个词内的不同字,并且句内采用下标逐渐递增的方式来表示不同词语,如实例给定分句中的首个词“奋斗”二字都用符号 c_1 表示。同时为了将标点与排比句内容进行区分,用 c_0 表示句末的分号。由于排比句具有词性对齐的特点,第二个、第三

个分句的分词结果与给定分句保持一致,对应的符号表示也相同。因此,根据给定分句,即可获得排比句的全局分词标识。此外,使用符号 $\langle /s \rangle$ 来对分句进行分割,使用 $\langle eos \rangle$ 标识句子的结束。这一过程的具体表示方式如式(1)所示。

$$C = \{c_1, c_1, c_2, c_3, c_3, c_0, \langle /s \rangle, \\ c_1, c_1, c_2, c_3, c_3, c_0, \langle /s \rangle, \\ c_1, c_1, c_2, c_3, c_3, c_0, \langle /s \rangle, \langle eos \rangle\} \quad (1)$$

随机采样策略 若直接采用前面得到的分词标识 C 进行训练,只能让模型学习到分句间对应位置词语的字数相同,难以学习到词语相同或词性相同的特征。因此,本节提出采用随机采样策略对分词标识进行改进,增强模型捕捉排比句词性对齐特征的能力。具体地,在给定分句中随机选择一个词语,并将第二、三个分句中与该词位置相对应的词语不用分词符号标识,而是直接使用词语本身。由于这个被选中的词语可能是排比词,也可能是排比句中对应位置词性相同的词语,如式(2)到式(4)所示,从而能让模型充分学习到排比句词性对齐的特征。

$$C = \{\text{奋,斗}, c_1, c_2, c_2, c_0, \langle /s \rangle, \\ \text{实,干}, c_1, c_2, c_2, c_0, \langle /s \rangle, \\ \text{发,力}, c_1, c_2, c_2, c_0, \langle /s \rangle, \langle eos \rangle\} \quad (2)$$

$$C = \{c_1, c_1, \text{有}, c_2, c_2, c_0, \langle /s \rangle, \\ c_1, c_1, \text{有}, c_2, c_2, c_0, \langle /s \rangle, \\ c_1, c_1, \text{有}, c_2, c_2, c_0, \langle /s \rangle, \langle eos \rangle\} \quad (3)$$

$$C = \{c_1, c_1, c_2, \text{目,标}, c_0, \langle /s \rangle, \\ c_1, c_1, c_2, \text{引,领}, c_0, \langle /s \rangle, \\ c_1, c_1, c_2, \text{方,向}, c_0, \langle /s \rangle, \langle eos \rangle\} \quad (4)$$

句内词序标识的设定 本文设置符号 P_i 表示排比句中每个字在分句内的位置。 P_i 采用逆序方式标识不同字,这能够让模型更容易注意到句子何时结尾,以保证生成分句的内容完整性。在排比句各分句间对应位置的字,符号 P 对应的下标也相同,故该标识可使模型捕捉到分句间词序的对应关系。具体表示方式如式(5)所示。

$$P = \{P_5, P_4, P_3, P_2, P_1, P_0, \langle /s \rangle, \\ P_5, P_4, P_3, P_2, P_1, P_0, \langle /s \rangle, \\ P_5, P_4, P_3, P_2, P_1, P_0, \langle /s \rangle, \langle eos \rangle\} \quad (5)$$

句间位置标识的设定 本文使用符号 S_i 表示排比句分句间的位置关系,即同一个分句内的汉字

① <https://stanfordnlp.github.io/CoreNLP/>

使用相同的符号表示。 S_i 采用升序的方式标识不同分句,即用 S_0 表示第一个分句内的每一个字, S_1 表示第二个分句内的每一个字,以此类推。句间位置标识的设置旨在加强模型对不同句子的区分,以学习到不同句子间的交互信息。具体表示方式如式(6)所示。

$$\begin{aligned} S = \{ & S_0, S_0, S_0, S_0, S_0, S_0, </s>, \\ & S_1, S_1, S_1, S_1, S_1, S_1, </s>, \\ & S_2, S_2, S_2, S_2, S_2, S_2, </s>, <eos> \} \end{aligned} \quad (6)$$

对每组排比句获取分词标识 C 、句内词序标识 P 和句间位置标识 S 后,按照式(7),将其对应的编码进行相加,得到词性对齐特征表示向量 F 。

$$F = E_C + E_P + E_S \quad (7)$$

2.2.2 词性对齐特征学习模块

在获取词性对齐特征表示向量 F 后,为了能够让模型进一步学习到排比句的词性对齐特征,本节设计了词性对齐特征学习模块,如图4所示。该模块主要由掩码多头注意力机制模块和全局多头注意力机制模块组成。

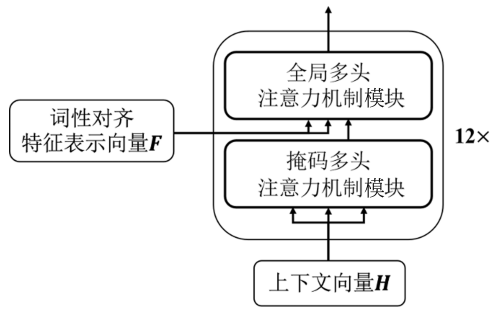


图4 词性对齐特征学习模块结构

模型首先需要将输入的字符、相对位置以及设定的特殊符号等信息依次通过词嵌入层,得到语义空间中相应的词向量,再将所有的词向量按位相加,最终得到当前时刻 t 的历史信息的隐式表示 H_t 。具体地,初始的隐层表示状态如式(8)所示。

$$H_t^0 = E_{w_t} + E_{C_t} + E_{P_t} + E_{S_t} + E_{g_t} \quad (8)$$

其中, H_t^0 表示第0层第 t 个时间步的隐层表示, E_* 表示输入 $*$ 的嵌入向量表示, w_t 、 C_t 、 P_t 、 S_t 分别表示输入、分词、句内词序、句间位置第 t 个位置的字或者标识符。 g_t 则是全局位置表示方法,这是Transformer结构中常用的一种全局位置表示方法。随着解码过程的逐步进行, H 会动态更新,即 H 会随着时间步 t 的变化而变化。

为了重点学习排比句词性对齐特征,我们将词性对齐特征表示向量 F 也作为模型的输入。在对

输入进行处理后,引入掩码多头注意力模块和全局多头注意力模块进行词性对齐特征学习。

掩码多头注意力模块 本文的主要目标是根据给定分句,生成能够与之组成排比句的其他分句。因此,在生成过程中,模型可以根据给定分句和已生成的内容来预测后面的每个字。基于该任务特性,本文引入了掩码多头注意力模块。掩码多头注意力模块是Transformer计算单元中原有的一个模块,具体计算过程如式(9)所示。

$$\begin{aligned} K^0, V^0, Q^0 &= H^0 W^K, H^0 W^V, H^0 W^Q \\ C_t^1 &= \text{LN}(\text{SLF-ATT}(Q_t^0, K_{\leq t}^0, V_{\leq t}^0) + H_t^0) \quad (9) \\ C_t^1 &= \text{LN}(\text{FFN}(C_t^1) + C_t^1) \end{aligned}$$

其中, W^K 、 W^V 和 W^Q 都是可学习的权重矩阵, $\text{LN}(\cdot)$ 、 $\text{SLF-ATT}(\cdot)$ 和 $\text{FFN}(\cdot)$ 分别表示层归一化、自注意力机制以及前馈神经网络模块。通过掩码多头注意力模块可以得到上下文向量 C_t^1 ,即已生成内容与所需预测词之间的语义关联信息。最后,将 C_t^1 作为全局多头注意力机制模块的输入。

全局多头注意力模块 与历史信息 H 不同,词性对齐特征表示向量 F 是固定不变的。训练和测试时都可以通过给定分句对整组排比句进行全局标识,即确定词性对齐特征表示向量 F 。因此,在排比句词性对齐特征学习以及解码过程中无需再更新 F 。在全局多头注意力模块中,将 C_t^1 作为查询向量,融合了分词、句内词序和句间位置信息的词性对齐特征表示向量 F 作为键值向量。通过该注意力块,模型可以学习到完整排比句的词性对齐特征,从而确定当前预测词语的词性、词序和词语长度,具体计算过程如式(10)所示。

$$\begin{aligned} Q^1, K^1, V^1 &= C_t^1 W^Q, F W^K, F W^V \\ H_t^1 &= \text{LN}(\text{GLOBAL-ATT}(Q_t^1, K^1, V^1) + C_t^1) \\ H_t^1 &= \text{LN}(\text{FFN}(H_t^1) + H_t^1) \quad (10) \end{aligned}$$

其中, $\text{GLOBAL-ATT}(\cdot)$ 表示全局多头注意力机制。上述公式完成了词性对齐特征学习模块第一层的计算,该模块经过12层堆叠,最终得到当前时间步 t 时刻包含词性对齐特征的向量 H_t^{12} 。

2.2.3 依存关系特征学习模块

排比句各分句依存关系一致,如图5所示,分句1中的“紧盯”与“抓”并列,“目标”是“紧盯”的直接宾语,“落实”是“抓”的直接宾语,分句2、分句3与分句1具有相同的依存关系。

基于对排比句各分句依存关系相同这一特性的考虑,本文提出了依存关系特征学习模块,该模块主要由BERT、依存关系编码和一个特殊的注意力模

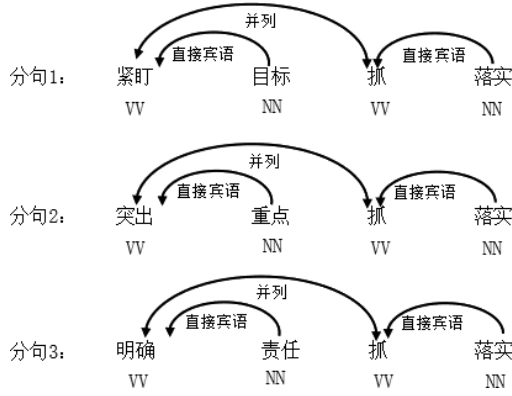


图5 排比句依存关系示例

块三个部分组成,如图6所示。

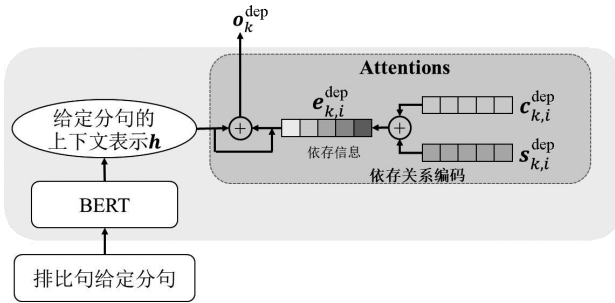


图6 依存关系特征学习模块结构

首先,通过 BERT 对给定分句 x (x 中有 l 个字符,第 k 个字符记为 x_k) 进行编码,得到它的上下文表示向量 h (h_k 是 x_k 的编码输出向量),即给定分句的语义信息,可以简单表示为式(11)。

$$h = \text{BERT}(x) \quad (11)$$

其次,在依存关系编码中,先使用自然语言处理工具 StanfordCoreNLP 获取给定分句的依存关系,再对依存关系进行编码。对于给定分句的每个字符 x_k ,提取一组与其关联的依存关系实例 $\text{DEP}_i = \{(c_{k,i}^{\text{dep}}, s_{k,i}^{\text{dep}}), 1 \leq i \leq u_k\}$,其中 DEP 表示 x_k 的依存关系集合; $c_{k,i}^{\text{dep}}$ 表示与 x_k 有依存关系的词; $s_{k,i}^{\text{dep}}$ 表示 x_k 与 $c_{k,i}^{\text{dep}}$ 的依存关系类型; u_k 表示依存关系个数; i 表示在集合 DEP 中的第 i 个元素。在得到 DEP 之后,对于每组实例 $(c_{k,i}^{\text{dep}}, s_{k,i}^{\text{dep}})$,再将 $c_{k,i}^{\text{dep}}$ 和 $s_{k,i}^{\text{dep}}$ 分别映射到嵌入层,获取嵌入向量 $c_{k,i}^{\text{dep}}$ 和 $s_{k,i}^{\text{dep}}$,并通过相加的方式得到依存关系的编码表示 $e_{k,i}^{\text{dep}}$,如式(12)所示。

$$e_{k,i}^{\text{dep}} = c_{k,i}^{\text{dep}} + s_{k,i}^{\text{dep}} \quad (12)$$

然后,通过注意模块计算得到依存关系对所在句的贡献权重矩阵,即分配给每个实例的注意力 $a_{k,i}^{\text{dep}}$,如式(13)所示。

$$a_{k,i}^{\text{dep}} = \frac{\exp(h_k \cdot e_{k,i}^{\text{dep}})}{\sum_{i=1}^{u_k} \exp(h_k \cdot e_{k,i}^{\text{dep}})} \quad (13)$$

最后,将 $a_{k,i}^{\text{dep}}$ 与 $e_{k,i}^{\text{dep}}$ 相乘,计算不同依存关系实例的加权和,并将 h_k 加到结果向量上,得到包含语义信息和依存关系信息的输出向量 o_k^{dep} ,如式(14)所示。

$$o_k^{\text{dep}} = h_k + \sum_{i=1}^{u_k} a_{k,i}^{\text{dep}} \cdot e_{k,i}^{\text{dep}} \quad (14)$$

为了在训练过程中减少模型在生成文本时的概率分布与真实文本之间的概率分布差异,本文使用负对数似然作为损失函数,如式(15)所示。

$$L^{\text{nl}}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^N \log P(\hat{y}_i | \hat{y}_{<i}) \quad (15)$$

其中, N 是文本序列的长度, $P(\hat{y}_i | \hat{y}_{<i})$ 是模型在给定前文 $\hat{y}_{<i}$ 的条件下,预测字符 $\hat{y}_i$ 的概率。通过最小化负对数似然函数,优化模型参数,使模型的生成结果更加接近真实数据分布。

3 实验

3.1 数据集

3.1.1 数据集构建

本文主要通过从文学作品中抽取和网站采集的方式进行排比句数据的收集。考虑到句子长度对模型性能的影响,本文去除了长度超过 60 个字(每个分句长度超过 20 个字)的排比句。

在构建数据集之前,本文利用自然语言处理工具 StanfordCoreNLP 对数据进行分词处理和依存关系分析,得到所有排比句的分词结果以及每个分句的依存关系。然后,对部分分词结果明显不一致的数据进行人工修正,防止此类不符合排比句特性的数据对模型生成效果造成不良影响,具体修正情况如表 1 所示。表 1 中左边存在分句对应位置分词结果不是对其进行修正时重新替换的满足分词一致要求的字符,右边加粗的则是进行修正时重新替换的满足分词一致要求的字符。如表 1 中第三组排比句修正示例,左边是原始分词结果,可以看到第一个分句中“为官做官”与后两个分句对应位置的分词不一致,我们将其替换成“当官任职”,进行修正后,排比句的三个分句的分词结果将保持一致,满足排比句的结构要求。

在对数据按上述过程进行清洗后,最终共得到 30 452 组排比句。以上数据集(Parallelisms Datasets)

的构建为后续排比句生成模型的训练与测试奠定了 基础。

表 1 分词结果不一致的排比句修正示例

原数据及其分词结果	修正结果
勤俭节约 是 干事创业 的 鲜活 教科书; 勤俭节约 是 拒腐防变 的 无形 安全门; 勤俭节约 是 砥砺人生 的 最好 营养剂;	勤俭节约 是 干事创业 的 鲜活 教科书; 勤俭节约 是 拒腐防变 的 无形 安全门; 勤俭节约 是 砥砺人生 的 安全 营养剂;
系统 推出 一 批 有 影响力 的 研究成果; 编辑 整理 一 批 权威性 强的 文献典籍; 策划 打造 一 批 十分 高能 的 平台载体;	系统 推出 一批 有 影响力 的 研究成果; 编辑 整理 一批 有 权威性 的 文献典籍; 策划 打造 一批 有 高能性 的 平台载体;
为官 做官 , 要 撸起 袖子 、 落好 棋子; 干事创业 , 要 甩开 膀子 、 干出 样子; 为民服务 , 要 俯下 身子 、 做枚 石子;	当官任职 , 要 撸起 袖子 、 落好 棋子; 干事创业 , 要 甩开 膀子 、 干出 样子; 为民服务 , 要 俯下 身子 、 做枚 石子;

3.1.2 数据集划分

本文的实验均使用上一节中构建的排比句数据集 Parallelisms Datasets。该数据集包含 30 452 组排比句。我们按照 8 : 1 : 1 的比例划分数据集,根据排比句类别和长度的分布,两次从数据集随机选择 3 045 组分别用作测试集和验证集,剩余的排比句用作训练集。具体划分情况如表 2 所示。

表 2 数据集划分情况

数据集	训练集	测试集	验证集
排比句	24 362	3 045	3 045

3.2 参数设置

本文将模型迭代次数设为 30 次,词性对齐特征学习模块中的掩码多头注意力块和全局多头注意力块中均采用 12 个注意力头,隐藏层大小为 768。在依存关系特征学习模块中,使用的 BERT 为 BERT-Base-Chinese^①,该 BERT 模型的隐藏层大小设置为 768,Dropout 设置为 0.1,注意力头的数量为 12。本文采用了 Adam 优化器对损失函数进行优化,结合学习率衰减策略和 10 000 个预热步骤来进行训练。

3.3 评价指标

为保证模型评价的客观性和公平性,本文结合 BLEU 值^[34]和 Distinct 值^[35]来对模型进行评估。BLEU 指标是用于评估机器翻译或其他自然语言生成任务中生成文本质量的标准指标之一。它通过衡量机器生成的文本与参考文本之间的字符重叠度来评估生成文本的准确性。BLEU 值越高表明生成的排比句与原排比句越接近,因此也有更高的可能性以取得较好的相关性和连贯性,即模型的性能越好。Distinct 指标通过计算生成文本中不重复的

N-gram 数量与总 N-gram 数量的比例,来度量生成文本的多样性。Macro-Distinct (MA-D)和 Micro-Distinct(MI-D)是 Distinct 的扩展,用于在不同层面上分析多样性,前者关注单个文本的不重复项比例,而后者则考虑所有生成文本的总和。Distinct 值越高,表明生成的排比句越多多样化,内容更丰富。

3.4 实验结果和分析

3.4.1 对比实验

目前,关于排比句生成的研究工作相对匮乏。为了更好地验证模型的有效性,本文选择了与排比句生成任务相关度较高,并修改后适合用于该任务的模型进行对比实验。

SongNet^[36] 一个基于预训练语言模型的固定格式控制下的文本生成模型。SongNet 在格式、押韵和句子完整性三个方面量身定制了一套符号集,并改进了注意力机制,使模型能够捕获有关全局格式的信息,最终生成严格符合格式和韵律的古诗词。

MixPoet^[37] 一个基于半监督变分自编码器的诗歌生成模型。MixPoet 主要通过后验网络和先验网络进行训练,以得到隐变量。后验网络中得到的隐变量被视为真值,目的是使先验网络中得到的隐变量尽可能接近真值,从而得到生成诗歌的首句编码。然后,通过这个首句编码生成诗歌的后三句。本文对该模型进行了修改,使其能够根据输入的排比句句生成后续分句。

对比实验结果如表 3 所示,从中可以看出,在测试集上,CPG-PosDep 模型生成的文本相比两个对比模型生成的文本的 BLEU 值更高,这表明本文模型生成的结果与人类写出的排比句更相似。一方

① <https://huggingface.co/google-bert/bert-base-chinese>

面,由于 SongNet 模型和 MixPoet 模型在生成文本时并未考虑排比句的结构特征,而本文模型从词性对齐和依存关系两方面控制生成文本的结构,这保证了生成句与参考句结构的一致性,使得模型在 BLEU 值的指标上得到了较大的提升。另一方面,由于本文的模型使用了基于预训练的语言模型,具有海量数据的语言知识,比 MixPoet 模型具有更好的文本相关性和连贯性。

在文本多样性方面,CPG-PosDep 模型比 SongNet 模型生成的文本多样性更好,MixPoet 模型也仅在 MI-D-2 值上略高于我们的模型,这表明本文模型生成的句子重复内容相对较少。总之,单纯地使用 SongNet 和 MixPoet 效果并不理想。这说明在排比句生成中充分考虑排比句特有的格式特性、设计相应的特征学习模块等获得了预期的效果,而这也是我们研究的出发点所在。

表 3 对比实验结果

Model	BLEU $\uparrow$				Distinct $\uparrow$			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	MA-D-1	MI-D-1	MA-D-2	MI-D-2
SongNet	30.20	25.18	22.47	20.21	71.23	6.56	80.37	40.35
MixPoet	44.72	31.23	23.95	19.42	62.23	6.20	84.97	55.15
CPG-PosDep	51.74	41.66	35.93	32.06	75.45	7.23	89.45	50.05

3.4.2 消融实验

为进一步证明融合信息对模型性能提升的有效性,本节分别对融合分词、句内词序和句间位置信息的词性对齐特征表示向量 F 中的每个组成部分 (C 、 P 、 S)以及依存关系特征学习模块进行消融实验,结果如表 4 所示。在实验中,我们逐一移除 CPG-PosDep 模型中词性对齐特征表示向量 F 的不

同组成部分。结果显示,移除任何一个组成部分后,模型的 BLEU 值均出现了显著下降。这一现象表明,词性对齐特征向量 F 中的每个部分在模型的文本生成过程中都发挥了不可或缺的作用。同时,依存关系特征学习模块的去除同样会使 BLEU 值有所下降,进一步证实了依存关系特征的融入对提升模型性能的重要性。

表 4 去除 F 各组成部分和依存关系特征学习部分的消融实验结果

Model	BLEU $\uparrow$				Distinct $\uparrow$			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	MA-D-1	MI-D-1	MA-D-2	MI-D-2
CPG-PosDep(- C)	40.91	28.14	21.66	16.77	73.73	6.98	87.90	46.81
CPG-PosDep(- P)	38.88	25.09	17.76	13.19	75.92	6.84	90.85	49.38
CPG-PosDep(- S)	38.71	24.84	17.38	12.63	76.23	6.81	89.97	47.26
CPG-PosDep (-依存关系)	48.36	37.53	33.58	30.69	75.32	7.18	89.60	51.58
CPG-PosDep	51.74	41.66	35.93	32.06	75.45	7.23	89.45	50.05

具体地,从 BLEU 指标的变化程度可以发现,去除词性对齐特征表示向量 F 中的任意组成部分 (C 、 P 、 S)对 BLEU 值的影响明显大于去除依存关系特征学习模块的影响。这种现象表明,词性对齐特征在文本生成任务中具有更直接和更强的贡献。这是因为词性对齐特征直接影响了生成文本的结构一致性和连贯性,使得生成文本能够更精准地模仿给定分句的句法结构。词性对齐特征向量 F 的各个部分在提升模型性能中的重要性,可能源于它们在不同层面对文本生成的约束作用。例如,分词

标识(C)确保了词汇的准确使用,句内词序标识(P)维护了句子内部的合理结构,而句间位置标识(S)则保证了整体文本的连贯性。这三者共同作用,形成了一种多层次的结构化约束,使得生成的文本在多个维度上与给定分句保持高度一致。因此,去除 F 的任何一部分都会导致 BLEU 值显著下降。相较之下,依存关系特征虽然也对文本生成的语法正确性和语义连贯性有重要影响,但其作用更多地可能表现在更高层次的语法结构上。因此,去除依存关系特征学习模块虽然会导致性能

下降,但其影响相对较小。消融实验结果分析进一步强调了在排比句生成中,词性对齐特征表示向量 $\mathbf{F}$ 的不可替代性,以及它在提高模型输出质量方面的关键作用。

从以上结果可以得出结论:引入的词性对齐特征表示向量 $\mathbf{F}$ (由 C 、 P 和 S 组成)与依存关系对模

型性能确实起到了至关重要的作用,每个部分在提升模型性能方面都发挥了独特的作用,而它们的组合则能够实现最佳性能。

3.4.3 实例分析

当给定一个分句时,模型需要根据给定分句生成另外两个分句,实验结果如表 5 所示。

表 5 不同模型生成的文本示例

	示例 1	示例 2
给定分句	敢想他人未曾想;	加强思想淬炼,打造思想上的金钟罩;
目标句	敢谋他人未曾谋; 敢干他人未曾干;	加强政治历练,练就政治上的明白人; 加强实践锻炼,成为作风上的实干派;
MixPoet	敢于未人显新走; 敢于未人越走走;	坚炼实治修炼,增造厉律上硬的能本; 坚炼实律修炼,增强纪律立硬的高本;
SongNet	有欲甚则邪心胜; 在严字上下功夫;	坚持一稳皆稳、一失不前的意识不变; 一个个对标一、奋发有为的创新激情;
CPG-PosDep	敢做他人未曾做; 敢闯他人未曾闯;	坚实组织基础,画好群众上的同心圆; 持续凝聚共进,筑牢党性上的防火墙;

从所给示例可以看出,模型 MixPoet 的生成句与给定分句具有部分相同的排比词,但分词结果不一致,也不能与给定分句构成排比关系;SongNet 生成的文本在格式上与给定句相差较大,更加无法与之构成排比句;而本文模型的生成句较为通顺,与给定分句之间存在相同的排比词“敢……他人未曾……”和“……,……的……”,分句间相应位置上的词性大致相同,依存关系也大体一致,能够组成完整的排比句。

总的来说,从所给示例中可以看出,本文模型取得了目前最好的结果。虽然 SongNet 和 MixPoet 生成的分句可以做到与给定分句的长度一致,但是它们各自具有明显的缺陷。SongNet 生成的句子较为通顺,却无法学习到排比句的特性,如各分句中具有排比词、对应相同位置词语的词性相同、依存关系一致等。MixPoet 只有在偶然情况下才可以学习到排比句的部分特性,并不能保证每个对应位置词语的词性都相同,生成句的流畅性也不好。而本文模型之所以能够取得最佳生成结果,是因为同时考虑了排比句的分词、词序、词语长度和依存关系等信息,让模型能够学习到排比句的特性,从而生成与给定分句构成排比关系的其他分句。

4 总结与展望

排比句讲究结构对齐、逻辑连贯、内容关联,使

得计算机自动生成排比句的难度较大。目前文本生成领域专门针对排比句生成的研究相对空缺。本文针对排比句生成任务,收集并构建了相关数据集 Parallelisms Datasets,结合排比句具有的语言学特征,提出了基于词性对齐与依存关系的中文排比句生成模型 CPG-PosDep。通过考虑排比句的各分句对应位置词语词性对齐和词语之间依存关系一致的特征,首先引入词性对齐特征编码和词性对齐特征学习模块,使模型在预测时可以考虑词性、词序和词语长度等全局信息。最后提出依存关系特征学习模块对给定分句的依存关系进行编码,并为预测提供相应指导。实验结果表明,本文提出的方法可以生成与给定分句词性对齐且依存关系相同的后续分句。

本文主要研究的是如何根据给定分句来生成能与之构成排比关系的后续分句,而在自动写作中人们常常希望能够通过较少的输入内容来生成符合要求的排比句。因此,研究如何仅给定关键词或主题来生成与之匹配的排比句,或者仅给定排比词就能生成通顺的排比句是将来工作的重点。今后的研究中,我们将从以下几个方面考虑:探究如何更好地学习排比句关键词与各分句间关系的表征,进而更好地理解关键词与各分句的关系,以及探究如何学习每个分句与排比词之间的关系,理解排比词与其所在分句的内在关联。

参考文献

- [1] 谭学纯, 濮侃, 沈孟璿. 汉语修辞格大辞典[M]. 上海: 上海辞书出版社, 2010.
- [2] 尚观胜. 用多彩修辞打开孩子奇妙的想象[J]. 课外语文, 2023, (10): 67-69.
- [3] 矣晓沅. 具有文学表现力的中文古典诗歌自动写作方法研究[D]. 北京: 清华大学博士学位论文, 2021.
- [4] 朱媛媛. 基于循环神经网络的对联生成模型研究[D]. 吉林: 吉林大学硕士学位论文, 2018.
- [5] 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.
- [6] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 4171-4186.
- [7] 熊李艳, 林晓乔, 钟茂生. 面向自动写作的中文排比句抽取方法[J]. 计算机应用研究, 2018, 35(6): 1751-1755.
- [8] 张伯江, 郭光. 消极修辞的灵活度[J]. 当代修辞学, 2019, (03): 1-10.
- [9] 谯宇同. 基于 BERT 和依存句法分析的小学作文优秀例句推荐[D]. 武汉: 华中师范大学硕士学位论文, 2021.
- [10] 穆婉青, 廖健, 王素格. 融合 CNN 和结构相似度计算的排比句识别及应用[J]. 中文信息学报, 2018, 32(02): 139-146.
- [11] DAI Y G, SONG W, LIU X J, et al. Recognition of parallelism sentence based on recurrent neural network[C]//Proceedings of the IEEE 9th International Conference on Software Engineering and Service Science, 2018: 148-151.
- [12] 朱晓亮, 谯宇同. 基于 BERT 模型的排比句自动识别方法[J]. 计算机应用与软件, 2021, 38(07): 153-158.
- [13] 刘广辉. 基于修辞句识别的中学语文作文评价系统的设计与实现[D]. 辽宁: 中国科学院大学(中国科学院沈阳计算技术研究所)硕士学位论文, 2022.
- [14] 高明虎, 于志强. 神经机器翻译综述[J]. 云南民族大学学报(自然科学版), 2019, 28(01): 72-76.
- [15] 李雪晴, 王石, 王朱君, 等. 自然语言生成综述[J]. 计算机应用, 2021, 41(05): 1227-1235.
- [16] 梁明轩, 王石, 朱俊武, 等. 知识增强的自然语言生成研究综述[J]. 计算机科学, 2023, 50(S1): 11-18.
- [17] Yan R, Li C T, Hu X, et al. Chinese couplet generation with neural network structures [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 2347-2357.
- [18] Wu X, Tosa N, Nakatsu R. New hitch haiku: An interactive renku poem composition supporting to-ol applied for sightseeing navigation system[C]//Proceedings of the International Conference on Entertainment Computing, 2009: 191-196.
- [19] He J, Zhou M, Jiang L. Generating Chinese classical poems with statistical machine translation models[C]//Proceedings of the 26th AAAI Conference on Artificial Intelligence, 2012: 1650-1656.
- [20] Zhang X X, Lapata M. Chinese poetry generation with recurrent neural networks[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014: 670-680.
- [21] Yi X, Sun M, Li R, et al. Chinese poetry generation with a working memory model[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018: 4553-4559.
- [22] Tian Y, Peng N. Zero-shot sonnet generation with discourse-level planning and aesthetics features[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, 2022: 3587-3597.
- [23] Yan Y, Wen D, Yang L, et al. Poetry generation combining poetry theme labels representations[C]//Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing, 2023: 1246-1255.
- [24] Zou X, Yin D, Zhong Q, et al. Controllable generation from pretrained language models via inverse prompting[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021.
- [25] Ormazabal A, Artetxe M, Agirrezabal M, et al. PoeLM: A meter-and rhyme-controllable language model for unsupervised poetry generation[C]//Proceedings of the Association for Computational Linguistics, 2022: 3655-3670.
- [26] Yang L, Shen Z, Zhou F, et al. TPoet: Topic-enhanced Chinese poetry generation[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2023, 22(6): 1-15.
- [27] GREENE E, BODRUMLU T, KNIGHT K. Automatic analysis of rhythmic poetry with applications to generation and translation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2010: 524-533.
- [28] JIANG L, ZHOU M. Generating Chinese couplets using a statistical MT approach[C]//Proceedings of the 22nd International Conference on Computational Linguistics, 2008: 377-384.
- [29] 张开旭, 孙茂松. 统计与规则相结合的古文对联应对

- 模型[J].中文信息学报,2009,23(01): 100-105.
- [30] GUO Z, ZHANG Y, LU W. Attention guided graph convolutional networks for relation extraction [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 241-251.
- [31] ZHANG J, ZHANG Z, ZHANG S, et al. VV-Couplet: An open source Chinese couplet generation system[C]//Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2018: 1756-1760.
- [32] GAO R, ZHU Y, LI M, et al. Encoder-Decoder couplet generation model based on ‘Trapezoidal Context’ character vector[J].The Computer Journal, 2020, 64(3): 286-295.
- [33] SONG Y. Chinese couplet generation with syntactic information[C]//Proceedings of the 29th International Conference on Computational Linguistics, 2022: 6436-6446.
- [34] 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, 2002: 311-318.
- [35] LI J, GALLEY M, BROCKETT C, et al. A diversity-promoting objective function for neural conversation models [C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 110-119.
- [36] LI P, ZHANG H, LIU X, et al. Rigid formats controlled text generation[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 742-751.
- [37] YI X, LI R, YANG C, et al. Mixpoet: Diverse poetry generation via learning controllable mixed latent space[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(05): 9450-9457.



钟茂生(1974—),博士,教授,主要研究领域为自然语言处理、智能教育与软件、机器学习与数据挖掘等。

E-mail: zhongmaosheng@sina.com



吴如萍(1998—),通信作者,硕士研究生,主要研究领域为自然语言处理、智能教育与软件。

E-mail: 1455791641@qq.com



刘蕾(2000—),硕士研究生,主要研究领域为自然语言处理、考试软件与系统。

E-mail: 1419816401@qq.com