Text Generation Algorithm based on Variant Probability Graph Model

Article *in* Journal of Computer Applications · June 2018

CITATION
1

READS
30

2 authors, including:

Tingzhen Liu
Tencent
20 PUBLICATIONS **10** CITATIONS

SEE PROFILE

文章编号:1001-9081(2018)S1-0099-05

基于变种概率图模型的文本生成算法

刘廷镇[1*],张 华[2]

(1.渤海大学附属高级中学,辽宁 锦州 121000; 2.渤海大学 海事学院,辽宁 锦州 121013)

(* 通信作者电子邮箱 firstsg@outlook.com)

摘 要:针对文本自动生成问题中通过抽象摘要生成文本的需求特点,提出了一种多层概率图模型用于文本生成。通过多层次摘要等方法使得其能从更高层次理解语义需求,逐层传播信度进行文本生成,并通过线性规划进行不同层间连接权值的训练。对其进行了系统实现后的实验结果表明,所建立的计算模型实现了根据关键词生成文章的目标,效果上显著优于基于seq2seq的文本生成模型,可以用于"概念-文本"生成的实际应用场景。最后,对该模型进一步改进和与现有其他机器学习模型融合的方法进行了展望。

关键词:语义结构;语言模型;自然语言生成;机器学习;概率图模型

中图分类号:TP181 **文献标志码**:A

Text generation algorithm based on variant probability graph model

LIU Tingzhen[1*], ZHANG Hua[2]

(1. *Senior School attached to Bohai University*, *Jinzhou Liaoning* 121000, *China*;
2. *Maritime College*, *Bohai University*, *Jinzhou Liaoning* 121013, *China*)

Abstract: Aiming at the demand characteristics of text generated by abstract abstract in automatic text generation problem, a multi-layer probability graph model for text generation was presented. By means of multi-level summarization, the semantic requirement could be understood from a higher level, and the text was generated by propagatiing trust degree layer by layer. The weights of different layers were trained by the linear programming. The experimental results of the system show that the computational model achieves the goal of generating the article according to the key words, which is significantly better than that of the text generation model based on seq2seq. So the model can be used for the actual scenario of concept-text generation. Finally, the further improvement of the model and the fusion with other machine learning models were prospected.

Key words: semantic structure; language model; Natural Language Generation (NLG); machine learning; probabilistic graph model

0 引言

自然语言生成(Natural Language Generation, NLG)技术是计算语言学研究的重点之一,其复杂性在于不同词语间丰富的语义信息使得表达充满了歧义,必须结合一定的上下文才能正确进行表层生成。另一方面,自然语言往往具有复杂的层次结构,语言粒度包含文档、段落、句子和词等多个不同的组成部分。在表示过程中同样需要考虑这些不同粒度的语法单元,为其选择合适的数据结构。传统的NLG方法中直接使用模板组织输入的信息,模板又分为很多种,如修辞结构理论(Rhetorical Structure Theory, RST)生成技术中的修辞关系集、Lexicalization过程中的间隔描述表[1],这样生成文本的效果就高度依赖于模板的好坏。虽然目前使用关系自举方法进行自然语言模板自动抽取的技术[2]已经广为应用,但所抽取的模板都较为简单,无法直接用于上下文具有丰富语义信息的NLG工作。抽取和处理复杂的结构化模板更是需要了解相关的语义学的知识,对于相关人员在该领域能力的要求较高。目前,基于对神经网络的火热研究,使用循环神经网络(Recurrent Neural Network,RNN)的NLG方法也进入了人们的视野[3]。RNN摒弃了传统方法的过度形式化,反而善于刻画文本中的顺序关系。但是对于端到端(end to end)的生成工作来说,虽然能够生成通顺的句子,由于其本质上仅仅编码序列单位间的概率,难以用来创作流畅的文章。同时在大训练集上巨大的计算开销和反复编码造成的低可解释性也是RNN用于低垂NLG领域的障碍所在。

自然语言生成可以说是文本摘要的逆过程,即从抽象的概念层次开始,通过执行一定的语义语法规则来生成文本[4]。如在基于概率潜在语义分析(Probabilistic Latent Semantic Analysis, PLSA)的文本摘要模型中,先通过猜测文本产生的方式,考虑文章是通过一个主题-词语的生成模型生成,再根据文章内容、以最大化似然函数为目的反过来求解被视为主题的隐变量,以实现文本摘要[5]。从这个角度考虑,本文算法将训练文本分割成不同感知窗口进行摘要生成摘要块,再将摘要块视为随机变量,同大小感知窗口得到的摘要块按邻接关系相互连接描述顺序关系,上下级感知得到的摘要块按包含关系相互连接形成层级,边的权值为进行该种传递的条件概率。通过这样描述语义单位间的概率依赖建立概率图模型(Probabilistic Graphical Model, PGM)。

传统的贝叶斯网络是以随机变量为节点的有向无环图。但通过分析RNN模型可知,RNN通过刻画序列单位间的顺序关系以模仿样本进行生成。基于这种思想,对于单层结构,将每个摘要块与其在文本中前后相连的摘要块连一条边以刻

收稿日期:2017-11-21;**修回日期**:2017-12-21。

作者简介:刘廷镇(1999—),男,辽宁锦州人,主要研究方向:自然语言处理; 张华(1978—),女,辽宁锦州人,讲师,硕士,主要研究方向:机器学习。

画序列关系。但这样结构中就会出现环，整体的训练和推断方式也发生改变，成为变种的概率图模型。针对此结构的特点，本文特别提出使用线性规划训练模型的权值。最终，当赋予数个摘要块在新文本中出现的先验信度时，这个信度值会沿着邻接边传递下去，形成一个宏观规划序列。该序列又会通过上下级边向下传递，以生成微观子序列。通过贝叶斯链式法则可以有效评估最终表层序列产生的概率并进行优选。由于 PGM 可以有效描述随机变量间的相互依赖[6]，使得其能够很好地刻画文本中不同层次间语义结构的生成关系。

该方法起到的作用类似于现今对话系统检索使用的融合模型——对问答句子中的关键字进行摘要，并用知识图谱中周围知识对正确的答句主题方向进行修正和增强[7]。实验结果表明该模型具有良好的可解释性和极高的运行效率。由于权值具有稀疏性，故能在高层次原生描述文本语义[8]，又没有陷入过度形式化导致训练信息难以获取的境地，不失为未来文本生成的一种有效方法。

为了对该模型进行分析、总结和比较，使用一些剧本和小说文本训练了模型，对模型中关键数据进行统计，并自定义一些关键字进行了激活生成，获得的新文本在宏观层面和表层上都较为流畅，表明该模型有能力生成接近于人类手写的文本。

1 句子摘要与相似度计算

最后算法运行的基础是对训练文本使用不同大小的滑动窗口进行不同层级的摘要，对于每个窗口进行摘要得到的摘要词集合称为摘要块 B。

定义 1 摘要块。摘要块为本层的概率图节点，其被如下五元组描述：

$$B = \{L_1, L_2, S, P, C\}$$

其中：L_1 为在 B 之前、与 B 相邻出现的所有摘要块集合；L_2 为在 B 之后、与 B 相邻出现的所有摘要块集合。S 为代表生成 B 的滑动窗口中所有下级摘要块的中间节点集合。如产生 B 的滑动窗口为句级，那么 S 即为被 B 所摘要的句子中所有单词的中间摘要节点。P 为 B 在其本级滑动窗口中出现在首端的频率。如产生 B 的滑动窗口为单词级，那么 P 即为 B 单词出现在句首的频率。C 为该摘要块在训练集中出现的次数。

如果产生 B 的滑动窗口为单词级，那么可以直接通过对该词语频数的统计计算 P 和 C。但如果产生 B 的滑动窗口为句级，由于在一篇文章中表达实义的句子基本不会完全相同地出现两次，这样 P 和 C 都一直为 1，统计量就失去了意义。所以需要一种方法计算两个同级窗口间的相似度，然后把相似度较小的窗口的摘要块视作同一个。对于这类问题，由于窗口大小不固定，一些专门为词、句等语法结构设计的相似度计算算法就不适用了。而我们仅需要一种完全基于统计量的方法描述两个窗口间的相似性，因此这里使用方宁[9] 提出的同粒度文本语境词汇关联计算方法的改进版作为相似度计算算法：

假设该级滑动窗口产生 m 个摘要块，记为 $B = B_1, B_2, \cdots, B_m$，上级滑动窗口对 B 区域进行摘要产生 n 个摘要块，记为 $\beta = \beta_1, \beta_2, \cdots, \beta_n$。因为所有不同层级的摘要都可以用一个摘要词描述，那么对于两个该级滑动窗口计算相似度，可以通过计算上级摘要 β 中摘要词的交并数进行。

对于 B_1 和 B_2 两个区域的相似度计算方法：

1) 计算 B 上级滑动窗口产生的摘要块 β_{B1} 和 β_{B2} 间相同的关键词个数，记为 $\beta_{B1} \cap \beta_{B2}$；

2) 计算 β_{B1} 和 β_{B2} 间所有的关键词个数，记为 $\beta_{B1} \cup \beta_{B2}$；

3) 如果 $\beta_{B1} \cap \beta_{B2} / \beta_{B1} \cup \beta_{B2} > \xi$（$\xi$ 为超参数）那么两摘要块间相似。

对于摘要算法，由于对除单词层外的其他层级进行摘要时，摘要的对象已经并非连贯的文本，而是一些摘要词的集合。故本模型同样使用完全基于统计量的 TextRank 算法[10] 进行摘要。摘要并不需要十分准确，一些并不能代表下层特征的词被摘要出来也是允许的。因为摘要的目的仅仅在于筛选出可能与下层有连接的单元描述并建立新的上层节点，而具体连接权值的大小还需要通过进一步训练。

2 变种概率图的构造原理

2.1 单层结构

以每个摘要块作为节点，与前后相连的摘要块连一条有向边，代表其概率的相关性。摘要块 B_1 与摘要块 B_2 间边的权值为在摘要块 B_1 后出现摘要块 B_2 的概率值。如第一层，滑动窗口为单词级，那么采样得到的摘要块即是文本中的所有单词，以每个单词为一个节点，与其前后相连的单词以一条边连接，建立图结构。由于是第一层，不存在下级摘要块 S，所以本层的摘要块实际为四元组。这样，模型的数据结构被初步建立起来。

2.2 多层化

仅仅这样做，同样会遇到 RNN 在端到端生成工作中遇到的问题——仅能生成通顺的句子，而难以对文章内容有整体的规划[3]。基于神经网络的记忆能力和编解码操作，RNN 可以记录很长范围中的顺序关系并进行进一步抽象[11]，而单层结构的序列图中每个词仅与前后的词连接，无法满足实际任务需求。为了解决这个问题，使用不同大小的滑动窗口分别对文章进行多级采样，并构造多层的概率图。

如基于先前构造的单词级概率图，放大滑动窗口的范围，每次对一个句子进行采样，这样得到了该句子的多个摘要词，对每个摘要词构造中间节点，每个中间节点与被采样句子中所包含的单词对应的单词级节点全连接，代表这些单词出现的概率与该中间节点相关，每条边的权值反映了该摘要词对其下层每个节点的语义传递作用。这些摘要词构成的节点的集合即为 S。S 为高层摘要块 B 的核心部分，与第一层相同，摘要块依然与本层前后相连的其他摘要块连一条边。

直观来讲，高层摘要块是对语义的一种抽象描述，当需要生成文本时，仅需要赋予与所需主题相关的摘要块或中间节点以一个先验信度，模型就会根据训练后的权值自动向下层传递、计算概率，最终将信息传递到作为单词级描述的第一层进行表层生成。

2.3 短接边与同义词边

至此，模型已经能从较高层面表示待生成文本的语义。但对于第一层，依然有一些语言上的基本问题需要解决，每个单词前后相连的边代表着该单词从涵义上讲下面将会出现某个单词。而停用词前后连接的词则与涵义完全无关，如短语“汤姆的战争”，如果模型以“汤姆—的—战争”形式建立边，那么“汤姆”与“战争”间的语义联系就被“的”这个节点割裂开了，而“的”前后连接什么单词并非是由于其语义，而仅仅是一种语法现象。我们想表达的涵义应当是“汤姆”和“战争”相连接时中间常有“的”的出现。因此对于停用词，使用一种短接的手法，对描述边的结构 L 加入元素——停用词集合。

定义 2 停用词集合。

停用词集合为 $l = \{l_1, l_2, \cdots\}$，其中：$l_n = \{W, P\}$，W 为

该边连接的两个节点中出现的停用词,P 为该停用词的出现概率。在生成过程中,在通过该条边激发的两个词之间插入概率较大的停用词 W_n 即可。

同时,由于同义词前后出现的词也很可能相同,在所有同义词间连接一条边使得它们能共享权值,加强生成的效果。

3　通过信度传播生成文本

完成了模型数据结构的创建工作后,还需要给出如何使其生成文本的算法。假设直接赋予一个现有的高层摘要块一个先验信度,那么它会先进行本层的传导以计算所有摘要块在所需文本中出现的概率。

算法 1　本层信度传导。设初始仅激活一个块 B_j,赋予先验信度为 α,B_i 当前求得概率值为 α_i。

1. 如果 $\alpha \geqslant \tau$ (τ 为偏置值,即可传递的最低的概率值,超参数)
2. $\quad \alpha_j \leftarrow \alpha_j + \alpha$
3. $\quad$ 对 L_{1_j} 连接的所有节点 B_i 执行以下操作
4. $\quad\quad$ 如果概率值没有通过该条边进行过传递
5. $\quad\quad\quad \alpha \leftarrow \alpha \times P(B_j \mid B_i)$, $j \leftarrow i$, 递归进行本算法
6. $\quad$ 对 L_{2_j} 连接的所有节点 B_i 执行以下操作
7. $\quad\quad$ 如果概率值没有通过该条边进行过传递
8. $\quad\quad\quad \alpha \leftarrow \alpha \times P(B_i \mid B_j)$, $j \leftarrow i$, 递归进行本算法

分析上述算法,同一条边被限定不能传递两次,这使得概率值不会在两个节点之间反复传播。边的权值为 $P(B_j \mid B_i)$, $P(B_i \mid B_j) < 1$ 使得传播的信度值不断减小,最终在一定范围内小于偏置值,传播停止。

本层节点的概率计算结束后,将会计算所有可能出现的摘要块序列的概率。

算法 2　生成摘要块序列并计算联合概率,设序列出现的概率为 p。

1. 遍历本层所有摘要块,对每个摘要块 B_j 执行以下操作
2. $\quad$ 如果 $\alpha_j \geqslant \upsilon$ 且 $P_j > 0$(υ 为节点选入摘要块序列的最低概率值,超参数)
3. $\quad\quad$ 新建摘要块序列, $p \leftarrow P_j \times f_2(a_j)$
4. $\quad\quad$ 标签 递归起点。
5. $\quad\quad\quad$ 将 B_n 加入其中摘要块序列
6. $\quad\quad\quad$ 对 L_{2_j} 连接的所有节点 B_i 执行以下操作
7. $\quad\quad\quad\quad$ 如果 $\alpha_i \geqslant \upsilon$
8. $\quad\quad\quad\quad\quad p \leftarrow p \times f_1(C_j, n_{B_j \rightarrow B_i}) \times f_2(\alpha_i)$ ($n_{B_j \rightarrow B_i}$ 为 B_j 到 B_i 的边的出现次数)
9. $\quad\quad\quad\quad\quad j \leftarrow i$
10. $\quad\quad\quad\quad\quad$ 到递归起点递归进行本算法
11. $\quad\quad$ 遍历完成后,通过类模拟退火公式 $e^{-\Delta/T}$ 近似地选择 p 最大的序列(Δ 为活跃度参数,越大越倾向于选择概率小的序列; T 为训练集大小)

分析上述算法不难看出,算法使用深度优先搜索的方式扩展序列中元素。在实际工程中,为防止摘要块序列扩展范围过大使得内存不足,还加入了其他的扩展条件进行剪枝,在此不作展开。此外,概率更新则使用了 f_1 和 f_2 两个函数。因为如果直接使用 $p \leftarrow p \times P(B_i \mid B_j)$,那么越长的句子 p 就越小,这显然与实际情况不符。函数 f_1 和 f_2 定义如下:

$$f_1(C, n) = \rho n / C$$

$$f_2(\alpha) = \alpha \Big/ \left(\frac{1}{m} \sum_{i=1}^{m} \alpha_i \right)$$

f_2 的形式很容易理解,其为该环境下条件概率公式的近似形式,代表着概率值大的节点在摘要块序列生成过程中具有更大的被选中概率。

f_1 与 f_2 类似,即如果 n/C(摘要块 B_j 后连接 B_i 的频率) 大于 $1/\rho$, $p > p \times f_2(\alpha)$。也就是说,出现频率更大的摘要块组合在摘要块序列生成过程中具有更大的被选中概率。经实验, $\rho = 2$ 是一个合适的取值。

高层的摘要块序列是关于文章叙述顺序的一种高层次规划,有了高层次规划,需要将其向下层传导以进行进一步的规划,一直到单词层生成摘要块序列(即表层生成) 为止。下面给出向下层传导概率值的算法:

算法 3　向下层传导概率值,设中间节点 S_j 连接到下层摘要块 b_j 的权值为 $\omega_{j,i}$,下层摘要块 b_j 当前求得的概率值为 α_j。

1. 遍历摘要块序列,对每个摘要块 B_j 执行以下操作
2. $\quad$ 遍历摘要块 B_j 所包含的中间节点 S,对每个中间节点 S_j 执行以下操作
3. $\quad\quad$ 遍历 S_j 连接到的下层摘要块集合,对每个下层摘要块 b_i 执行以下操作
4. $\quad\quad\quad$ 激发 b_i, 即 $\alpha_j \leftarrow \omega_{j,i}$

需要注意的是,摘要块序列中不同摘要块向下传递概率的过程相互独立,即摘要块 B_1 向下传递,得到最终结果(B_1 所对应的单词序列) 后将下层节点概率值全部归零,再进行 B_2 的传递。因为生成文本段落的序列关系仅通过高层摘要块序列反映,低层摘要块序列只负责其所对应的微观规划和表层生成,故不需要前一次的传递结果来影响后一次的生成。

4　层间传递权值训练

完成具体如何生成文本的算法后,剩下的工作就是权值的训练。其中,同层边的概率已经通过频率赋予,而没有进行调整的是摘要块向下级节点连接边的权值。对于概率图模型的参数估计问题,通常使用极大似然估计求解,基于此思想,训练过程应最大化训练集中序列的联合概率。从句级-单词级的角度、结合本模型的环境特点考虑,如果一个句级摘要块向下传导、单词级节点同层激发后,句级摘要块所包含的单词节点的概率值 $\alpha \geqslant \upsilon$,不被包含的单词节点 $\forall b_j \in B$, $h_j(\omega_{1,1}, \cdots, \omega_{m,k}) \geqslant \upsilon$,必然可以生成与训练集中文本完全相同的序列,即该条件为最大化训练集中序列联合概率的充分条件。那么该问题可视为:

对句级摘要块 B 的层间传递权值矩阵 $\boldsymbol{\omega}$,使得 $\forall b_j \in B$, $h_j(\omega_{1,1}, \cdots, \omega_{m,k}) \geqslant \upsilon$。

考虑到计算量,对于后一个条件,仅调整未训练激活结果中 $h_j \notin B \wedge h_j(\omega_{1,1}, \cdots, \omega_{m,k}) \geqslant \upsilon$ 的项目,使得其不等号方向调换(h_j 为在未训练 $\boldsymbol{\omega}$ 情况下同层传导得到 α_j 的计算过程)。又从算法 1 可以看出, h_j 的函数值仅为 $\omega_{1,1}, \cdots, \omega_{m,k}$ 的线性组合,那么整个问题转化为 $m \times k$ 个目标变量的线性规划问题。

由于不能保证在所有摘要块的约束条件都加入后,不等式组依然有解,所以要对约束条件进行改写以扩大可行域。由于 h_j 中 $\omega_{1,1}, \cdots, \omega_{m,k}$ 的系数必定为正数,故引入一组变量 s 进行如下改写:

对约束条件 $h_j(\omega_{1,1}, \cdots, \omega_{m,k}) \geqslant \upsilon$,改写为件 $h_j(\omega_{1,1}, \cdots, \omega_{m,k}) \geqslant \upsilon - s_j$;

对约束条件件 $h_j(\omega_{1,1}, \cdots, \omega_{m,k}) \leqslant \upsilon$,改写为 $h_j(\omega_{1,1}, \cdots, \omega_{m,k}) \leqslant \upsilon + s_j$。

为使对可行域的改变最小,设置目标函数为 $\sum_j S_j$,求最小值,其中 $s_j \in [0, +\infty)$。由于可以肯定所有中间节点对其

全连接到的下层节点都有语义连接，故 $\omega_{i,j} \in [\upsilon, +\infty)$，该规划问题必有最优解。

这样，从底层到上层，逐层使用该方法训练权值，即可得到所有层间传递权值。

5 实验结果分析

模型是否可以用作“概念-文本”的实际生成工作需要通过两方面评价：一是生成的文本是否贴合摘要块描述的精确度评价，由于该评价标准即为线性规划训练的目标，故评价的主要意图是验证训练是否达到效果；二是生成的文本从自然语言角度上讲是否通顺的流畅度评价，意图是评价模型生成的文本是否具有实际应用价值。

使用小说《炼狱之鸦》(局部)分别对变种概率图模型和基于 LSTM 的单词级 seq2seq 模型使用 i7 7700 CPU 进行训练，两模型的基本属性如表 1。

5.1 通过精确度指标评价训练有效性

由于“概念-文本”的文本生成问题并不同于机器翻译、图像标注等问题有着明确的 Gold-sequence 可以作为评价依据，为了有效进行评价，需要引入一个基本假设：使用与训练文本中某单句的摘要对应的摘要块时，模型应当生成与该单句相似的句子。

依据该假设，可以根据类似 BLEU 指标[12]的计算方法，定义该问题下的召回率作为精确度的评价指标：

定义 3 召回率。

使用与训练文本中某单句摘要对应的摘要块激活模型，设训练文本中单句的词 2 元组数为 N_1，生成单句中词 2 元组数为 N_2，二者共现的 2 元组数为 C。召回率计算公式为：

$$R = \frac{C}{N_1} \times 1; N_2 < N_1$$

$$R = \frac{C}{N_1} \times e^{\frac{N_1}{N_2}-1}; N_2 \geqslant N_1$$

其中，第一项 C/N_1 衡量生成单句和样例重合的 2 元组数在样例中所占比例。但如果生成单句过长，会影响这种判断的准确性，因此引入第二项作为惩罚因子，即当 $N_2 \geqslant N_1$ 时，根据 N_2 与 N_1 的差距对第一项计算值予以折扣。R 的取值范围为 $(0,1]$：当 R 为 1 时，生成文本和训练文本完全相同。R 越小说明二者相似度越低。

选取训练集中 130 个句子进行对比，召回率平均值为 0.883，部分句子对比结果如表 2。可以看出，随机选取的 5 个测试单句中：3 个生成单句都与训练文本完全相同；另外 2 个中的 1 个是因使用了同义词影响了召回率，含义上并无区别，仅有一个出现偏差。该结果说明了模型训练方法的有效性。

表 1 两模型配置与运行的基本属性对比

模型	层数	输入数据	并行模式	迭代次数	运行时间/s
变种概率图文本生成模型	两个摘要层	激活代表元素为“找到”“希望”“寒冷”“公园”“刀锋”等的二级摘要块	单线程	(完成线性规划)	102.7
seq2seq 模型	LSTM 层 + dropout 层 + dropout 层 + LSTM 层 + dropout 层 + 全连接层	以“希望”为首词向下预测	CPU 并行	700	1 305.3

表 2 不同激活数据的召回率

激活的摘要块	训练文本中单句	生成的单句	召回率
观察、周围、空无一人	最后再观察一下周围的动静，空无一人。	最后再观察一下周围的动静，空无一人。	1.000
井之头、公园、万籁俱寂	不久就来到了井之头公园。夜幕下的公园里，万籁俱寂。	欢迎来到了井之头公园，里万籁俱寂。	0.286
手腕、刀、向下	是那种每当我用力把架在手腕上的刀向下按时，就会感到的一阵酥麻痛感。	是那种每当我用力把架在手腕上的刀向下按时就会感到一阵麻痹。	0.761
无法、倾诉 、思绪	甚至无法，对谁倾诉此刻的万千思绪。	甚至无法，对谁倾诉此刻的万千思绪。	1.000
重生、罪	想要重生的念头，是一种罪吗?	想要重生的念头，是一种罪吗?	1.000

5.2 与 seq2seq 模型的流畅度对比

为了分析模型生成的文本的情节连贯性和表层流畅性，使用如表 3 的数据分别输入两对比模型。

表 3 对两模型的输入数据

模型	输入数据
变种概率图文本生成模型	激活代表元素为“找到”“希望”“寒冷”“公园”“刀锋”等的二级摘要块
seq2seq 模型	以“希望”为首词向下预测

用两模型分别生成如下文本，均进行了经过了人工分段和去除个别循环重复词句。

1) seq2seq:

希望夜幕凝神，毕竟试练之余痛感越发。

默默抹了抹桐木箱。利用有手腕溢出光芒。母亲维持让七盏死骨清理销声匿迹。

如今母亲搬起维持大都市三宝台与将角落。丝毫选对念头，明明刀刃各种各样，昭和却好至极。

孤身邂逅真是会必成物是人非。雪落方才早，先是紧邻，原本过多教唆，万般无奈。新宿毕竟迁徙零散，可精神科等说工作四月四肢沉重至极。无关二战新宿。桐木箱画简单，感染因此迅速空档。

昭和积雪仿佛已帮她怪模怪样。鲜红万千思绪不难空档。但是祈祷倾诉万千得到此时。幸福石板路染上纯雪，写生娃娃此刻又夹着试练和吉祥等备好少女习俗。成年人仪式听到用度才迟钝。即使屋外景色曾捶打短刃，途中来日无多。倒不如一瞬身子银白盖住暖意。

2) 变种概率图:

希望你能找到了那幅她画的写生。

找到了那声音的主人，

找到了那幅她画的写生。

我歪过头。

毕竟，当你决定打开这扇门继续前行时，就必须抛弃所有的希望，你能找到那声音的主人。

夜幕下的公园中，孤身走在前方的你像是感到寒冷，耸起肩膀。感到一阵酥麻痛感。像是刺骨的寒意。雪花飘零的公园里万籁俱寂。夹着一些白色。让人感到寒冷。夹着雪粒儿的夜风让人感到刺骨的寒意。可比起疼痛，你最先感到刺骨的冰冷的刀锋。

我变得只剩下了空壳。

捏着剩下的泥土从怀里取出了一片事先备好的纸。泥土捧

在手中,填进了箱子周围。因为女孩人偶已被放回到了三宝台里。

要完成这个人偶了。

所以贪婪的吸收各种各样的东西,绝没有那么简单。

街上的路灯光被池水反射,摇摇曳曳。刀刃反射着道光芒。让人不禁想用只画笔记录下眼前的一切。

可以看出,seq2seq 使用十余倍变种概率图模型的时间进行训练和生成,对训练文本风格学习尚可,生成的部分句子词性上也比较通顺,但含义错配问题严重,且前后句之前逻辑联系缺乏。而变种概率图生成的文本前后段落具有逻辑联系,并且生成了可完整叙事的独立段落(如“夜幕下的公园”一段) ,其他部分也使用了独立生成的修辞短语,并非对训练文本的机械重复。措辞和叙事手法说明成功学习了训练文本的独白风格,基本达到了我们依照训练文本风格指定摘要词生成新文本的目的。在可解释性方面,查看训练后的摘要块连接可以发现,它确实反映了不同层级的语义联系。可以说,变种概率图在此类问题上的表现一定程度上优于 seq2seq 模型。

通过实验也发现变种概率图生成存在的一些缺陷,如在初期生成中经常机械重复同一短语,因为该短语中的词具有较高的概率值,通过联合概率计算后可以提高序列总概率。因此在生成序列的过程中对重复扩展同一个摘要块的情况作了限制,并将限制数目作为超参数。以上文本为置该参数为 2 时生成,但依然有局部连续重复的情况。另外对于以上文本出现了跨句子反复使用同一修辞(“刺骨的”) 和时态错配(“希望你能找到了那幅写生”) 等现象,此类问题可以通过扩大训练集解决。其他问题,如停用词的错误扩展(“反射着道光芒”) 可以通过具体工程模型中超参数的调整得到一定程度的解决。

6　结语

在对现有自然语言生成方法进行总结和分析的基础上,本文提出了基于图的文本生成模型并进行了系统实现。该模型在应用上并非与其他生成方法互斥,针对其他模型的特点对该模型进行改进,或使用一些方法将不同模型进行融合可能在解决特定领域问题上取得更好的成果。如利用词典展开释义[13] 丰富节点间的连接; 与其他自然语言判别模型结合,进行超参数和层间传递权值的进一步调整; 将“注意力机制”[14] 引入为边的特性,根据已经生成的前文自动暂时调整一些边的连接状态,以帮助解决生成长句子时容易重复同一词句的问题; 与神经网络融合,让神经网络生成自定义摘要块以简化指定所需文本特征的难度; 甚至进行半监督增量[15] 。这些都是可以将来进一步研究的问题。

参考文献:

[1] 司畅,张铁峰. 关于自然语言生成技术的研究[J]. 信息技术,2010(9):108 -110.

[2] 陈思佳. 实体关系抽取技术研究[D]. 北京: 北京邮电大学,2014.

[3] KARPATHY A. The unreasonable effectiveness of recurrent neural networks [EB/OL]. [2017-11-01]. http://karpathy. github. io/2015/05/21/rnn-effectiveness/.

[4] 张建华,陈家骏. 自然语言生成综述[J]. 计算机应用研究,2006,23(8):1 -3.

[5] ZHENG Y C. Text segmentation based on PLSA-TextTiling model[J]. Applied Mechanics & Materials, 2014, 556 -562:4018 -4022.

[6] 王双成,冷翠平,李小琳. 小数据集的贝叶斯网络结构学习[J]. 自动化学报,2009,35(8):1063 -1070.

[7] ZHANG Y, LIU K, HE S, et al. Question answering over knowledge base with neural attention combining global knowledge information [J]. arXiv, 2016, 2016: arXiv:1606.00979.

[8] 秦胜君,卢志平. 稀疏自动编码器在文本分类中的应用研究[J]. 科学技术与工程,2013,13(31):9422 -9426.

[9] 方宁. 基于认知的文本语境生成和度量研究[D]. 上海: 上海大学,2009.

[10] MIHALCEA R, TARAU P. TextRank: Bringing order into texts [C]// Proceedings of EMNLP 2004. Stroudsburg: Association for computation al Linguistics, 2004: 404 -411.

[11] TRAN V K, NGUYEN L M. Natural language generation for spoken dialogue system using RNN encoder-decoder networks [C]// CoNLL 2017: Proceedings of the 21st Conference on Computational Natural Language Learning. [S. l.]: Association for Computational Linguistics, 2017: 442 -451.

[12] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation [C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2002: 311 -318.

[13] 王小林,王义. 改进的基于知网的词语相似度算法[J]. 计算机应用,2011,31(11):3075 -3077.

[14] CINAR Y G, MIRISAEE H, GOSWAMI P, et al. Time series forecasting using RNNs: an extended attention mechanism to model periods and handle missing values[J]. arXiv, 2017, 2017: arXiv:1703.10089.

[15] 唐晓亮. 基于神经网络的半监督学习方法研究[D]. 大连: 大连理工大学,2009.

(上接第 38 页)

[4] JIA Y, SHELHAMER E, DONAHUE J, KARAYEY S, et al. Caffe: convolutional architecture for fast feature embedding[C]// Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM, 2014: 675 -678.

[5] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C/OL]// Proceedings of the 32nd International Conference on Machine Learning. 2015 [2017-11-01]. http://proceedings. mlr. press/v37/ioffe15. pdf.

[6] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines [C]// Proceedings of the 27th International Conference on Machine Learning. [S. l.]: Omnipress, 2010: 807 -814.

[7] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J/OL]. arXiv, (2014) [2017-11-01]. http://arxiv. org/abs/1409.1556.

[8] LIN M, CHEN Q, YAN S. Network in network[J/OL]. arXiv, (2013) [2017-11-01]. http://arxiv. org/abs/1312.4400.

[9] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]// ECCV 2014: Proceedings of the 2014 european Conference on Computer Vision. LNCS 8689. Berlin: Springer, 2014: 818 -833.

[10] YI D, LEI Z, LIAO S, et al. Learning face representation from scratch[J/OL]. arxiv, (2014-11-28) [2017-11-15]. https://arxiv. org/pdf/1411.7923. pdf.

[11] HUANG G B, RAMESH M, BERG T,et al. Labeled faces in the wild: a database for studying face recognition in unconstrained environments [R]. Amherst: University of Massachusetts, 2007: 7 -49.