北京大学学报(自然科学版) 第 56 卷 第 1 期 2020 年 1 月
Acta Scientiarum Naturalium Universitatis Pekinensis, Vol. 56, No. 1 (Jan. 2020)
doi: 10.13209/j.0479-8023.2019.103

基于主题约束的篇章级文本生成方法

黄炎[1,2] 孙海丽[1] 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; † 通信作者, 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这一问题, 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 3 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。在 3 个真实数据集上分别与 Char-RNN, SC-LSTM 和 MTA-LSTM 基准模型进行对比, 并对 3 个方面的改进进行独立验证。实验结果表明, 所提方法在人工评判和 BLEU 自动评测上均优于基准模型, 生成的文本能更好地贴合主题。

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

Discourse-Level Text Generation Method Based on Topical Constraint

HUANG Yan[1,2], SUN Haili[1], XU Ke[1,3], YU Xiaoyang[1], WANG Tongyang[1,†], ZHANG Xinfang[1], 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 form three aspects including topic distribution, attention scoring function and topic coverage generation. In experiments, the proposed method was compared with benchmark models such as Char-RNN, SC-LSTM and MTA-LSTM on three real datasets, three improvement strategies were verified and analysed independently. Experimental results show that proposed model is more efficient than benchmark models on human and BLEU metrics, and the generated text can catch the topic more 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的具有挑战性的研究任务, 旨在让计算机学会像人类一样写出高质量的自然语言文本, 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、从数据到文本生成以及从图像到文本生成[1]。

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。Sutskever 等[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。Graves[5]

深圳市科技计划基础研究项目(JCYJ20180306124612893, JCYJ20170818160208570, JCYJ20170307160458368)资助
收稿日期: 2019–05–22; 修回日期: 2019–09–23

将长短时记忆网络与RNN模型相结合，用于序列生成，学习长时依赖关系。为了模拟人们日常写作过程中的短语复用行为，Gu等[6]提出在序列到序列模型中引入copy机制。Logeswaran等[7]基于循环神经网络，针对语句连贯性建模，支持语句的重新排序，并生成语言连贯的文本内容。Kiddon等[8]提出用一种神经清单模型，来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。

基于深度学习生成的文本内容通常存在主题不明确以及语句不通顺等问题。通过给定的一个或多个主题词，显性地将主题词加入生成文本的适当位置[9]，根据主题词，挑选与主题相似的词语或语句，重新组合，生成新文本[10−11]，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[12−14]。这些方法可以在一定程度上缓解以上问题，但仍存在主题单一、主题分布不可控或主题词覆盖生成问题。

本文提出一种基于主题约束的篇章级文本自动生成模型。该模型首先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[15]，依据前文与每个主题的相似度调整注意力评分，用来平衡多个主题的影响。为了提高主题词在生成文本中出现的可能性，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自动评测方法进行验证，实验结果表明该方法具有可行性和有效性。

1 篇章级文本生成框架

本文提出的篇章级文本自动生成模型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。模型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，篇章级文本生成框架如图1所示。

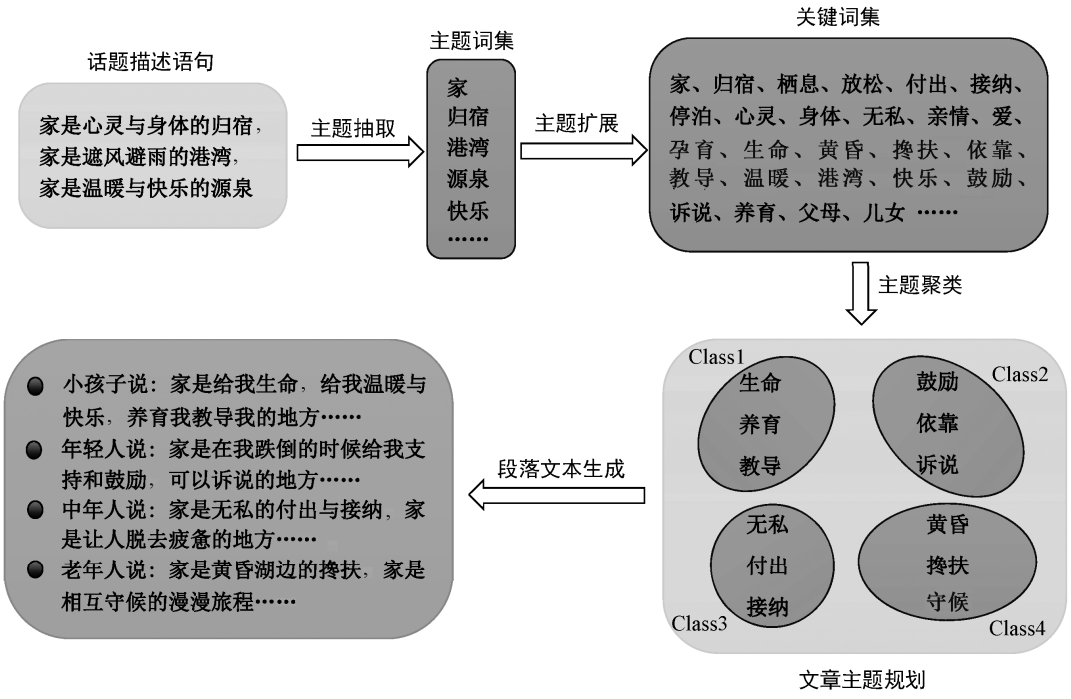

图 1 篇章级文本生成框架

Fig. 1 Discourse-level text generation framework

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”，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，根据文献[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主题词相近的词语，如“生命”、“养育”、“亲情”、“依靠”等，构成关键词集；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。与文献[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的方法不同，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。本例中的4个聚类分别生成4个对应的段落，每个段落阐述主题一个方面的内容。

2 基于主题约束的文本生成模型

2.1 任务定义及建模

对基于主题聚类约束生成段落文本任务的定义如下：给定一个包含 k 个关键词的主题聚类 $T=\{\text{topic}_1, \text{topic}_2, \ldots, \text{topic}_k\}$，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 $\text{Dict}=\{\text{word}_1, \text{word}_2, \ldots, \text{word}_m\}$，生成的段落文本内容 $\text{Text}=\{x_1, x_2, \ldots, x_l\}$，则任务可以形式化地描述为最大化条件概率 $P(\text{Text}|T)$ 的概率语言模型：

$$\text{maximize } P(x_1, x_2, \ldots, x_l|\text{topic}_1, \text{topic}_2, \ldots, \text{topic}_k),$$

subject to:

$$x_i \in \text{Dict},\ 1 \le i \le l;\ t_i \in \text{Dict},\ 1 \le i \le k;\ k \le l \le m。\quad (1)$$

2.2 文本生成模型

本文采用基于循环神经网络模型，模拟人类大脑注意力行为，对主题聚类信息添加注意力机制来约束文本的生成，并在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3个方面提出相应的改进策略。

2.2.1 基于注意力机制的 RNN 模型

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 $\text{topic}_i (1 \le i \le k)$ 和段落中每个词汇 $x_j (1 \le j \le l)$ 的向量表示。将段落词汇按序列输入RNN模型，假设 t 时刻的输入为 x_t，令 t 时刻的输出等于下一时刻的输入，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。

条件概率语言模型(式(1))的计算方式如下：

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

下一个词汇的出现概率定义如下：

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

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

我们采用基于主题的注意力机制来约束LSTM神经网络生成的文本，即通过计算主题词向量与生成词汇之间的相似度来评分并排序，挑选评分较高的词汇作为LSTM输出，使生成的文本内容更偏向我们选择的主题。对每个主题词 $\text{topic}_1 \sim \text{topic}_k$，通过Attention[13]计算转化为相应的注意力得分 $\alpha_1, \alpha_2, \ldots, \alpha_k$，则 t 时刻的主题向量表示 T_t 为

$$T_t = \sum_{j=1}^{k} \alpha_{tj} \text{topic}_j,$$

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。

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

$\eta(\cdot)$ 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

为了使模型生成的文本内容能够覆盖所有主题，引入 k 维主题覆盖向量 $C_t = [C_{t,0}, C_{t,1}, \ldots, C_{t,k}]$[13]，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，Feng等[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, …,1.0]，没有考虑文本对不同主题的概率分布情况。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:

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

其中，score (topic$_j$)为通过 Twitter LDA 算法从原始文本中抽取主题 topic$_j$ 计算得到的得分，该得分表示主题在文本中的概率分布。

通过 C_t 调整注意力评分来确保主题覆盖率:

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

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:

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

$$\phi_j=N\cdot\sigma(\boldsymbol{U}_f[T_1,T_2,\ldots,T_k]),\ \boldsymbol{U}_f\in\mathbb{R}^{kd_w},$$

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，$\boldsymbol{U}_f$ 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

现有的基于注意力机制[13−15]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}(式(2)和(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，计算公式如下:

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

其中，similarity(·)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。所以，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

在基于主题约束的文本自动生成任务中，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。我们参照文献[12]，为每个词汇 w_i 的生成概率添加一个附加项，以提高主题词的生成可能性:

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

其中，

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

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

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

$$Z=\sum\nolimits_{w\in V}\mathrm{e}^{g_V(h_t)}+\sum\nolimits_{w\in K}\mathrm{e}^{g_K(h_t)},$$

$g_V(h_t)$ 和 $g_K(h_t)$ 分别是两个不同参数的全连接层，V 指响应词汇表 Dict, K 指主题词表 T。

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出现的概率。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:

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

其中

$$x_i\in\text{Dict},\ 1\le i\le l;\ t_i\in\text{Dict},\ 1\le i\le k;\ k\le l\le m。$$

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 和 MTA-LSTM 模型进行对比。为了独立地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 和 ZhiHu 数据集[13]进行模型的训练和验证测试。ESSAY 数据集通过爬取和整

理中国高考优秀作文得到，ZhiHu数据集通过爬取知乎网站的文章和相应主题词得到。通过爬取百度百科网站的自然科学语料构建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3种语料，分别去除包含数字、英文的样本，过滤掉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50~200之间。3种数据集的统计结果见表1。

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，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，然后计算每个主题词的概率分布。针对ESSAY数据集训练word2vec词向量，每个词向量的维度设为300。由于ZhiHu和BaiKe语料已有很多预训练词向量集合，因此，我们直接使用Li等[18]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，我们将双层LSTM作为RNN模型的基础结构。

SC-LSTM[16] 基于语义控制的LSTM结构进行统计语言生成，引入对话行为的one-hot主题向量覆盖机制，使生成的文本包含特定的主题信息。

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

为了独立地验证文本主题分布、注意力评分方法和主题覆盖生成3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3种改进的对比实验，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

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的基础结构，隐含层包含600个神经元，RNN生成词汇个数限制在200以内。batch大小设为32，学习率为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通过对基于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(·)采用余弦相似度计算，β=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与样本中文本长度一致。

3.4 评估指标

采用人工评估和计算机自动评估两种评估方式。

1) 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0分，最高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，主题完整性用于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，信息量用于衡量生成文本所表达的信息是否充分完整。

2) BLEU评估：BLEU评估算法[19]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和0.2。

3.5 结果对比分析

我们选取ESSAY数据集的测试结果进行人工评估，结果如表2所示。表中的得分是采用8个人工打分在所有样本上的平均值。

从表2可以发现，由于模型Char-RNN没有对生成文本的主题进行约束，所以在主题相关性和主题完整性上的效果比较差。与MTA-LSTM和SC-LSTM相比，本文模型在这两个主题相关维度上都

表 2 人工评分结果

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
本文模型	**3.78**	**4.06**	4.11	3.83	**3.51**	**4.23**

有明显的提升, 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, 在主题词蕴含上的性能提升说明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, 本文方法相对于 MTA-LSTM 并没有明显的提升, 但是在信息量上提升幅度较大, 可能是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。

对 3 个数据集分别计算 BLEU 评分, 结果如表 3 所示。可以看出, 本文采用的模型在 ESSAY 和 BaiKe 数据集上比其他 3 个模型都有较大程度的提升, 且取得最佳性能。在 ZhiHu 数据集上相对于 MTA-LSTM 模型的提升不明显, 原因可能是 ZhiHu 数据集较小, 且主题词来源于知乎网站的人工标注, 部分主题词没有出现在原文中, 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 3 种独立改进策略, Att-LSTM-1 在主题相关性和主题完整性上有提升,

表 3 BLEU 评分结果

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
本文模型	**3.58**	1.78	**3.25**

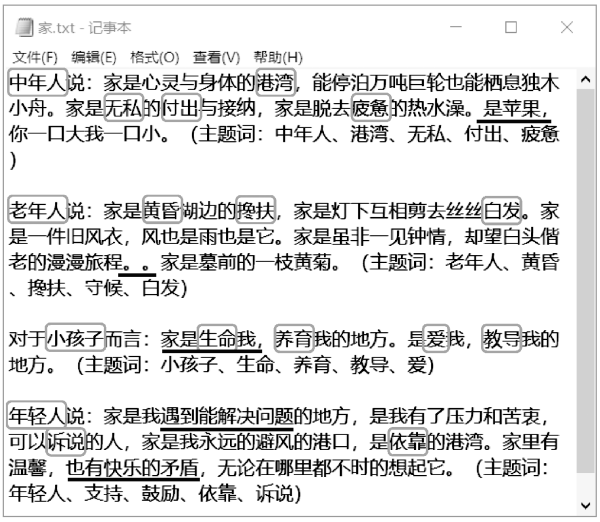

图 2 主题“家”的文本生成结果

Fig. 2 Text generated based on topic “home”

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幅度较大, 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, 证明了主题词覆盖策略的高效性。3 种改进模型在 BLEU 评分上均优于 Att-LSTM 模型, 进一步验证了 3 种改进策略的一致有效性。

针对第 1 节中关于“家”的主题描述语句, 进行篇章文本生成单例测试, 结果如图 2 所示。可以看出, 针对主题“家”共生成 4 个段落, 每段结尾括号中输出该段相应的聚类主题词, 用灰色方框标记主题词出现的位置。就生成的结果而言, 主题词的覆盖率较高, 也能够贴合“家”的主题思想。但是, 依然存在一些问题, 例如黑色横线标记的地方, 存在缺乏主语、语法结构错误以及语句不通顺、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, 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, 本文提出的模型在文本生成性能上相比基准模型有明显的提升, 且能够很好地贴合所要表达的主题思想。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, 实现篇章级文本的自动生成。该领域仍存在很多待解

决的问题，例如难以控制生成文本的情感色彩，长文本的自动生成，以及主题段落的连贯性，等等，这也是我们未来的研究内容。

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. Journal of Artificial Intelligence Research, 2018, 61(1): 65–170

[2] Duma D, Klein E. Generating natural language from linked data: unsupervised template extraction // Proc of the 10th International Conference on Computational Semantics. Stroudsburg, 2013: 83–94

[3] Zhou Qingyu, Yang Nan, Wei Furu, et al. Selective encoding for abstractive sentence summarization // Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, 2017: 1095–1104

[4] Sutskever I, Martens J, Hinton G E, et al. Generating text with recurrent neural networks // Proc of the 28th International Conference on Machine Learning. New York, 2011: 1017–1024

[5] Graves A. Generating sequences with recurrent neural networks [EB/OL]. (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 // Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, 2016: 1631–1640

[7] Logeswaran L, Lee H, Radev D R, et al. Sentence ordering and coherence modeling using recurrent neural networks // Proc of the 32th AAAI Conference on Artificial Intelligence. Menlo Park, 2018: 5285–5292

[8] Kiddon C, Zettlemoyer L, Choi Y. Globally coherent text generation with neural checklist models // Proc of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, 2016: 329–339

[9] Ghazvininejad M, Shi Xing, Choi Y, et al. Generating topical poetry // Proc of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, 2016: 1183–1191

[10] Wang Zhe, He Wei, Wu Haiyang, et al. Chinese poetry generation with planning based neural network // Proc of the 26th International Conference on Computational Linguistics. New York, 2016: 1051–1060

[11] Qin Bing, Tang Duyu, Geng Xinwei, et al. A planning based framework for essay generation [EB/OL]. (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 // Proc of the 31th AAAI Conference on Artificial Intelligence. Menlo Park, 2017: 3351–3357

[13] Feng Xiaocheng, Liu Ming, Liu Jiahao, et al. Topic-to-essay generation with neural networks // Proc of the 27th International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence. San Francisco, 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. 计算机系统应用，2018，27(8): 259–264

[15] Bahdanau D, Cho K, Bengio Y, et al. Neural machine translation by jointly learning to align and translate // Proc of the 6th International Conference on Learning Representations. San Diego, 2015: 1–15

[16] Wen T, Gasic M, Mrksic N, et al. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems // Proc of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, 2015: 1711–1721

[17] Zeiler M D. ADADELTA: an adaptive learning rate method [EB/OL]. (2012–12–22)[2018–10–28]. https://arxiv. org/pdf/1212.5701.pdf

[18] Li Shen, Zhao Zhe, Hu Renfen, et al. Analogical reasoning on Chinese morphological and semantic relations // Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, 2018: 138–143

[19] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation // Proc of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, 2002: 311–318