

基于主题约束的篇章级文本生成方法研究

黄炎^{1,2} 徐科^{1,3} 余晓阳¹ 王同洋^{1,†} 张新访¹ 路松峰^{1,2}

1. 华中科技大学计算机科学与技术学院, 武汉 430074; 2. 深圳华中科技大学研究院, 深圳 518063;
3. 中南民族大学计算机科学学院, 武汉 430074; † 通讯作者, E-mail: platanus@hust.edu.cn

摘要 针对计算机自动生成的文本缺乏主题思想问题, 提出一种基于主题约束的篇章级文本自动生成方法。该方法围绕用户输入的主题描述语句提取若干主题词; 然后对主题词进行扩展和主题聚类, 形成文章主题规划; 最后, 利用每个聚类中的关键词信息约束每个段落的文本生成。该模型从文本主题分布、注意力评分方法和主题覆盖生成三个方面对现有基于注意力机制的循环神经网络文本生成模型进行了改进。实验部分在三个真实数据集上分别与 Char-RNN、SC-LSTM 及 MTA-LSTM 基准模型对比并对三种改进进行了独立验证分析, 实验结果表明该方法在人工评判和 BLEU 自动评测上均优于基准模型, 生成的文本能够较好地贴合主题。

关键词 文本自动生成; 主题约束; 循环神经网络; 长短时记忆网络; 注意力机制

中图分类号 TP391

A Research of Discourse-level Text Generation Method Based on Topical Constraint

HUANG Yan^{1,2}, XU Ke^{1,3}, YU Xiaoyang¹, WANG Tongyang^{1,†}, ZHANG Xinfang¹, LU Songfeng^{1,2}

1. School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074;
2. Shenzhen Huazhong University of Science and Technology Research Institute, Shenzhen, 518063;
3. School of Computer Science, South-Central University for Nationalities, Wuhan, 430074;
† Corresponding Author, E-mail: platanus@hust.edu.cn

Abstract To solve the topic missing problem of text generated by computers, this paper proposed a new discourse-level text generation method based on topical constraint. Providing a short topic description, the approach extracted several topic words from the text, then extended and clustered the keywords to form topical planning which were used to restrain the generation of each paragraphs. The model improved the attention based recurrent neural network from three aspects including topic distribution, attention scoring function and topic coverage generation. In experiments, the paper compared the proposed method with benchmark models such as Char-RNN, SC-LSTM and MTA-LSTM on three real datasets, three improvement strategies are verified and analysed independently, experimental results show that the model is more efficient than benchmark models on human and BLEU metrics, the generated text can catch the topic effectively.

Key words automatic text generation; topical constraint; RNN; LSTM; attention mechanism

文本自动生成是自然语言处理领域的一项重要研究和挑战, 旨在让计算机像人类一样学会写作写出高质量的自然语言文本, 作为核心技术广泛应用于机器翻译、文本摘要、问答系统和对话系统等方面。按照输入类型划分, 文本自动生成^[1]分为从文本到文本生成、从意义到文本生成、从数据到文本生成、从图像到文本生成。

针对这一研究领域, 国内外学者提出了很多研究方法和模型。传统的基于规则和模板的生成方法^[2]和基于信息抽取的方法^[3]生成的文本格式固定, 缺乏语义信息, 不能产生内容丰富、风格多样的文本。随着深度学习技术的发展, 众多基于深度神经网络模型的文本生成方法被提出来。Hinton 等人^[4]将 HF 优化算法结合循环神经网络模型应用于字符级语言建模任务, 验证了 RNN 模型生成文本的有效性。文献[5]将长短时记忆网络与 RNN 模型结合用于序列生成, 以学习长时依赖关系。为了模拟人们日常写作过程中短语复用行为, 文献[6]提出在序

列到序列模型中引入 copy 机制。文献[7]基于循环神经网络针对语句连贯性进行建模，支持语句的重新排序并生成语言连贯的文本内容。论文[8]提出用一种神经清单模型解决循环神经网络生成文本过程中缺乏对已经生成内容的记忆问题。

然而，基于深度学习生成的文本内容通常存在主题不明确和语句不通顺等问题。已有解决办法包括通过给定的一个或者多个主题词，显性地将主题词加入到生成文本的适当位置^[9]，根据主题词挑选主题相似的词语或语句重新组合生成新文本^{[10][14]}，或者采用基于注意力机制的 RNN 模型根据主题词控制新文本内容的生成^{[12][13]}。这些方法可以在一定程度上缓解以上问题，但是仍存在主题单一、主题分布不可控或主题词覆盖生成问题。

为了解决上述问题，本文提出一种基于主题约束的篇章级文本自动生成模型。该模型针对用户输入的主题描述语句，采用 Twitter LDA 概率主题模型提取若干主题词；然后采用 Word2Vec 词向量相似度计算方法对主题词进行关键词扩展；接着对扩展后的关键词集进行主题聚类，形成文章主题规划；最后，采用基于注意力机制的循环神经网络，利用每个聚类中的关键词信息约束每个段落生成的文本内容。根据每个主题聚类生成一个段落，生成主题多方面的文本信息，实现篇章级文本生成。该模型依据主题关联程度为每个关键词设置不同的权重，从而控制生成文本的主题分布。另外，模型采用一种改进的注意力评分机制，依据前文与每个主题的相似度调整注意力评分，可以平衡多个主题的影响。为了提高主题词在生成文本中的出现可能性，针对主题词添加一个生成概率附加项。

本文基于作文语料、知乎语料和百度百科语料进行文本生成实验，并与 Char-RNN、SC-LSTM 及 MTA-LSTM 基准模型进行对比，采用人工评判和 BLEU 自动评测方法，实验结果表明了该方法可行而且有效。

1 篇章级文本生成框架

本文所讨论的篇章级文本自动生成任务主要依据用户输入的一段主题描述性文本进行理解和分析，计算机自动生成包含多个段落的篇章结构文本内容。要求生成的文本整体语义结构完整，每个段落能够表达关于主题的一个不同方面的内容。

针对该任务，本文采用的篇章级文本生成框架如图 1 所示。首先，针对用户输入的主题描述文本“家是心灵与身体的归宿，家是遮风避雨的港湾，家是温暖与快乐的源泉”，采用 Twitter LDA 概率主题模型提取若干主题词，如“家”、“归宿”、“港湾”、“源泉”、“快乐”等。然后参考论文[11]进行主题扩展，基于 Word2Vec 预训练好的词向量采用余弦相似度计算与这几个主题词相近的词语，如“生命”、“养育”、“亲情”、“依靠”等构成关键词集；接着采用 K-means 方法对关键词集进行主题聚类，每个聚类包含若干个关键词，形成文章的主题规划，如本例生成 4 个聚类，每个聚类表示主题的某一方面信息。最后，与论文[11]中进行主题理解后选择相似语句并重新排序生成新文本不同，本文根据每个聚类中的关键词，采用基于注意力机制的循环神经网络模型，生成对应段落的文本内容。本例 4 个聚类分别生成 4 个对应的段落，每个段落阐述主题的一个方面内容。

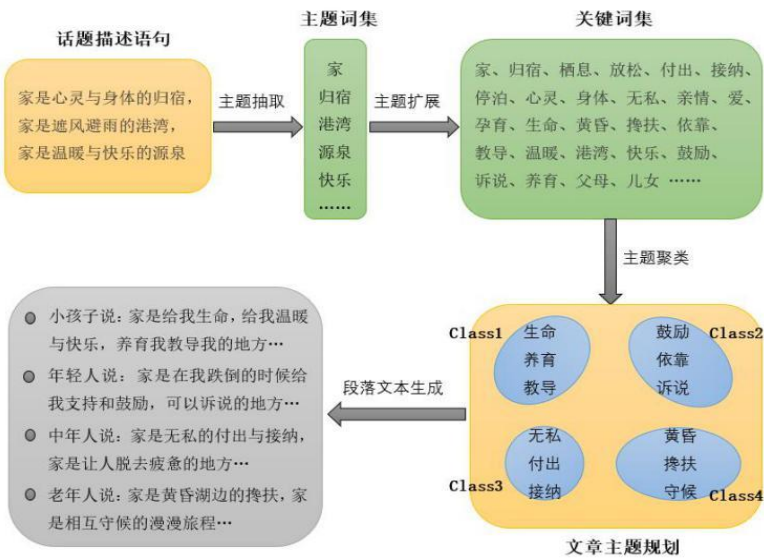


图 1 篇章级文本生成框架图

Fig.1 Discourse-level Text Generation Framework

2 基于主题约束的文本生成模型

2.1 任务定义及建模

基于主题聚类约束生成段落文本任务定义如下：给定一个包含 k 个关键词的主题聚类 $T = \{topic_1, topic_2, \dots, topic_k\}$ ，根据该主题聚类中的关键词自动生成一个长度为 ℓ 的段落文本内容。假设生成的段落文本全部来源于词汇表 $Dict$ ，其中包含 m 个不同词汇，即 $Dict = \{word_1, word_2, \dots, word_m\}$ 。生成的段落文本内容 $Text = \{x_1, x_2, \dots, x_\ell\}$ 。那么，任务可以形式化描述为最大化条件概率 $P(Text|T)$ 的概率语言模型：

$$\begin{aligned} & \text{maximize } P(x_1, x_2, \dots, x_\ell | topic_1, topic_2, \dots, topic_k) \\ & \text{subject to:} \\ & \quad x_i \in Dict, 1 \leq i \leq \ell \\ & \quad t_i \in Dict, 1 \leq i \leq k \\ & \quad k \leq \ell \leq m \end{aligned} \quad (1)$$

2.2 文本生成模型

鉴于深度神经网络模型的强大表示学习能力和以上高效性实证研究，本文依然采用基于循环神经网络模型为基础，并模拟人类大脑注意力行为，将主题聚类信息以注意力机制约束文本的生成。在此基础上本文分别就生成文本的主题分布、注意力评分方式和主题词覆盖生成问题提出相应的改进策略。

2.2.1 基于注意力机制的 RNN 模型

首先，我们通过 word2vec 预训练词向量获取词汇表 $Dict$ 中每个词汇的向量表示，并从中得到每个关键词 $topic_i, 1 \leq i \leq k$ 和段落中每个词汇 $x_j, 1 \leq j \leq \ell$ 的向量表示。接着，将段落词汇按序列输入到 RNN 模型，假设 t 时刻的输入为 x_t ，我们令 t 时刻的输出等于下时刻的输入，即 $y_t = x_{t+1}$ 。由于 LSTM^[16]（长短时记忆网络）能够学习语句序列的长时依赖关系并且在解码阶段具有较好的性能，本文采用双层 LSTM 作为 RNN 模型每一时刻的网络结构。

如此，我们的条件概率语言模型公式(1)通过如下方式计算：

$$\begin{aligned} P(Text|T) &= P(x_1, x_2, \dots, x_\ell | T) \\ &= P(x_1 | T) \prod_{t=2}^{\ell} P(x_t | T, x_2, \dots, x_{t-1}) \\ &= P(x_1 | T) \prod_{t=2}^{\ell} P(x_t | T, x_{t-1}, h_{t-1}) \end{aligned}$$

下一个词汇的出现概率定义如下：

$$\begin{aligned} P(x_t | T, x_{t-1}, h_{t-1}) &= \text{softmax}(g(h_t)) \\ h_t &= f(h_{t-1}, x_{t-1}, T_t) \end{aligned}$$

其中 $g(\cdot)$ 是一个全连接层， $f(\cdot)$ 由上述 LSTM 网络结构决定。

此外，我们采用基于主题的注意力机制^[15]来约束 LSTM 神经网络生成的文本，具体而言计算主题词向量与生成词汇之间的相似度评分并排序，挑选评分较高的词汇作为 LSTM 输出，从而使得生成的文本内容更偏向我们选择的主题。对每个主题词 $topic_1 \sim topic_k$ ，通过 Attention^[13] 计算转化为相应的注意力得分 $\alpha_1, \alpha_2, \dots, \alpha_k$ 。那么 t 时刻的主题向量表示 T_t 可以表示如下：

$$T_t = \sum_{j=1}^k \alpha_{tj} topic_j$$

其中 $topic_j$ 表示第 j 个主题词的词向量， α_{tj} 表示 t 时刻第 j 个主题词的注意力得分。

$$\begin{aligned} \alpha_{tj} &= \frac{\exp(g_{tj})}{\sum_{i=1}^k \exp(g_{ti})} \\ g_{tj} &= \eta(h_{t-1}, topic_j) \\ &= v_a^T \tanh(W_a h_{t-1} + U_a topic_j) \end{aligned} \quad (2)$$

$\eta(\cdot)$ 通常采用 MLP（多层感知器）实现并采用 \tanh 作为激活函数。

2.2.2 文本主题分布

为了使得模型生成的文本内容能够覆盖所有主题，我们参考文献[13]引入 k 维主题覆盖向量 $C_t = [C_{t,0}, C_{t,1}, \dots, C_{t,k}]$ ，对应主题聚类 T ，其中 k 是主题词个数， $C_{t,j}$ 表示 t 时刻第 j 个主题词的权重。然而，文献[13]将所有主题词同等对待，初始时刻 C_t 的权重全部设为 1.0，即 $C_0 = [1.0, 1.0, \dots, 1.0]$ ，没有考虑文本对不同主题的概率分布情况。因此，本文提出一种基于文本主题分布的改进方法，为每一个主题 $topic_j$ 设置 $[0,1]$ 之间的初始权重：

$$C_{0,j} = \frac{\text{score}(\text{topic}_j)}{\max_{1 \leq i \leq k} \text{score}(\text{topic}_i)}$$

其中 $\text{score}(\text{topic}_j)$ 是通过 Twitter LDA 算法从原始文本中抽取主题 $topic_j$ 所计算出来的得分，该得分表示主题在文本中的概率分布。

我们通过 C_t 调整注意力评分来确保主题覆盖率：

$$g_{tj} = C_{t-1,j} v_a^T \tanh(W_a h_{t-1} + U_a \text{topic}_j) \quad (3)$$

模型在生成文本时不断调整主题权重 $C_{t,j}$ 以确保表达不充分的主题得到充分表达：

$$C_{t,j} = C_{t-1,j} - \frac{\alpha_{t,j}}{\phi_j}$$

$$\phi_j = N \cdot \sigma(U_f[T_1, T_2, \dots, T_k]), U_f \in \mathbb{R}^{kd_w}$$

其中 $\alpha_{t,j}$ 表示 t 时刻第 j 个主题词的注意力得分， N 表示长度为 ℓ 的文本 $Text$ 中有意义（非 PAD）词汇个数， d_w 表示词向量维度， U_f 是一个 $[k, d_w]$ 的矩阵， T_k 表示主题词向量， $\sigma(\cdot)$ 是 sigmoid 函数。

2.2.3 注意力评分方法

现有基于注意力机制^{[15][13]}的生成模型通常采用多层感知器来计算主题词的注意力得分 g_{tj} ，见公式(2)(3)。本文提出一种新的注意力评分方法，在原来基础上添加一个基于上文与主题相似度的惩罚项，公式如下：

$$g_{tj} = C_{t-1,j} [v_a^T \tanh(W_a h_{t-1} + U_a \text{topic}_j) - \beta * \text{similarity}(h_{t-1}, \text{topic}_j)]$$

其中 $\text{similarity}(\cdot)$ 是相似度计算函数，本文采用余弦相似度， $\beta \in (0,1)$ 用于平衡减号前后两项的值。如此，前文在 $topic_j$ 上的关注越高，后文生成的内容在 $topic_j$ 上的关注度则适当降低，从而可以平衡多个主题的影响。

2.2.4 主题词覆盖生成

基于主题约束的文本自动生成任务中我们通常需要生成的文本内容与给定的主题词强相关甚至直接包含部分主题词。我们参考文献[12]为每个词汇 w_i 的生成概率添加一个附加项：

$$P(x_t|T, x_{t-1}, h_{t-1}) = P_V(x_t|T, x_{t-1}, h_{t-1}) + P_K(x_t|T, x_{t-1}, h_{t-1})$$

以提高主题词的生成可能性。

$$P_V(x_t = w|T, x_{t-1}, h_{t-1}) = \begin{cases} \frac{1}{Z} e^{g_V(h_t)}, & w \in V \cup K \\ 0, & w \notin V \cup K \end{cases}$$

$$P_K(x_t = w|T, x_{t-1}, h_{t-1}) = \begin{cases} \frac{1}{Z} e^{g_K(h_t)}, & w \in K \\ 0, & w \notin K \end{cases}$$

其中，

$$h_t = f(h_{t-1}, x_{t-1}, T_t)$$

$$Z = \sum_{w \in V} e^{g_V(h_t)} + \sum_{w \in K} e^{g_K(h_t)}$$

$g_V(h_t)$ 和 $g_K(h_t)$ 分别是两个不同参数的全连接层，这里为了符号与参考文献[12]保持一致， V 指代响应词汇表 $Dict$ ， K 指主题词表 T 。

因此，每个词汇的生成概率取决于是否是关键词，对于非关键词，生成概率不发生变化，对于关键词，生成概率会加上一个附加项 $P_K(x_t|T, x_{t-1}, h_{t-1})$ 以提高生成文本中关键词的出现概率。同时，该方法可以优化首个词汇的选择，使其更准确，从而提高后续内容和整体文本的生成效果。

2.3 模型训练

本文模型最终转化为以下优化问题求解：

maximize $P(x_1|T) \prod_{t=2}^{\ell} P(x_t|T, x_{t-1}, h_{t-1})$

subject to:

$x_i \in Dict, 1 \leq i \leq \ell$

$t_i \in Dict, 1 \leq i \leq k$

$k \leq \ell \leq m$

我们通过 AdaDelta^[17]算法求解该优化问题，并迭代训练同时获得使模型最优的基于注意力机制的 LSTM 模型参数和主题建模生成参数。

3 实验验证与对比分析

为了验证本文基于主题约束的段落文本生成模型和篇章级文本生成框架的有效性，我们分别在作文语料、知乎语料和百度百科语料上进行模型训练和预测段落文本生成效果，并与 Char-RNN、SC-LSTM 及 MTA-LSTM 模型对比分析。同时，为了独立验证每个改进策略对模型的影响效果，我们对每个改进算法进行了实证研究和分析。最后，基于训练好的模型采用篇章级文本生成框架生成文章并进行人工验证分析。

3.1 数据集

我们使用 ESSAY 数据集和 ZhiHu 数据集^[13]进行模型的训练和验证测试，其中，ESSAY 数据集通过爬取和整理中国高考优秀作文而来，ZhiHu 数据集通过爬取知乎网站上的文章和相应主题词而来。同时，通过爬取百度百科网站上的自然科学语料构建了 BaiKe 数据集。考虑到数字和英文语料的不足，难以学习到有效的向量表示，针对三种语料，我们分别去除了包含数字、英文的样本，过滤掉了特殊字符和表情符号，将英文标点符号全部转换为中文标点符号，使得最终的样本只包含中文词汇和中文标点，并限制每个样例的文本长度在 50 至 200 之间。三种数据集的统计结果如下表 1 所示：

表 1 数据集统计表

Table 1 Statistics of datasets

数据集	训练集	测试集	分词	词向量
ESSAY	385662	5000	Jieba	244951
ZhiHu	38789	3000	HanLP	56900
BaiKe	200000	5000	HanLP	182586

针对过滤之后的样本，我们分别采用 Jieba 分词和 HanLP 分词工具进行中文文本分词，并采用 Twitter LDA 算法提取 5 个关键词作为主题词并计算每个主题词的概率分布。另外，针对 ESSAY 数据集训练了 word2vec 词向量，每个词向量的维度设为 300。由于 ZhiHu 和 BaiKe 语料已经存在很多预训练词向量集合，为了方便，我们直接使用文献[19]提供的预训练词向量集合。

3.2 对比模型

我们分别与以下基准模型进行段落文本生成实验对比：

Char-RNN: 基于循环神经网络的逐字符文本生成模型，将输入文本中的每个字符按照顺序进入 RNN 网络模型，每个字符传入 RNN 之后输出紧跟其后的一个字符。该模型无需对原始文本进行分词，我们采用双层 LSTM 作为 RNN 模型基础结构。

SC-LSTM: 该模型来源于文献[16]，基于语义控制的 LSTM 结构进行统计语言生成，模型引入对话行为的 one-hot 主题向量覆盖机制，使得生成的文本包含特定主题信息。

MTA-LSTM: 文献[13]提出的基于主题的作文生成模型，采用多主题覆盖向量 coverage vector 调整注意力评分来确保主题覆盖率并在解码阶段不断调整主题权重。

为了独立验证文本主题分布、注意力评分方法和主题覆盖生成三种改进对基于注意力机制的 LSTM 模型 Att-LSTM 生成文本的影响，我们设置了三种改进的分开对比实验，即基于文本主题分布改进的 Att-LSTM-1 模

型、基于改进注意力评分的 **Att-LSTM-2** 模型和基于主题词覆盖生成的 **Att-LSTM-3** 模型。

3.3 实验参数设置

实验采用文本分词词汇对应的 300 维词向量作为每个时刻的模型输入，采用双层 LSTM 作为 RNN 基础结构，隐含层包含 600 个神经元，RNN 生成词汇个数限制在 200 以内。同时，batch 大小设为 32，学习率 0.0015，dropout 概率设为 0.4，迭代训练 100 次。每个样本包含 5 个主题词，coverage vector 采用基于 Twitter LDA 算法提取的主题词得分进行归一化得到。注意力评分模型中 $similarity(\cdot)$ 采用余弦相似度计算， $\beta=0.15$ 。采用训练好的模型进行文本预测生成时，为了减小文本长度对计算 BLEU 评估得分的影响，把生成的文本长度设为样本中文本的长度。

3.4 评估指标

我们采用人工评估和计算机自动评估两种评估方式：

人工评估：分别让 8 个研究生分别从主题相关性、主题完整性、主题词蕴含、句子通顺度、语句连贯性和信息量 6 个维度进行评分，最低分 0 分，最高分 5 分，然后计算每个样本的平均得分。其中，主题相关性用于衡量生成文本是否与所要求主题相关，主题完整性判断是否囊括每个主题信息，主题词蕴含用于判断生成文本中是否包含主题词，语句连贯性和句子通顺度分别用于判断前后两句是否连贯和单个句子是否通顺，另外，信息量用于衡量生成文本所表达的信息是否充分完整。

BLEU 评估：BLEU^[18]评估算法是文本自动生成领域普遍采用的一种评估方法，本文实现了 BLEU-4 得分并将 1-gram、2-gram、3-gram 和 4-gram 的权重分别设为[0.3,0.3,0.2,0.2]。

3.5 结果对比分析

表 2 人工评分结果表

Table 2 Manual evaluation results

模型	人工评分					
	主题相关性	主题完整性	主题词蕴含	句子通顺度	语句连贯性	信息量
Char-RNN	0.59	0.74	1.21	3.23	2.58	2.15
SC-LSTM	2.61	2.98	2.15	3.82	2.77	3.54
MTA-LSTM	3.26	3.54	2.67	3.96	3.35	3.32
Att-LSTM	2.63	2.48	2.37	3.43	2.86	3.65
Att-LSTM-1	3.22	3.87	3.45	4.06	3.17	3.48
Att-LSTM-2	3.47	2.96	3.17	3.69	2.99	3.86
Att-LSTM-3	3.19	3.27	4.38	3.45	3.31	3.42
Our Model	3.78	4.06	4.11	3.83	3.51	4.23

我们选取 ESSAY 数据集的测试结果进行了人工评估，结果如表 2 所示，表中每列分别表示主题相关性、主题完整性、主题词蕴含、句子通顺度、语句连贯性和信息量其中一个维度的得分，采用 8 个人工打分在所有样本上取平均值。从表中结果可以发现，由于模型 Char-RNN 没有对生成文本的主题进行约束，在主题相关性和主题完整性上效果比较差，我们的模型相对 MTA-LSTM 模型和 SC-LSTM 模型在这两个主题相关维度上都有明显提升。从主题相关性和主题完整性上的性能提升可以说明本文提出的文本主题分布和注意力评分方法可以取得成效，从主题词蕴含上的性能提升可以看出本文采用的主题覆盖生成方法确实能提高主题词在生成文本中的出现概率。在句子通顺度和语句连贯性上，本文所提出方法相对 MTA-LSTM 方法并没有明显提升，但是在信息量上提升较大，分析可能由于覆盖了更多的关键词和主题信息，导致信息量也随之增长。

分别对三个数据集计算了 BLEU 评分，结果如表 3 所示。可见，本文所采用模型在 ESSAY 数据集和 BaiKe 数据集上相对其他三个模型都有较大提升且取得最佳性能，在 ZhiHu 数据集上相对 MTA-LSTM 模型的提升不明显。分析可能因为 ZhiHu 数据集较小且主题词来源于知乎网站上的人工标注，部分主题词并没有出现在原文中，相应的词向量也存在缺失导致训练不准确。

通过对比 Att-LSTM 模型和三种独立改进策略，Att-LSTM-1 在主题相关性和主题完整性上的提升证明了主题分布策略的有效性；Att-LSTM-2 在主题相关性上提升较大，说明改进的注意力评分机制切实可行；Att-LSTM-3 模型在主题词蕴含指标上取得最高得分，同样证明了主题词覆盖策略的高效。三种改进模型在 BLEU 评分上

均优于 Att-LSTM 模型，进一步论证了三种改进策略的一致有效性。

表 3 BLEU 评分结果表

Table 3 BLEU evaluation results

模型	BLEU 评分		
	ESSAY	ZhiHu	BaiKe
Char-RNN	1.25	0.39	1.44
SC-LSTM	2.46	1.31	2.73
MTA-LSTM	3.29	1.66	2.96
Att-LSTM	2.37	1.34	2.55
Att-LSTM-1	3.35	1.56	3.14
Att-LSTM-2	3.17	1.63	2.91
Att-LSTM-3	3.55	1.83	3.23
Our Model	3.58	1.78	3.25

我们针对第一章中关于“家”的主题描述语句进行了篇章文本生成单例测试，测试结果如图 2 所示。

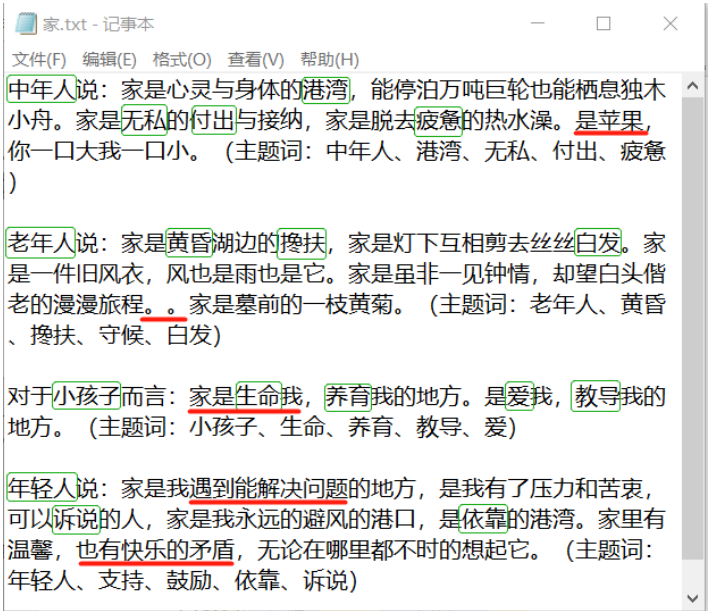


图 2 主题“家”的文本生成结果

Fig.2 Text generated based on topic “home”

从图中结果可见，针对主题“家”一共生成了 4 个段落，每段结尾括号中输出了该段相应的聚类主题词，我们用绿色方框在文中标记了主题词出现的位置。就生成的结果而言，主题词的覆盖率较高，也能够贴合所要表达的“家”这个主题思想。但是，依然存在一些问题。如我们用红色横线标记的地方，存在一些缺乏主语、语法结构错误和语句不通顺不连贯的情况。

4 结束语

本文针对基于主题思想的文本自动生成问题，提出了一种篇章级文本自动生成框架和基于主题约束的段落文本生成模型，模型对现有基于注意力机制的循环神经网络文本生成模型进行了改进和优化，在 ESSAY、ZhiHu 和 BaiKe 数据集上的实验结果证明本文提出模型在文本生成性能上相比基准模型有明显提升，且能够很好地贴合所要表达的主体思想。同时，依据我们提出的框架，每个主题聚类生成一个段落，可以生成主题多方面的文本信息，实现篇章级文本自动生成。该领域仍存在很多待解决的问题和挑战，例如难以控制生成文本的情感色彩、长文本的自动生成、主题段落的连贯性等，这也将是我们接下来的研究目标。

参考文献

- [1] Gatt A, Krahmer E. Survey of the state of the art in natural language generation: core tasks, applications and evaluation [J]. Journal of Artificial Intelligence Research, 2018, 61(1): 65-170.
- [2] Duma D, Klein E. Generating natural language from linked data: unsupervised template extraction [C]// Proc of the 10th International Conference on Computational Semantics. Stroudsburg, PA: ACL Press, 2013: 83-94.
- [3] Zhou Qingyu, Yang Nan, Wei Furu, et al. Selective encoding for abstractive sentence summarization [C]// Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL Press, 2017: 1095-1104.
- [4] Sutskever I, Martens J, Hinton G E, et al. Generating text with recurrent neural networks [C]// Proc of the 28th International Conference on Machine Learning. New York: ACM Press, 2011: 1017-1024.
- [5] Graves A. Generating sequences with recurrent neural networks [J/OL]. ArXiv: Neural and Evolutionary Computing, 2013. (2013-08-04)[2018-10-28]. <https://arxiv.org/pdf/1308.0850.pdf>.
- [6] Gu Jiatao, Lu Zhengdong, Li Hang, et al. Incorporating copying mechanism in sequence-to-sequence learning [C]// Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL Press, 2016: 1631-1640.
- [7] Logeswaran L, Lee H, Radev D R, et al. Sentence ordering and coherence modeling using recurrent neural networks [C]// Proc of the 32th AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2018: 5285-5292.
- [8] Kiddon C, Zettlemoyer L, Choi Y. Globally coherent text generation with neural checklist models [C]// Proc of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2016: 329-339.
- [9] Ghazvininejad M, Shi Xing, Choi Y, et al. Generating topical poetry. [C]// Proc of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2016: 1183-1191.
- [10] Wang Zhe, He Wei, Wu Haiyang, et al. Chinese poetry generation with planning based neural network [C]// Proc of the 26th International Conference on Computational Linguistics. New York: ACM Press, 2016: 1051-1060.
- [11] Qin Bing, Tang Duyu, Geng Xinwei, et al. A planning based framework for essay generation [J/OL]. ArXiv: Computation and Language, 2015. (2015-12-18)[2018-10-28]. <https://arxiv.org/pdf/1512.05919.pdf>.
- [12] Xing Chen, Wu Wei, Wu Yu, et al. Topic aware neural response generation [C]// Proc of the 31th AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2017: 3351-3357.
- [13] Feng Xiaocheng, Liu Ming, Liu Jiahao, et al. Topic-to-essay generation with neural networks [C]// Proc of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publisher, 2018: 4078-4084.
- [14] 姜力, 詹国华, 李志华. 基于递归神经网络的散文诗自动生成方法[J]. 计算机系统应用, 2018, 27(8): 259-264. (Jiang Li, Zhan Guohua, Li Zhihua. Automatic generation method of prose poem based on recurrent neural network [J]. Computer Systems & Applications, 2018, 27(8): 259-264.)
- [15] Bahdanau D, Cho K, Bengio Y, et al. Neural machine translation by jointly learning to align and translate [C]// Proc of the 6th International Conference on Learning Representations. San Diego, CA: ICLR, 2015: 1-15.
- [16] Wen T, Gasic M, Mrksic N, et al. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems [C]// Proc of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2015: 1711-1721.
- [17] Zeiler M D. ADADELTA: an adaptive learning rate method [J/OL]. ArXiv: Machine Learning, 2012. (2012-12-22)[2018-10-28]. <https://arxiv.org/pdf/1212.5701.pdf>.
- [18] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation [C]// Proc of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL Press, 2002: 311-318.
- [19] Li Shen, Zhao Zhe, Hu Renfen, et al. Analogical reasoning on chinese morphological and semantic relations [C]// Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL Press, 2018: 138-143.