

# 基于外部知识筛选的主题文本生成技术研究

王 沛, 杨 频, 程芃森, 代金鞘, 贾 鹏

(四川大学网络空间安全学院, 成都 610065)

**摘要:** 在自然语言生成任务中, 主题文本生成是一项富有挑战性的工作, 其主要难点在于: 源信息量远小于目标生成的信息量. 为了解决该问题, 本文提出一个基于外部知识筛选的主题文本生成模型 Trans-K, 通过引入与主题词相关的外部知识来丰富源信息, 进而提高生成文本的质量. 本文为了解决引入外部知识的“一词多义”问题, 提出一种基于线性变换的主题向量计算方法, 用于筛选和主题词语义一致的外部知识; 提出一种基于注意力机制的外部权重计算方法, 为每个外部词设定一个主题权重, 使其更贴合文本语义; 为了解决主题词(含候选词)在生成文本中反复出现的问题, 提出一种基于多头注意力机制的内部权重计算方法. 在 EASSY 数据集上的实验表明, 与基线相比, Trans-K 生成文本质量的各项指标更优. 此外, 人类评估表明, 该模型可生成与主题更相关、语言更连贯、且符合语义逻辑的文本.

**关键词:** 自然语言生成; 主题文本生成; Transformer; HowNet; 知识增强

**中图分类号:** TP391.1 **文献标志码:** A **DOI:** 10.19907/j.0490-6756.2024.012003

## Research on topic text generation technology based on external knowledge filtering

WANG Pei, YANG Pin, CHENG Peng-Sen, DAI Jin-Qiao, JIA Peng

(School of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China)

**Abstract:** In the natural language generation task, topic text generation is a challenging task, the main difficulty is that the amount of source information is much smaller than the amount of information generated by the target. To solve this problem, this paper proposes a topic text generation model called Trans-K based on external knowledge filtering, which enriches the source information by introducing external knowledge related to topic words, thereby improving the quality of the generated text. In this paper, in order to solve the "polysemy" problem of introducing external knowledge, a topic vector calculation method based on linear transformation is proposed to filter external knowledge consistent with the semantics of the topic words. An external weight calculation method based on attention mechanism is proposed, which sets a topic weight for each external word to make it more suitable for text semantics. In order to solve the problem that topic words including candidate words, appear repeatedly in the generated text, an internal weight calculation method based on the multi-head attention mechanism is proposed. Experiments on the EASSY dataset show that Trans-K is superior to various indicators of the quality of generated text compared to the baseline. In addition, human evaluations show that the model can generate more topic-relevant, linguistically coherent, and semantically logical's text.

收稿日期: 2022-09-20

基金项目: 四川省科技厅重点研发项目(2021YFG0156)

作者简介: 王沛(1998-), 女, 重庆人, 硕士研究生, 研究方向为信息内容安全. E-mail: 1335316220@qq.com

通讯作者: 杨频. E-mail: yangpin@scu.edu.cn

**Keywords:** Natural language generation; Topic text generation; Transformer; HowNet; Knowledge enhancement

## 1 引言

随着信息时代发展,人工智能技术对人类生活产生了巨大影响.基于人工智能的自然语言生成(Natural Language Generation, NLG)技术被广泛应用于新闻写作、自动对话等领域,现已成为人工智能的研究热点之一<sup>[1]</sup>.主题文本生成(Topic-to-Essay Generation, TEG)是近几年提出的一个新方向,它是指以一组用户指定的主题词集合作为输入,输出一段与主题相关、连贯的文本<sup>[2]</sup>.此技术可快速将舆论话题中的几个热点关键词扩展成一段文本,也能为其他文本生成任务的个性化、多样化发展提供更多可能性.

然而,相较于对话系统、机器翻译、摘要生成等文本生成任务,主题文本生成更具挑战性.一个主要原因是:TEG 任务的源信息过少,而目标生成的信息量较多.如果输入与输出间存在着巨大的语义信息差,则容易造成生成的句子枯燥且质量低下<sup>[2]</sup>.从图 1 不难看出,机器翻译输入和输出的信息量大致相同,文本摘要的输入明显大于输出,而 TEG 的输入则远小于目标生成的信息量.

为了解决源信息匮乏的问题,研究者们提出引入外部知识来丰富源信息,但现有的研究方法仍存在以下问题:(1) 外部词的“一词多义”问题:对于同一个主题词,从知识图谱中检索出的外部词可能具有多个不同的含义.如果将这些有歧义的外部知识直接融入解码器,可能会引入与主题无关的信息.虽然已有研究<sup>[3]</sup>提出将平均主题词向量和不同的外部词计算相似度,以减少歧义性,但平均值法是模糊和非排他性的<sup>[2]</sup>.(2) 生成文本的主题表达不全面:在生成文本的过程中,可能会出现某些词反复出现,而另一些词没有出现的情况.虽然现有主题覆盖机制<sup>[2]</sup>可加强对未表达话题词的关注度,但它多用于循环神经网络(Recurrent Neural Network, RNN)<sup>[4]</sup>,且一般只作用于主题词,未对引入的外部词进行处理.

为了解决上述问题,本文提出了一个融合外部知识的主题文本生成模型 Trans-K,该模型基于 Transformer<sup>[5]</sup>框架,通过自注意力机制捕获文本中的长依赖关系;同时取消了编码器中的位置编码,以解决 TEG 任务的时序问题.本文的贡献点

如下:(1) 提出了一种基于线性变换的主题向量计算方法,用于筛选和主题词语义一致的外部词,解决了外部词的“一词多义”问题;(2) 提出了一种基于注意力机制的外部权重计算方法,为外部词设定主题权重,使引入的外部知识更加贴合文本语义,避免引入不必要的噪声;(3) 提出了一种基于多头注意力机制的内部权重计算方法,在解码器中更新多头自注意力分数,使模型更多地关注未表达的主题词,提高生成文本的主题相关性;(4) 在公用数据集 EASSY 上的自动和人工评估结果均表明,本文提出的模型优于先进的基线模型.

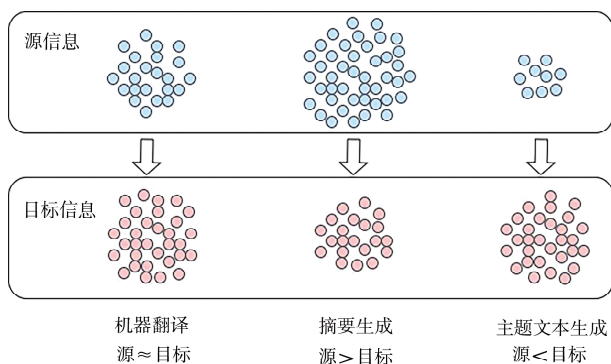


图 1 三种文本生成任务的信息流动图

Fig. 1 Information flow diagram of three text generation tasks

## 2 相关工作

2018 年, Feng 等<sup>[2]</sup>首次提出了 TEG 任务.论文设计了 TAV、TAT 和 MTA 三个模型,逐步引入平均主题词向量、注意力机制和主题覆盖机制.其中 MTA 模型成为该领域的经典基准模型之一.2019 年, Yang 等<sup>[6]</sup>提出通过动态记忆机制将来自外部知识库的常识整合到生成器中,从而提高生成文章的新颖性和多样性,并采用基于多标签鉴别器的对抗性训练来进一步提高主题一致性.2021 年, Luo 等<sup>[3]</sup>提出了一个基于义原的主题文本生成模型,使用一种非当前主题词的相似性度量法来过滤义原信息;2021 年, Liu 等<sup>[7]</sup>提出了一种具有综合知识增强功能的主题文本生成模型,通过教师-学生的 CVAE 网络实现内部知识增强,同时通过一个主题知识图编码器实现外部知识增强.2021 年, Pascual 等<sup>[8]</sup>提出了一种即插即用解码方法

K2T. 将词汇表中的概率分布向与目标约束词语义相似的单词添加一个偏移. 虽然该方法较直观, 但可能使词汇的概率分布变得混乱.

然而, 上述方法都基于 RNN 实现, 该网络结构在 TEG 任务中存在着一些不足. 一方面, 由于 TEG 任务的源信息过少, 随着目标文本长度的增加, RNN 很难捕获文本中的长依赖关系<sup>[9]</sup>; 另一方面, RNN 适合挖掘时序数据的语义<sup>[10]</sup>, 但主题词集合不是严格的时序数据. 采用时序模型来获取输入词语的上下文向量, 有些不符合常理. 2022 年, He 等<sup>[11]</sup>提出了一个基于 Transformer 的分层主题文本生成模型, 可在一定程度上缓解 RNN 的长依赖问题, 但该方法仍然没有很好地改善源信息匮乏的问题.

与主题文本生成任务类似的一个任务是中国诗歌的生成. 2014 年, Zhang 等<sup>[12]</sup>首次将 RNN 引入到中文诗歌生成任务中. 2016 年, Wang 等<sup>[13]</sup>提出了一种两段式的生成方法, 先生成几个子主题, 再利用 RNN 逐次生成对应行的诗句. 2019 年, Liu 等<sup>[14]</sup>提出了一种修辞控制编码器, 用于现代汉语诗歌生成任务. 2021 年, Shao 等<sup>[15]</sup>按照风格、情感、格式和主要关键字, 对数十万首诗歌进行分类, 并通过掩码自注意力机制关联多个标签, 进而生成风格、情绪可控的诗歌.

然而, 中文诗歌具有明显的结构化规则和平仄规律, TEG 任务的生成目标却是非结构化的长文本. 直接沿用诗歌生成任务的研究思路, 往往会导致主题漂移问题, 给 TEG 任务带来挑战.

### 3 模型描述

主题文本生成任务可表示为: 给定一个由  $k$  个主题词组成的集合  $T = \{t_1, t_2, \dots, t_k\}$ , 生成一个包含  $T$  中主题的句子  $Y = \{y_1, y_2, \dots, y_n\}$ , 且  $Y$  内的词语数  $n$  远远大于主题词数量  $k$ . 模型的训练目标是从主题-文本对的数据集中, 获得具有  $\theta$  参数的最优模型 $\hat{\theta}$ , 使条件概率最大化:

$$\hat{\theta} = \arg \max_{\theta} P_{\theta}(Y|T) \tag{1}$$

本文提出了一个融合外部知识的主题文本生成模型 Trans-K, 图 2 给出了模型的整体结构图. 首先, 基于 Transformer 构建编码器和解码器, 特别的是, 在编码器中取消了位置编码, 因为输入的主题词集合不具备严格的时序性; 其次, 基于线性变换的主题向量计算方法, 筛选出和主题词语义一致的外部词; 再次, 为筛选后的外部词设置外部权重, 减少不必要的噪声; 最后, 根据外部知识和编码器隐藏层状态, 更新解码器的内部权重, 并将其应用到多头自注意力机制中, 使模型更多地考虑未表达的词语.

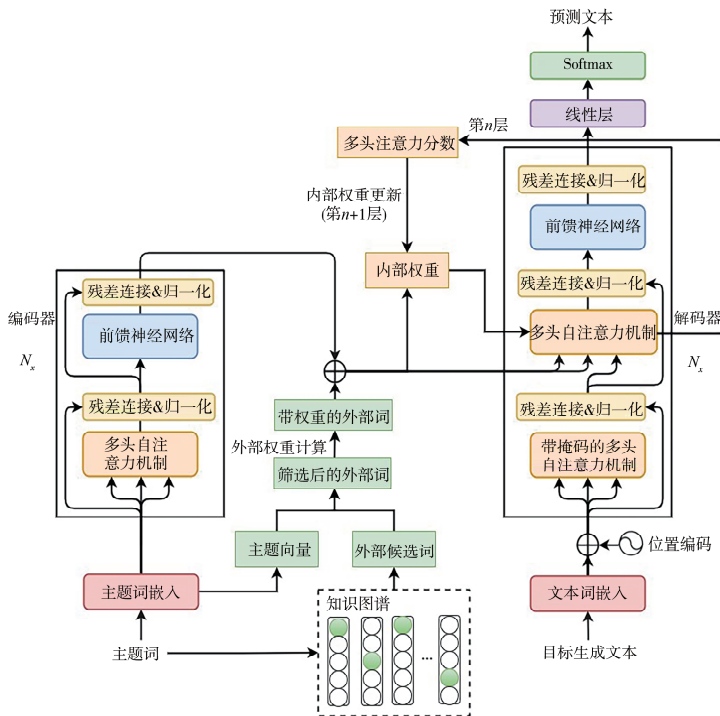


图 2 Trans-K 模型整体结构图

Fig. 2 Overall structure of Trans-K model

### 3.1 基于线性变换的主题向量计算方法

将主题词集合  $T = \{t_1, t_2, \dots, t_k\}$  中的某一主题词  $t_i$  作为索引, 在外部知识图谱中进行检索, 可得到  $L$  个候选词. 检索结果表示为

$$\text{candidate}_i = \{t_i: [c_{i1}, c_{i2}, \dots, c_{iL}]\} \quad (2)$$

然而, 中文里普遍存在着“一词多义”的现象, 例如: “苹果”一词可检索出“水果”和“电脑品牌”两种完全不同的扩展含义. 因此, 本文将当前主题词集合的主题向量和各候选词进行相似度计算, 以避免引入与主题无关的信息.

文献[3]提出将所有非当前主题词的平均词向量作为主题向量, 但 this 方法是模糊和非排他性的<sup>[2]</sup>. 如: 主题词组  $A: [0.1, 0.3, 0.5]$  和  $B: [-0.1, 0.8, 0.2]$  都可得到平均词向量  $[0.3]$ . 为了解决这个问题, 本文提出一种基于线性变换的主题向量计算方法.

(1) 将一组主题词按照字典序号从小到大的方式进行排序, 排序后的主题词组表示为  $T_{\text{sort}}$ . 这是为了避免在后续操作中, 出现同样的词语因排列顺序不同而拼接结果不同的情况;

(2) 将所有排序后的主题词依次拼接起来, 得到一个新的主题向量  $T_{\text{merge}}$ . 由于关键词在生成文本中的实际顺序与输入顺序无关, 因而采用固定方式拼接不会对结果造成影响.

$$T_{\text{merge}} = \text{concatenate}(T_{\text{sort}}) \quad (3)$$

其中, 函数 concatenate 表示将输入的有序向量集合进行行拼接. 假设每个主题词的向量维度为  $x$ , 则  $T_{\text{merge}}$  的维度为  $k \times x$ .

(3) 利用线性变换, 对  $T_{\text{merge}}$  进行降维, 使其与各候选词  $c_{il}$  的维度保持一致, 降低后续计算的复杂度. 压缩维度的公式如下.

$$T_{\text{linear}} = A \cdot T_{\text{merge}} + b \quad (4)$$

其中,  $T_{\text{linear}}$  表示当前主题词组的主题向量;  $A$  表示权重矩阵;  $b$  表示偏置矩阵.

完成上述操作后, 再计算主题向量和各候选词之间的相关度, 找出与当前主题最相关的候选词. 计算如下所示.

$$s_i = \{\text{similarity}(c_{il}, T_{\text{linear}})\}_{l=1}^L, i = 1, 2, \dots, k \quad (5)$$

$$m_i = (c_{il} | \max(s_i)), i = 1, 2, \dots, k \quad (6)$$

式(5)中,  $c_{il}$  表示第  $i$  个主题词  $t_i$  对应的第  $l$  个候选词;  $s_i$  表示第  $i$  个主题词与其  $L$  个候选词的相关度集合; 式(6)中,  $m_i$  表示最大相关度对应的候选词, 即与当前主题词语义最贴合的外部词.

最终, 找出和  $k$  个主题词一一对应的最相关外部词, 表示为集合  $M$ .

$$M = \{m_i\}_{i=1}^k \quad (7)$$

图 3 展示了上述步骤的执行流程, 其中无序主题词里的数字表示该主题词在词典中对应的序号.

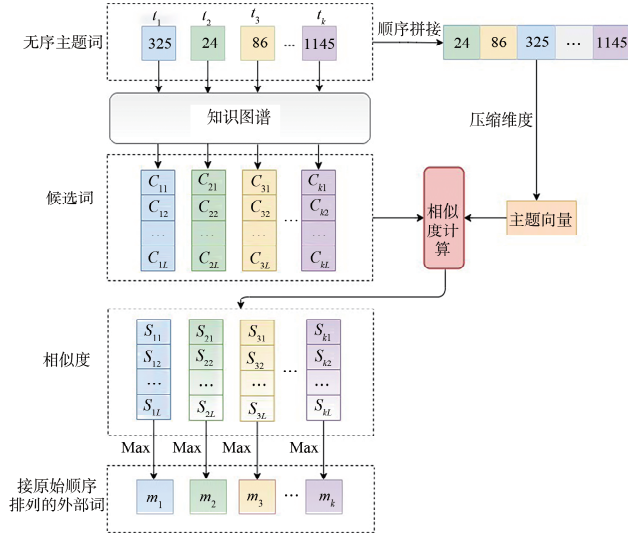


图 3 基于线性变换的候选词筛选过程

Fig. 3 Filtering process based on linear transformation

表 1 给出一个经过筛选后的示例. 由表 1 可知, 第一组主题词显然围绕苹果自身的含义展开, 自然地, “苹果”对应的外部词为“水果”; 而对于第二组主题词, 主题语义和“科技”、“电脑品牌”等含义更相关, 因此“苹果”对应的外部词为“电脑”, “现代”对应的是“经济”, 而不是“车”.

表 1 主题词及筛选后的外部词示例

Tab. 1 Examples of topic words and filtered external words

主题词	外部词
苹果 春天 好吃 生活 健康	水果 时间 喜欢 活着 有利
苹果 公司 科技 梦想 现代	电脑 商业 知识 期望 经济

### 3.2 基于注意力机制的外部权重计算方法

虽然上述过程可以筛选出和主题词语义一致的外部知识, 但如果将原主题词和外部词简单拼接后直接传入解码器, 则可能会引入不必要的噪声.

例如: 给定一组主题词[“苹果”、“春天”], 目标生成句子为“春天的苹果很好吃”. 经过上一步骤, 可得到最贴合原主题词语义的外部词组[“树”、“春”]. 其中, “春”和目标句的语义很相关, “树”却不太相符. 在 Tranformer 中, 自注意力机制虽然可以使源序列和目标序列“自身”所蕴含的信息更加

丰富,但没有对外部词和目标序列间的关系进行处理.因此,我们希望能够为每个外部词设定一个主题权重,减小“树”对应的权重,将更多的“注意力”放到“春”上.

受注意力机制<sup>[16]</sup>启发,本文提出一种计算候选词外部权重的方法,可根据期望表达的语义对权重进行动态调整.

将固定的外部词向量列表  $M$ ,转换为可根据输入信息而变化的动态词向量  $M'$ ,计算公式如下.

$$M' = \{m_i * a_i\}_{i=1}^k \quad (8)$$

其中,  $a_i$  表示给第  $i$  个外部词  $m_i$  分配的权重,由以下公式推导得出:

$$a_i = \frac{\exp(g_i)}{\sum_{i=1}^k \exp(g_i)} \quad (9)$$

$$g_i = v_a^T \tanh(W_a m_i + U_a H^{\text{enc}}) \quad (10)$$

其中,  $v_a$ 、 $W_a$  和  $U_a$  是可学习的参数矩阵或向量;  $H^{\text{enc}}$  是主题词组  $X$  经过编码器后的隐藏状态;  $g_i$  是  $m_i$  上的注意力得分.

最后,将编码器输出的主题词隐藏状态  $h_i^e$  与赋权值后的候选词  $m_i'$  进行拼接,并将其传入解码器.本文采用一一对应的拼接方式,而不是统一附着在  $H^{\text{enc}}$  之后,此时的词向量表示为

$$H^{\text{enc-m}} = \{h_1^e, m_1', h_2^e, m_2', \dots, h_k^e, m_k'\} \quad (11)$$

同时,考虑外部知识后的 TEG 任务的目标(1)可以修改为

$$\hat{\theta} = \arg \max_{\theta} P_{\theta}(Y|T, M') \quad (12)$$

### 3.3 基于多头注意力机制的内部权重计算方法

为了保证模型生成文本的主题完整性和相关性,避免某些词反复出现,而另一些词没有出现在生成文本中的情况.受主题覆盖向量<sup>[2]</sup>的启发,本文提出一种基于多头注意力机制的内部权重计算方法,以便在 Transformer 解码器内部,动态调整多头注意力分数,使模型可以更多地考虑未表达的主题词和候选词.

内部权重表示主题词(含候选词)在下一层中需要表达的程度,用于调整注意力策略.该权重向量的更新过程如式(13)所示.

$$C_{n+1} = C_n - \frac{1}{\phi} s_n(Q, K, V) \quad (13)$$

其中,  $\phi$  由公式(14)可得;  $s_n(Q, K, V)$  是解码器在第  $n$  层的多头自注意力的分数,将其和  $V$  相乘可以得到缩放点乘注意力<sup>[17]</sup>,计算如公式(15)所示.

$$\phi = N \cdot \sigma(U_f H^{\text{enc-m}}) \quad (14)$$

$$s_n(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)C_n \quad (15)$$

式(14)中,  $N$  为解码器的层数;  $\sigma$  为 sigmoid 函数;  $U_f$  为参数矩阵;  $H^{\text{enc-m}}$  为编码器输出的隐藏状态和带权重候选词的拼接向量,即  $\{h_i^e, m_i'\}_{i=1}^k$ ; 式(15)中,  $K=V=H^{\text{enc-m}}$ , 维度转换为  $d_k$ ,  $Q=H_{n-1}^{\text{dec}}$ , 表示第  $n-1$  层解码器的输出.

此时,考虑内部权重后的 TEG 任务的目标(12)可以修改为

$$\hat{\theta} = \arg \max_{\theta} P_{\theta}(Y|T, M', C_n) \quad (16)$$

## 4 实验

### 4.1 基本设置

4.1.1 数据集 本文采用文献[2]提出的 ESSAY 数据集,它是 TEG 任务中的一个高质量公开数据集,包括 320 万篇中国初中或高中作文,每个句子有 5 个主题.我们选用了其中长度在 40~50 间的 50 000 条数据作为训练集,5000 条作测试集.文本最大长度设为 50,不足的部分用 '<pad>' 填充,词表外的词语用 '<unk>' 填充.

4.1.2 模型设置 本文使用 Word2Vec<sup>[18]</sup> 进行词嵌入,得到低维实数向量表示<sup>[19]</sup>.选用 HowNet<sup>[20]</sup> 知识图谱作为外部知识库,它在自然语言处理方面,更贴近语言的本质特点,在融入学习模型方面,也具有无可比拟的优势<sup>[21]</sup>.我们将主题词的邻近词作为候选信息,并去除和索引相同的词,候选词的最大数量设为 5.采用余弦相似度方法<sup>[22]</sup> 计算主题向量与各候选词之间的相似度.模型的参数设置如表 2 所示.

表 2 模型参数表

Tab. 2 Model parameter table

参数名	数值
batch_size	128
epochs	80
d_model	512
num_heads	8
num_layers	6
dff	2048
dropout	0.1
词典大小	48 255

### 4.2 基准模型

本文选择了以下基准方法与论文提出的模型

Trans-K 作比较:

(1)TAV-LSTM:利用所有主题词的平均加权 and 嵌入来表示主题语义,使用长短期记忆网络(Long Short-Term Memory, LSTM)<sup>[23]</sup>作为编/解码器<sup>[2]</sup>;

(2)TAT-LSTM:使用注意力机制对主题词和生成词之间的语义关系进行建模<sup>[2]</sup>;

(3)MTA-LSTM:使用主题覆盖向量来调整注意机制,使模型更加关注未表达的主题词<sup>[2]</sup>;

(4)MTAK-LSTM:自定义模型.在 MTA-LSTM 模型的基础上引入外部知识,具体的做法是将 HowNet 中提取的外部知识和源主题词一一拼接后传入解码器,即传入模型的主题词向量  $T = \{t_1, m_1, t_2, m_2, \dots, t_k, m_k\}$ .

### 4.3 评价指标

#### 4.3.1 自动评估

(1)BLEU:双语评估替补(BLEU)是机器翻译的一个自动评估指标<sup>[24]</sup>.使用训练集作为参考,计算 BLEU 值来评估生成的文本.本文选取 BLEU-2、BLEU-3 和 BLEU-4 的分数进行比较.分数越高,生成文本的精度(流畅性)越好.

(2)Back-BLEU:使用生成的文本作为参考,计算 BLEU 值来评估训练集中的文本.本文选取 Back-BLEU-2 的值进行比较,表中缩写为 B-BLEU.分数越高,生成文本的召回率(多样性)越好.

4.3.2 人工评估 由 5 名中文专家对每个模型生成的 100 个随机样本进行主观评价.包括:完整性(生成文本是否完整)、相关性(生成文本与主题词是否相关)、流畅性(在语法和句法上是否结构良好)、连贯性(是否具有主题和逻辑结构)四个评价维度<sup>[2]</sup>.每个维度给出一个 1~5 分之间的分数,并计算平均值作为最终得分.

#### 4.4 实验结果分析

自动评估结果如表 3 所示.结果表明,Trans-K 的各项指标均表现为最优.证明了本文提出的方法可以有效提高生成文本的质量.特别的是,与最佳基准模型相比,Trans-K 在 BLEU-4 上提高了 58.12,在 Back-BLEU 上提高了 20.53.证明生成文本的流畅性和多样性都得到了显著提升.

经过进一步观察,不难发现:随着  $n$ -gram 匹配片段的增大,上述模型的 BLEU 值均有不同程度的减小.其中,基于 LSTM 基准模型的下降速率较快,基于 Trans-K 的下降趋势却相对平缓.这可能

是因为在源信息远小于生成信息量的情况下,Transformer 是基于词语间的相对距离来构建模型的,能够较 LSTM 模型更好地捕获文本中的长依赖关系.另外,一个有趣的发现是,在 MTA-LSTM 模型上引入外部知识后(MTAK-LSTM)的各指标都不如原模型.

表 3 自动评估结果

Tab.3 Automatic evaluation results

模型	BLEU-2/%	BLEU-3/%	BLEU-4/%	B-BLEU/%
TAV-LSTM	78.4	50.67	26.02	56.23
TAT-LSTM	78.55	51.05	26.18	69.56
MTA-LSTM	79.14	51.47	26.3	69.41
MTAK-LSTM	75.67	48.41	24.18	66.53
Trans-K	93.22	88.38	84.42	90.09
Impv-Best	14.08	36.91	58.12	20.53

表 4 给出了人类评估结果,从中可以得出类似的结论.显然,Trans-K 模型优于基线,尤其是在主题一致性方面.例如:与最佳基线相比,该模型的相关性得分提高了 0.72,连贯性得分提高了 0.69.这是因为外部知识库提供的额外背景信息可以在一定程度上丰富源信息的语义知识,进而帮助模型生成更加相关、连贯的句子.

表 4 人工评估结果

Tab.4 Manual evaluation results

模型	完整	相关	流畅	连贯	平均
TAV-LSTM	2.18	2.34	2.69	2.42	2.41
TAT-LSTM	2.35	2.65	2.84	2.77	2.65
MTA-LSTM	2.61	2.83	3.15	3.23	2.96
MTAK-LSTM	2.57	2.84	3.11	2.81	2.83
Trans-K	3.25	3.56	3.78	3.92	3.63
Impv-Best	0.64	0.72	0.63	0.69	0.67

下面对比各模型的训练效果.

图 4 是训练过程的 Loss 值下降图,可以看出 TAT-LSTM 模型的收敛效果最差,Trans-K 的收敛效果优于其他基准模型.当模型趋于平稳时,Trans-K 的 Loss 值明显低于其他模型的 Loss 值.

图 5 较直观地对比了训练 80 轮后,各模型生成文本的准确率.准确率由生成文本和参考文本的张量进行逐元素比较得到,计算公式如下.

$$ACC = \frac{\text{corrects}}{\text{real}} \quad (17)$$

其中,corrects 表示两个张量中,索引位置和对应

数值均相同的元素数量;real 表示在参考文本张量中,去除 '<pad>' 填充位的元素数量。

由图 5 可知,当模型训练 80 轮时,Trans-K 生成文本的准确率最高,TAV-LSTM 的表现最差,引入外部知识后的模型 MTAK-LSTM 在准确率方面较 MTA-LSTM 有轻微的提升。

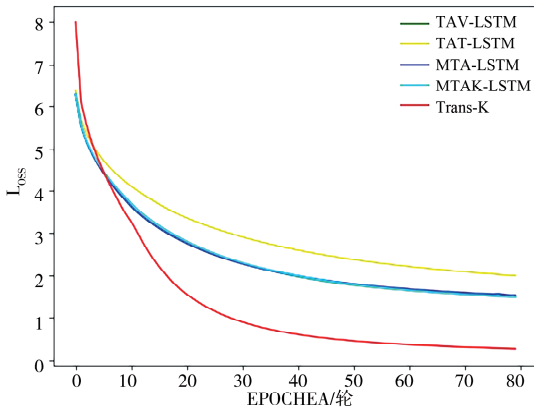


图 4 训练过程的 Loss 值下降图

Fig. 4 Loss value decline diagram during training

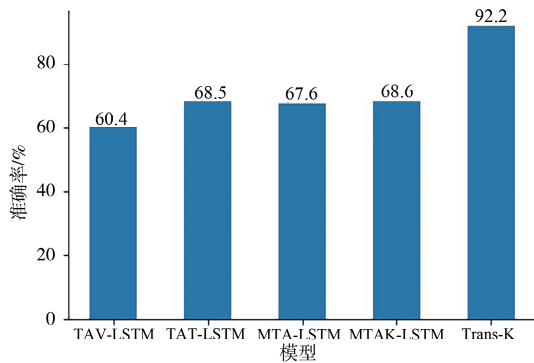


图 5 训练 80 轮后的准确率对比图

Fig. 5 Comparison of accuracy after 80 rounds of training

## 4.5 消融实验

4.5.1 关键组成部分研究 为了更直观地验证原文所述方法的可行性和重要性,在本节进行以下消融实验:

(1) Transformer: 最基础的 Transformer 模型,但编码器取消了位置编码。

(2) (w/o) LM: 筛选和主题词语义一致的候选词时,采用文献[3]提出的方法,将所有非当前主题词的平均词向量作为主题向量,并将其与候选词做余弦相似度计算。

(3) (w/o) EW-1: 取消基于注意力机制的外部权重,即 $\alpha_i = 1$ 。

(4) (w/o) EW-2: 计算基于注意力机制的外部权重时,将编码器的输出替换为目标句经过位置

编码后的隐藏状态,即 $g_i = v_a^T \tanh(W_a m_i + U_a H^{\text{pos}})$ 。

(5) (w/o) IW: 取消基于多头注意力机制的内部权重,即 $C_n = 0$ 。

上述模型的参数设置和 Trans-K 一致。表 5 和表 6 分别给出了消融实验的自动和人类评估结果。

表 5 消融实验自动评估结果

Tab. 5 Automatic evaluation results of ablation experiment

模型	BLEU-2/%	BLEU-3/%	BLEU-4/%	B-BLEU/%
Transformer	86.53	73.62	63.07	89.79
(w/o) LM	91.24	86.08	81.87	90.09
(w/o) EW-1	92.56	87.81	84.03	90.08
(w/o) EW-2	92.12	86.16	81.2	90.09
(w/o) IW	92.36	86.25	81.22	90.08
Trans-K	93.22	88.38	84.42	90.09

分析消融实验的自动评估结果,由 Transformer 的结果可知,基础的 Transformer 模型效果已经优于表 3 中的最优基准模型 MTA-LSTM,证明了自注意力机制的有效性;另一方面,它的 BLEU 值明显低于 Trans-K 的结果,其中 BLEU-4 值减小了 21.35,说明引入外部常识知识能够为模型提供更加丰富的语义知识,在一定程度上补充源信息。

由(w/o)LM 和 Trans-K 的结果对比可知,两者的 Back-BLEU 大致相同,但前者的 3 个 BLEU 值均小于完整模型,说明本文采用的方法可以更好地表达主题语义,采用平均词向量作为主题向量的方法存在语义模棱两可、不明确的可能性。

由(w/o)EW-1 和 EW-2 的结果,讨论外部词的噪声对模型生成效果的影响。对比(w/o)EW-1 和 Trans-K 的评估结果,发现前者的各项指标均低于后者,说明为外部词设定外部权重可以有效减少不必要的噪声,使引入的外部知识更加贴近主题;而(w/o)EW-2 和 Trans-K 的结果则证明了,编码器输出的隐藏状态与经过位置编码的目标生成文本相比,前者更适合与候选词做注意力计算。

最后,对比(w/o)IW 和 Trans-K 的评估结果,完整模型的各项指标均优于未加入多头覆盖向量的模型,证明了基于多头注意力机制的内部权重可以提高模型生成文本的质量和多样性。

消融实验的人工评估结果也表明,本文提出的

方法可以使模型性能得到有效提升。

表 6 消融实验人工评估结果

Tab. 6 Manual evaluation results of ablation experiment

模型	完整	相关	流畅	连贯	平均
Transformer	2.93	3.17	3.24	3.36	3.18
(w/o) LM	3.07	3.29	3.32	3.58	3.32
(w/o) EW-1	3.22	3.51	3.7	3.83	3.57
(w/o) EW-2	3.1	3.37	3.51	3.62	3.4
(w/o) IW	3.17	3.23	3.66	3.78	3.46
Trans-K	3.25	3.56	3.78	3.92	3.63

4.5.2 外部词检索方式研究 HowNet 里所有词语(sense)的含义可以由更小的语义单位,即“义原”(sememe)构成<sup>[20]</sup>.可总结出以下三种外部词检索方式:

(1) Sememe-T:选择义原树最顶端的义原作为主要的语义信息。

(2) Sememe-A:选择义原树所有的义原作为扩展语义信息。

(3) Sense:选择主题词的邻近词作为外部词,即本文采用的检索方式。

将 3 种方法分别应用到 Trans-K 模型,除检索方式不同,其他设置均相同.结果如表 7 所示.和表 5 中的 Transformer 相比,3 个模型的各项指标均有一定的提升,证明无论采用哪种检索方式,引入外部知识都是有意义的;后两种检索方式的结果大致相同,说明外部词的检索方式对模型生成效果没有太大的影响,本文提出的方法具有普适性。

表 7 采用不同检索方式的模型效果

Tab. 7 Model effects with different retrieval methods

检索方式	BLEU-2/%	BLEU-3/%	BLEU-4/%	B-BLEU/%
Sememe-T	87.63	75.84	66.29	90.08
Sememe-A	92.34	88.3	85.09	90.09
Sense	93.22	88.38	84.42	90.09

## 5 结 论

本文提出了一个融合外部知识的主题文本生成模型 Trans-K,其编码器和解码器基于 Transformer 结构,缓解了传统采用 RNN 方法带来的长依赖问题,减少了时序特性对生成结果的影响.该模型首先从外部常识知识库中检索出和主题词有关的多个候选词,再通过基于线性变换的主题向量

计算方法,筛选出和主题词语义一致的外部词;其次,通过基于注意力机制的外部权重计算方法,为外部词赋予主题权重,并将编码器的隐藏状态和赋值后的外部词一一拼接后传入解码器;最后在解码器计算多头自注意力分数时,引入内部权重。

在公用数据集 EASSY 上训练模型,自动评估和人工评估结果均表明,Trans-K 模型优于当前主题文本生成任务中的基线模型;在消融实验中证明了模型中各关键模块的有效性,并对比了不同检索候选词方法对模型的影响。

未来,我们尝试将更多类型或风格的外部知识引入到 TEG 任务中,例如:新闻、小说等.除此之外,还可进一步思考如何提升生成文本的多样性,考虑引入修辞手法、语法结构等特征,生成更加符合人类书写风格的内容。

## 参考文献:

- [1] Zhang J H, Chen J J. Summarization of natural language generation [J]. Appl Res Comp, 2006, 23: 1. [张建华, 陈家骏. 自然语言生成综述[J]. 计算机应用研究, 2006, 23: 1.]
- [2] Feng X, Liu M, Liu J, *et al.* Topic-to-essay generation with neural networks [C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, IJCAI-18. Stockholm: Morgan Kaufmann, 2018: 4078.
- [3] Luo D, Ning X, Wu C. Sememe-based topic-to-essay generation with neural networks [EB/OL]. [2022-12-20]. <https://iopscience.iop.org/issue/1742-6596/1861/1JPCS>.
- [4] Mikolov T, Kombrink S, Deoras A, *et al.* RNNLM-recurrent neural network language modeling toolkit [C]// Proceedings of the 2011 ASRU Workshop. Hawaii: IEEE, 2011: 196.
- [5] Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need [C]// Proceedings of the Advances in neural information processing systems 30. Long Beach, California: Morgan Kaufmann, 2017: 5998.
- [6] Yang P, Li L, Luo F, *et al.* Enhancing topic-to-essay generation with external commonsense knowledge [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019: 2002.
- [7] Liu Z, J Wang, Li Z. Topic-to-essay generation with comprehensive knowledge enhancement [C]// Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Da-

- tabases. Bilbao: ECML PKDD, 2021: 302.
- [8] Pascual D, Egressy B, Meister C, *et al.* A Plug-and-Play Method for Controlled Text Generation [C]//Empirical Methods in Natural Language Processing. [S. l.]: ACL, 2021: 3973.
- [9] Li Z C. Text generation algorithm based on keyword semantic control [D]. Beijing: Beijing University of Posts and Telecommunications, 2019. [李作潮. 基于关键词语义控制的文本生成算法研究[D]. 北京: 北京邮电大学, 2019.]
- [10] Sun C H, Hu B, Zou Y X. A BP-LSTM trend forecast model for stock index [J]. J Sichuan Univ(Nat Sci Ed), 2020, 57: 27. [孙存浩, 胡兵, 邹雨轩. 指数趋势预测的 BP-LSTM 模型[J]. 四川大学学报(自然科学版), 2020, 57: 27.]
- [11] He W, Rao Y. Transformer-based hierarchical topic-to-essay generation [EB/OL]. [2022-12-20]. <https://www.sciencedirect.com/science/article/pii/S1877050922005920>.
- [12] Zhang X, Lapata M. Chinese poetry generation with recurrent neural networks[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha: ACL, 2014: 670.
- [13] Wang Z, He w, Wu H, *et al.* Chinese poetry generation with planning based neural network[C]//Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics; Technical Papers. Osaka: [s. n.], 2016: 1051.
- [14] Liu Z, Fu Z, Cao J, *et al.* Rhetorically controlled encoder-decoder for modern chinese poetry generation [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019: 1992.
- [15] 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. [S. l. :s. n. ], 2021: 4784.
- [16] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2022-12-20]. <https://arxiv.org/abs/1409.0473>.
- [17] Wang J W, Yang X C, Ju S G, *et al.* Text classification model based on convolutional neural network and self-attention mechanism [J]. J Sichuan Univ (Nat Sci Ed), 2020, 57: 469. [汪嘉伟, 杨煦晨, 琚生根, 等. 基于卷积神经网络和自注意力机制的文本分类模型[J]. 四川大学学报(自然科学版), 2020, 57: 469.]
- [18] Goldberg Y, Levy O. word2vec Explained: deriving Mikolov *et al.*'s negative-sampling word-embedding method [EB/OL]. [2022-12-20]. <https://arxiv.org/abs/1402.3722>.
- [19] Liu G F, Huang X Y, Liu X Y. Document sentiment modeling based on topic attention hierarchy memory network [J]. J Sichuan Univ(Nat Sci Ed), 2019, 56: 833. [刘广峰, 黄贤英, 刘小洋, 等. 基于主题注意力层次记忆网络的文档情感建模[J]. 四川大学学报(自然科学版), 2019, 56: 833.]
- [20] Dong Z, Qiang D. How net and the computation of meaning [M]. [S. l.]: World Scientific, 2006.
- [21] Niu Y, Xie R, Liu Z, *et al.* Improved word representation learning with sememes[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver: ACL, 2017: 2049.
- [22] Wang C L, Yang Y H; Deng F, *et al.* Institute of Computer A Review of Text Similarity Approaches [J]. Inform Sci, 2019, 37: 158. [王春柳, 杨永辉, 邓霏, 等. 文本相似度计算方法研究综述[J]. 情报科学, 2019, 37: 158.]
- [23] Sundermeyer M, R Schlüter, Ney H. LSTM neural networks for language modeling[C]// Thirteenth annual conference of the international speech communication association. Portland, Oregon, USA: [s. n.]. 2012: 194.
- [24] Papineni K, Roukos S, Ward T, *et al.* Blue: a method for automatic evaluation of machine translation [C]// Proceedings of the Meeting of the Association for Computational Linguistics. Philadelphia: ACL, 2002: 311.