计算机科学
COMPUTER SCIENCE

知识增强的自然语言生成研究综述

梁明轩, 王石, 朱俊武, 李阳, 高翔, 焦志翔

引用本文

梁明轩, 王石, 朱俊武, 李阳, 高翔, 焦志翔知识增强的自然语言生成研究综述[J]. 计算机科学, 2023, 50(6A): 220200120-8.

LIANG Mingxuan, WANG Shi, ZHU Junwu, LI Yang, GAO Xiang, JIAO Zhixiang. Survey of Knowledge-enhanced Natural Language Generation Research [J]. Computer Science, 2023, 50(6A): 220200120-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220200120

知识增强的自然语言生成研究综述

梁明轩[1,2]　王　石[2]　朱俊武[1]　李　阳[1,2]　高　翔[1,2]　焦志翔[1,2]
1 扬州大学信息工程学院　江苏 扬州 225000
2 中国科学院计算技术研究所　北京 100190
(yzulmx@163.com)

摘　要　自然语言生成(Natural Language Generation,NLG)任务是自然语言处理(Natural Languge Processing,NLP)任务中的一个子类,并且是一项具有挑战性的任务。随着深度学习在自然语言处理中的大量应用,其已经变成自然语言生成中处理各种任务的主要方法。自然语言生成任务中主要有问答任务、生成摘要任务、生成评论任务、机器翻译任务、生成式对话任务等。传统的生成模型依赖输入文本,基于有限的知识生成文本。为解决这个问题,引入了知识增强的方法。首先介绍了自然语言生成的研究背景和重要模型,然后针对自然语言处理归纳介绍了提高模型性能的方法,以及基于内部知识(如提取关键词增强生成、围绕主题词等)和外部知识(如借助外部知识图谱增强生成)集成到文本生成过程中的方法和架构。最后,通过分析生成任务面临的一些问题,讨论了未来的挑战和研究方向。

关键词:自然语言生成;知识增强;深度学习;知识图谱;关键词提取;主题词
中图法分类号　TP391

Survey of Knowledge-enhanced Natural Language Generation Research

LIANG Mingxuan[1,2],WANG Shi[2],ZHU Junwu[1],LI Yang[1,2],GAO Xiang[1,2] and JIAO Zhixiang[1,2]
1 College of Information Engineering,Yangzhou University,Yangzhou,Jiangsu 225000,China
2 Institute of Computing Technology,Chinese Academy of Sciences,Beijing 100190,China

Abstract　Natural language generation(NLG) task is a subclass of natural language processing(NLP) tasks and is a challenging task. With the massive application of deep learning in natural language processing,it has become the main method for handling various tasks in natural language generation. The main natural language generation tasks are question and answer tasks,summary generation tasks,comment generation tasks,machine translation tasks,generative dialogue tasks,etc. Traditional generative models rely on input text to generate text based on limited knowledge,and knowledge enhancement methods are introduced to solve this problem. Firstly,the research background and important models of natural language generation are introduced. Then,methods to improve model performance are introduced for natural language processing induction,and the methods and architectures based on the integration of internal knowledge(such as extracting keywords to enhance generation,surrounding subject words,etc.) and external knowledge(such as enhanced generation with the help of external knowledge graph) into the text generation process are introduced.. Finally,the future challenges and research directions are discussed by analyzing some problems faced by the generation task.

Keywords　Natural language generation,Knowledge enhancement,Deep learning,Knowledge graph,Keyword extraction,Subject headings

1　前言

随着人工智能、机器学习、深度学习研究的不断深入,衍生出了很多用于自然语言生成的方法[1]。文本到文本(Text-to-Text)的生成方法是以一段序列、文章作为输入,经过模型的训练,生成出所需的文本,如生成文章的结尾、论文的摘要、新闻的评论等等。除此之外,也包含了数据到文本(Data-to-Text)、图像到文本(Image-to-Text)等。

在自然语言生成中,深度学习的应用起到了重大作用[2],其中衍生出了一代经典的序列到序列(Seq-to-Seq)模型[3]。该模型是基于 Encoder-Decoder 的框架提出,将序列输入到编码器编码,通过解码器解码获得目标序列的方式类似于压缩解压的过程,中间难免会有语义的缺失。因此有些模型会加入 LSTM 和 GRU 等记忆网络变阵,以弥补 RNN 在处理 Long Term Memory 的不足。基于这些模型提出的注意力机制(Attention)和拷贝机制(指针神经网络 Pointer-Generator-

基金项目:国家 242 信息安全计划项目(2021A008);北京市科技新星计划交叉学科合作课题(Z191100001119014);国家重点研发计划重点专项(2017YFC1700300,2017YFB1002300);国家自然科学基金(61702234)
This work was supported by the National 242 Information Security Program(2021A008),Beijing NOVA Program (Z191100001119014),National Key Research and Development Program of China(2017YFC1700300,2017YFB1002300) and National Natural Science Foundation of China (61702234).
通信作者:王石(wangshi@ict.ac.cn)

Network),也促进了文本的生成。而这些方法都是从输入文本的内部处理角度去生成,没有外部知识的加持,导致模型生成的文本格式单一且枯燥。然而,我们希望生成的文本序列是富有感情、多样化的。

神经网络的目的是模拟人脑思考的方式解决问题。人脑思考、理解、解决问题的过程是,内部知识的约束和外部知识的扩充,或者通过先验知识能判断出解决的方案[4-5];知识增强包括内部知识增强和外部知识增强[8]。比如,天阴了预示着快下雨;我们要爱护动物,(狗,属于,动物)→ 所以我们也要爱护狗。虽然有一些预训练的模型(如谷歌提出的 GPT-2[6]模型)和大数据集量下预训练好的 Transformer[7]在各项测试中都有很好的表现,但是这些模型仍是使用大量的数据集才能实现的结果,在一些特定场景尤其是小数据集训练场景下表现不是很好。接下来主要介绍近期的调研:

(1)对 NLG 任务的各种模型作了全面的概述,基于知识增强模型架构的方法介绍了注意力机制、指针拷贝网络、记忆网络、图网络。

(2)介绍了两种知识增强的优化方法,即内部知识增强和外部知识增强,并介绍两种增强方法的具体应用和解决方案。

(3)概述了在 NLG 中充分应用知识增强的几大挑战和难点,为未来探索的方向提供讨论和建议。

2 生成模型的一般方法

2.1 基线模型

生成模型一般都是按照 Encoder-Decoder 的模式,在递归神经网络 RNN 模型的基础上提出。2014 年的 Sequence-to-Sequence 模型(Seq2Seq),编码器接收输入文本,并将其转化为向量表示,这个向量可以看作序列的语义表示;解码器负责接收编码器压缩过后的语义信息,并将向量表示的文本转化为所需文本表示。解码器本质是一个概率分布的统计,在输入文本和已生成的文本条件下预测下一个词。其中生成过程的数学表达如下:

$$P(Y|X)=\prod_{t=1}^{T}P(y_t|X,y_1,y_2,\cdots,y_{t-1}) \tag{1}$$

encoder 的隐藏向量的通用公式如下:

$$h_i=f_{\text{encoder}}(e(x_1),h_{i-1}) \tag{2}$$

将整个输入序列经过 encoder 后最终的隐藏向量记为 c,在 Seq2Seq 模型中,这个 c 将记入 decoder 的计算中:

$$s_t=f_{\text{decoder}}(e(y_{t-1}),s_{t-1},c) \tag{3}$$

$$p(y_t|y_1,\cdots,y_{t-1},X)=f_{\text{MLP}}(e(y_{t-1}),s_t,c) \tag{4}$$

文本的生成可以看作是多分类任务,通过交叉熵进行优化。整个序列的损失函数如下:

$$J_{\text{generate}}=-\frac{1}{N}\left(\sum_{i=1}^{N}y_i\log(\hat{y}_i)\right) \tag{5}$$

生成模型的训练引入了 Teacher Forcing[9]的训练优化方法,RNN 模型的生成阶段往往使用前一步的输出作为下一步的输入进行预测。这种训练方法往往出现一种问题:某一个单元出现完全错误结果,会使得后面所有单元的学习效果变差。对此,Teacher Forcing 提出在训练过程中忽略前一步的输出,将真实标签作为输入,但是这样会导致模型泛用性降低,只能贴合训练集使得模型脆弱。由此提出两点解决方案:集束搜索(Beam Search),通过设置 beam size 来保证生成多个候选序列。课程学习(Curriculum Learning),加入一个概率来决定是使用真实标签还是上一步的输出作为这一步的输入。

2.2 常见神经网络

近年来的模型核心思想都是 Encoder-Decoder 的架构,比如自动编码器 Auto Encoder,还有一种特殊的生成器对抗生成网络 GAN[10],两者都是通过学习一些数据产生类似的数据,但是其核心思想是不同的。

2.2.1 自动编码器 Auto Encoder

将一段文本编码再解码,X 是输入序列,$\hat{X}$ 是输出序列,模型目的是使 X 和 $\hat{X}$ 尽量相似。模型架构如图 1 所示。

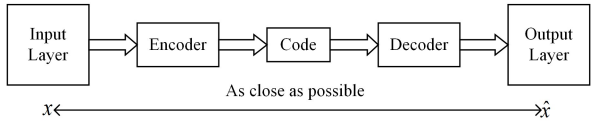

图 1 自编码器架构

Fig. 1 Autoencoder architecture

通过编码器得到隐藏向量 Code 的过程可被看作一个降维的过程,与 PCA[11]相似,再通过解码器生成的文本做到与原来尽量相似,但是当神经网络的参数复杂到一定程度时会出现过拟合风险。受此启发,如果使得输入的文本中混入一些随机的噪声,就能让模型更具有鲁棒性。Vincent 等人提出的 DAE(Denoising AutoEncoder)[12]就利用矩阵范式约束编码器,使得编码器可以学习到更完整的特征。在后来的模型中,这种随机的噪声被称为[Mask]。Devlin 等人提出的 Bert[13]是基于 Transformer[14]中编码器部分来构建的一种模型,整个架构就是基于 DAE 的,不过这部分在文章里被称为 MLM(Masked Language Model),随机地将一些单词掩盖,然后预测被掩盖的词。

2.2.2 变分自动编码器 VAE

Kingma 等[15]提出了 VAE 模型和 SGVB Estimator。其中 VAE 模型(见图 2)通过对输入文本加噪音,使得模型输入的 X 在经过编码器层后分出原有编码(Mean)和噪音编码(Variance),噪音编码通过给正态分布(Normal Distribution)分配一个权重再与原有编码相加得到 code 层的 $C_1,C_2,\cdots,C_n$,$C_i=\exp(\sigma_i)\times e_i+m_i$。此处的指数计算(exp)是保证分配的权重是正值。最后 VAE 还添加了一个必要的损失函数 $\sum_{i=1}^{n}(\exp(\sigma_i)-(1+\sigma_i)+(m_i)^2)$。

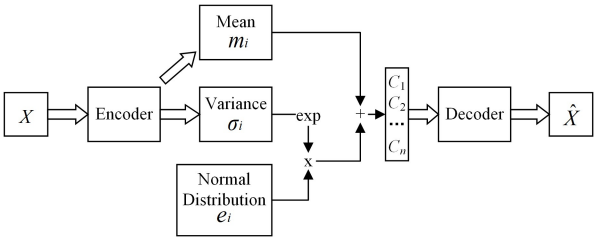

图 2 VAE 是在 AE 的基础上加入了噪音控制

Fig. 2 VAE adds noise control to AE

其中,$z\sim q_\Phi(z|x)$ 是建模的变分分布,$x\sim p_\theta(x|z)$ 是条件分布,隐变量先验概率为 $p_\theta(z)$。

SGVB 估计是对 VAE 求 ELBO 的方法,这种下界的巧妙之处在于将最后从连续分布中采样最大 x 的概率转化为求解下界的 L_b 的最大值,如图 3 所示。

$$L_b=E_{q_\Phi(z|x)}[\log p_\theta(x|z)]-KL[q_\Phi(z|x)||p_\theta(z)] \tag{6}$$

式(6)中期望可通过蒙特卡洛估计计算,从 $q_\Phi(z|x)$ 中依据 z 采样 L 个点:

$$E_{q_\varnothing(z|x)}[\log p_\theta(x|z)]\approx\frac{1}{L}\sum_{l=1}^{L}\log p_\theta(x|z^{(l)}) \tag{7}$$

由于式(7)对于 $\varnothing$ 不可导，通过重参数化技巧假设 $z^{(l)}=g_\varnothing(x,\varepsilon^{(l)})$，$\varepsilon^{(l)}\sim p(\varepsilon)$，其中 $p(\varepsilon)$ 和 $g_\varnothing$ 形式已知，则式(7)可变为：

$$E_{q_\varnothing(z|x)}[\log p_\theta(x|z)]=\frac{1}{L}\sum_{l=1}^{L}\log p_\theta(x|g_\varnothing(x,\varepsilon^{(l)})) \tag{8}$$

计算时假设 $p_\theta(z)\sim\mathrm{N}(z;0,I)$，$q_\varnothing(z|x)\sim\mathrm{N}(z;\mu,\sigma)$，$z$ 的维度是 J，通过转换得最终 Loss 为：

$$L(\varnothing,\theta,x)=\frac{1}{L}\sum_{l=1}^{L}\log p_\theta(x|g_\varnothing(x,\varepsilon^{(l)}))+\frac{1}{2}\sum_{j=1}^{J}(1+\log(\sigma_j^2)-\mu_j^2-\sigma_j^2) \tag{9}$$

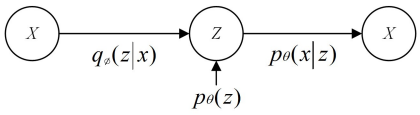

图 3　VAE 条件建模过程

Fig. 3　VAE condition modeling process

2.3　知识融合方法

生成任务中，为了使输入序列的编码融合更多语义，亦或者为了使解码过程多结合上下文，往往需要一种方法来增强模型理解输入序列。我们称这种方法为知识融合方法。

通过整合知识融合的方法，几种通用流行的方法经常被用于模型中：注意力机制、指针拷贝网络、图网络。

2.3.1　注意力机制

注意力机制[15,17]是为了学习输入序列和输出序列之间的软对齐。生成任务中输入序列中只有某些单词与预测下一个词相关，因此需要引入注意力机制帮助模型对那些单词的关注度给予足够的权重。一般地，计算注意力需要用到编码器输出最终状态 $\bar{h}_s$ 和目标隐藏状态 h_t。表 1 列出了自然语言生成的注意力机制一般方法。

在生成器的解码阶段，通常使用自注意力机制对输入序列内各个单词赋予权重。Vaswani 等[14]提出的多头自注意力机制让每一个单词都学习多个权重，从而得到的矩阵能包含某个字在其他字多个角度(比如状态，包含关系…)的语义信息。由于设置很多头会导致冗余，Hao 等[23]提出 MG-SA 方法，让 1/4 的 head 作自注意力机制，其他的 head 作 n-gram 短语划分。这样的划分可以捕捉到不同粒度的语义信息，增强了编码器的编码能力。

表 1　自然语言生成的注意力机制方法

Table 1　Attention mechanism approaches for natural language generation

名称	隐藏状态	注意力机制	文献
加法注意力 Bahdanau Attention	$s_t=f(s_{t-1},c_t,y_{t-1})$	$c_t=\sum_{j=1}^{T_x}\alpha_{tj}h_j$，$\alpha_{tj}=\frac{\exp(e_{tj})}{\sum_{k=1}^{T_x}\exp(e_{tk})}$	[16-17]
	$e_{tj}=a_t(s_{t-1},h_j)=\mathrm{softmax}(\tanh(s_{t-1}W_{\mathrm{decoder}}+h_jW_{\mathrm{encoder}})W_{\mathrm{alignment}})$ 是 t 时刻解码器隐藏状态和编码器输出的隐藏状态的注意力得分函数，本质上是一种加法性质的注意力机制，将解码器隐藏状态和编码器所有位置输出通过线性变换进行对齐，得到上下文向量。本质上是两层全连接网络，激活函数为 tanh		
点乘注意力 Luong Attention	$s_t=\tanh(W_c[c_t;h_t])$	$a_t(s)=score(h_t,\bar{h}_s)=\frac{\exp(score(h_t,\bar{h}_s))}{\sum_{s'}\exp(score(h_t,\bar{h}_{s'}))}$	[5,18]
	本质上是一种乘法性质的注意力机制，通过计算解码器隐藏状态和编码器的输出乘法结果，得到上下文向量，由 3 种得分函数进行对齐，最后将这些向量输入 softmax 函数得到权重		
自注意力 Self-Attention	$s_t=\sum_{i=1}^{t}\alpha_iH_i$	$\alpha=\mathrm{softmax}(W_1\tanh(W_2H^T))$	[19,21]
	自注意力用于计算来源于相同目标的隐藏向量，通过计算不同注意力系数下的句子嵌入得到加权的单词表征		
多头注意力 Muti-Head Attention	$s_t=\sum_i\alpha_iV_i$	$\alpha=\mathrm{softmax}(score(Q,K_i))=\frac{\exp(score(Q,K_i))}{\sum_j\exp(score(Q,K_j))}$	[22-23]
	将 QKV 通过参数矩阵映射到多头注意力层，各自都做注意力机制，重复几次再拼接起来，在序列内寻找内部关系，是自注意力机制的一种拓展		

而在编码器与解码器的连接处通常使用 Bahdanau 注意力机制和 luong 注意力机制，将输入序列压缩成一个隐藏向量，再在解码阶段根据这个隐藏向量进行预测。而注意力机制计算的则是编码器的隐藏向量和解码器当前隐藏状态的一个相关性，相似度计算的方式有以下 4 种。

$$score(h_t,\bar{h}_s)=\begin{cases}h_t^T\bar{h}_s, & \text{dot}\\ h_t^TW_a\bar{h}_s, & \text{general}\\ v_a^T\tanh(W_a[h_t;\bar{h}_s]), & \text{cocat}\\ \bar{h}_s^T\tanh(Wh_t+U\bar{h}_s), & \text{perceptron}\end{cases} \tag{10}$$

2.3.2　指针拷贝网络

拷贝网络(CopyNet)[54]采用类似于注意力机制的方法，具有 Seq2Seq 的结构，因此解码器的隐藏状态通常为 $s_t=f(y_{t-1},s_{t-1},c_t)$。一般地，此处 y_{t-1} 是上一个时间点生成的词，CopyNet 对其作了注意力机制，将得到的 $\gamma(y_{t-1})$ 拼接到上一时间点生成的词向量 $[e(y_{t-1});\gamma(y_{t-1})]$ 后：

$$\gamma(y_{t-1})=\sum_{\tau=1}^{T_x}\rho_{t\tau}h_\tau$$

$$\rho_{t\tau}=\begin{cases}\frac{1}{K}p(x_\tau,c|s_{t-1},M), & x_\tau=y_{t-1}\\ 0, & \text{otherwise}\end{cases} \tag{11}$$

此处 $\gamma(y_{t-1})$ 的计算方法类似于注意力机制，当上一时刻的输出在原文本中出现过，系数才有非零值。

指针拷贝网络(Pointer-Generator-Network)[54]中，指针类似于 C++ 的指针，通过从源文本中复制单词，保留产生

新词的能力。通常 RNN 无法处理 OOV(Out Of Vocabulary)问题,无法从输入文本中复制。Coverage Vector 利用注意力分布关注目前已被覆盖的词,并且网络再注意这些词的时候予以惩罚(损失函数)。通过式(10)的 perceptron 对齐函数对编码器隐藏状态h_i和解码器状态s_t计算权重,得$h_t^* = \sum_i a_i^t h_i$,注意力系数a_i^t看作源文本的单词概率分布,有最终生成概率:

$$\begin{aligned} P_v(w) &= \text{softmax}(W'(W[s_t; h_t^*]+b)+b' \\ P_{\text{gen}} &= \sigma(W_{h^*}^{\mathrm{T}} h_t^* + w_s^{\mathrm{T}} s_t + w_x^{\mathrm{T}} x_t + b_{\text{gen}}) \\ P(w) &= P_{\text{gen}} P_v(w) + (1-P_{\text{gen}}) \sum_{i:w_i=w} a_i^t \end{aligned} \tag{12}$$

计算的生成概率$P_{\text{gen}} \in [0,1]$决定从原文本复制单词或者词汇表生成,让模型具有复制能力,不再局限于词汇表的生成。

2.3.3 图网络

图神经网络是一种将知识转换为图形结构化数据的神经网络。图神经网络学习图中每个节点的嵌入,并用边连接各个节点构成图的结构。节点嵌入的方式利用了输入节点嵌入和图结构。

当给定目标序列时,首先通过构建每个序列间的关系建立依赖树,表示为$(w_i, r_{i,j}, w_j)$,$w_i w_j$为单词节点,$r_{i,j}$表示w_i w_j依赖关系。接着对相邻关系进行编码,记图为$G=(V,E)$,V是实体节点集合,E是边的集合。关系集合维度为$V \times V$。最后,建立依赖图,用v_i和v_j表示$w_i w_j$的实体节点,添加从节点v_i到节点v_j的无向边,类型为$r_{i,j}$。

通常使用 GNN 的方法将图的结构信息集成到文本生成中,依靠节点间边的信息记录两点的依赖关系,通过基于滤波器加强学习,利用输入节点嵌入和图结构,例如谱图卷积(GCN)[24]通过拉普拉斯对称矩阵对特征谱分解:$L = D^{-1/2} L D^{-1/2}$,D是L的度矩阵。最终可得到激活的节点向量:

$$H^l = \sigma(\widetilde{D}^{-1/2} \widetilde{A} \widetilde{D}^{-1/2} H^{l-1} W^{l-1}) \tag{13}$$

其中,W是可训练参数,$\widetilde{A} = A + I_n$,I_n是单位矩阵,H是节点的嵌入。

3 知识增强促进生成

3.1 通过主题增强生成

在新闻评论生成的过程中,标题是很重要的资源,是对正文内容的压缩。如果标题太短,将无法获取重要资源。Li 等[22]指出,与传统新闻不同,在线新闻平台上的文章含有大量噪音,文章中的许多句子甚至与新闻的主要主题无关。对此,他们提出提取文章的关键词作为新闻的主题。这些关键词是理解文章故事的最重要的关键词,其中大多数是命名实体。文章提出以主题为中心点形成聚类,再利用(TextRank)文本等级关键词提取算法[25]提取出关键词。利用谱图滤波器嵌入图结构,在解码阶段将上下文向量c_i和解码器隐藏状态t_i拼接,计算生成下一个词的概率:

$$y_i = \text{softmax}(W_0(\tanh(W([t_i; c_i])+b))) \tag{14}$$

除了用主题词来增强生成外,还有一种给定主题,将其作为上下文向量嵌入模型,来对生成进行更强的监督。Liu[27]和 Jin 等[28]将主题词事先编码,在解码阶段将主题词和编码的序列一起解码,在生成概率部分加入加权和。Narayan 等[29]提出将单词嵌入和主题词拼接,一起嵌入作为编码器输入,让输入的序列和主题知识一起参与增强生成。Xing 等[31]也提出用预训练的 LDA 模型生成主题,让主题词作为先验知识,结合注意力机制和偏向生成概率主导文本的生成。

3.2 通过关键词增强生成

关键词往往代表一段序列最主要、最概括的部分,关键词提取出序列中最具代表性的单词,不同于词汇表。关键词的提取技术主要有 TF-IDF 和 TextRank 等,其以词频和词为节点建立共现关系,提取关键词。不管是对话系统还是摘要,关键词都有助于指导抓住序列中的要点,为接下来的生成过程提供重要线索。

Mou 等[32]提出基于关键词的对话生成 Seq2BF 模型,首先在给定请求后预测一个关键词作为回复的主题,根据请求每个词预估一个 PMI 指数,最高的作为关键词,认为关键词应当显式地出现在模型对请求的回复上。然后将请求编码,得到一个上下文向量,解码时会将关键词作为第一个词,再利用上下文向量继续解码。Serban 等[33]运用了多个关键词作为驱动,利用关键词粗糙地表达意思进行编码,再与原序列编码上下文向量共同预测词。

关键词的融合方法有 3 种[34]:拼接融合、门控融合、分层融合。其中,拼接融合是将句子的上下文向量和关键词向量直接拼接;门控融合中,c_r是编码器上下文向量,关键词向量为c_{key}:

$$\begin{aligned} gate &= \sigma(W c_r + U c_{\text{key}}) \\ c &= gate \odot c_r + (1-gate) \odot c_{\text{key}} \end{aligned} \tag{15}$$

分层融合将解码器状态s分别和编码器上下文及关键词向量融合,使原序列和关键词建立联系。

$$\begin{aligned} \beta_r &= \sigma(W_r s + U_r c_r) \\ \beta_{\text{key}} &= \sigma(W_{\text{key}} s + U_{\text{key}} c_{\text{key}}) \\ c &= \beta_r c_r + \beta_{\text{key}} c_{\text{key}} \end{aligned} \tag{16}$$

在文本生成中,通常在得到解码器的隐藏状态后加权相加两个隐藏状态c和s,经过 softmax 层得到最终概率最大的词。在提取关键词的步骤中,往往使用 ground-truth 关键词微调,将提取的关键词作为输入,用 ground-truth 的方法收敛。

3.3 通过知识图谱增强生成

知识图谱(Knowledge Graph)本质上是一种表达实体之间关系的语义网络,它用一种结构化的表示实现对外部客观规律的归纳总结。知识图谱一般表示为知识三元组:实体、关系、语义描述;也有表现为主语谓语宾语的格式,例如(实体,属性,属性值)(实体 1,关系,实体 2)的形式,来存储语义信息。因此知识图谱以图的形式展现,通过遍历连接可以轻松地借助详细的语义信息促进文本生成的任务。

知识图谱提供了一种先验知识的方法,模拟人脑思考的过程,根据新接收到的信息,结合头脑中的先验知识,选出最大可能性的结果,例如输入序列"天阴了",结合知识三元组(阴天,Relate-to,下雨),最终生成的序列"天阴了,要下雨了。"。因此知识图谱加入到文本生成任务中会指导生成往正确的方向,节点的嵌入和关系的路径选择在文本生成任务中必不可少。将知识图谱融入到生成任务中的难点,主要是知识图谱的嵌入和知识图谱的推理。

知识图谱定义为由实体、关系组成的有向图$G=\{g_1, g_2, \cdots, g_N\}$,$g_i = (h, r, t)$,如(台湾省,属于,中国)。TransE[35]模型将实体和关系表示在同一空间下,关系只有一个属性且不可逆,关系表示为两个实体的距离,视为平移

向量，所以两个实体可以通过小位移低误差连接，$h+r\approx t$，即$\overrightarrow{台湾省}+\overrightarrow{属于}=\overrightarrow{中国}$。将知识图谱嵌入并融合进文本生成的策略是将输入序列 X 和知识图谱 G 拼接处理[4-5,36]。

在问答任务中，针对一个事实问题，需要找到一种知识图谱三元组来回答，而如果语料库知识图谱不全，例如请求为“梁朝伟主演的无间道的导演是谁?”此时所现有三元组只有(梁朝伟，主演，无间道)(无间道，导演，刘伟强)，则需要多跳推理路径来回答这个请求。Zhang 等[37]的 VRN 变分推断网络认为在问答任务中的首要任务就是识别请求中的主题实体(Topic Entity)，之前例子中主题实体就是“梁朝伟”，对给定的请求和抽取的主题实体，从主题实体开始经过 N 跳找到结果a_i。通过主题实体计算请求答案的概率一般构造为：

$$\begin{aligned}P(a|y,q)&=\frac{\exp(f_q(q)^N g(G_{y\to a}))}{\sum_{a'\in V(G_y)}\exp(f_q(q)^N g(G_{y\to a'}))}\\ g(G_{y\to a})&=\frac{\sum_{a_j\in Para(a),(a,r,a_j)\in G_y}\sigma(V\times g(G_{y\to a_j}),\vec{e}_r)}{\#\,parent(a)}\end{aligned}\tag{17}$$

其中，$f_q(q)^N$ 是对请求的编码，V 是实体集合，$\vec{e}_r$ 是关系的 one-hot 形式编码。VRN 本质上将提取的主题实体作为隐变量的变分计算方式，计算请求的答案的全概率 $P(a|q)=\sum_y p(y|q)P(a|y,q)$。对于知识图谱上连接缺失的问题，部分三元组没有包含到先验知识图谱中，由此 Sun 等人提出了 GraftNet[38]模型将文档级作为一个节点，设计 GNN 进行推理，对主题实体提取以其为中心的子图，用 GNN 的迭代可以感知多跳实体的信息，但是这种利用规则的抽取方式往往会获得一些与答案无关的子图。作者在之后提出的 PullNet 模型[39]要求采用 GCN 迭代的方式判断知识图谱中实体是否需要扩展，这种动态的子图扩展方法减少了无关子图数量，由此提升性能。EmbedKGQA 模型[40]认为需要对候选的实体进行再次的筛选即关系匹配，设置一个 Complex 得分函数来评定可能的候选项。

3.4 通过加入情感极性增强生成

在生成任务中，大多数改进都是在生成结果的准确率上改进，很少有任务专注于情感强度控制生成结果。对于这项任务，有两方面挑战：1)如何获得具有情感极性的语料库；2)如何将情感极性融入进生成模型。

Luo 等[40]提出的端到端模型中包括了 3 种情感分析器，先对语料库的序列做情感分析任务，将情感标签放入每段序列开头。除了语义嵌入作为输入，也有序列的情感嵌入，通过一个线性层将情感极性转化为矩阵形式，再将嵌入映射到一个真实值，在生成器中加入高斯核层用于鼓励生成情感接近的词，在训练时计算情感分析的损失和生成过程的损失。在测试过程，用细粒度的情感极性控制生成结果，如表 2 所列。

表 2 同一个上下文，在不同情感强度下生成的文本

Table 2 Texts generated in the same context with different emotional intensities

情感强度	生成文本
0.1	She still lost the game and was very upset.
0.3	She almost won the game, but eventually lost.
0.5	The game ended with a draw.
0.7	She eventually won the game.
0.9	She won the game and was very proud of her team.

Qiao 等[41]提出基于变分自动编码器的 Data-To-Text 任务的模型 SCTKG，在编码阶段将主题序列、情感嵌入 $e(s)$、上下文向量编码作为条件向量 c，通过识别网络(训练阶段)、先验网络(推理阶段)计算潜在变量 z，解码阶段附加主题知识图谱和情感嵌入促进生成文本，最后引入 GAN 中的鉴别器对抗训练。SCTKG 融入情感极性的方式是以情感嵌入作为解码器隐藏向量 d 的一部分：

$$\begin{aligned}d_0&=W_d[z,c,e(s)]+b_d\\ P_t&=\mathrm{softmax}(W_0[d_t;e(s);g_t]+b_0)\end{aligned}\tag{18}$$

其中，g_t 是根据 luong 注意力机制计算的图注意力机制，通过拼接的方式融入，最后的损失函数与 CVAE 相同，同式(9)。

4 评估标准、数据集与工具

4.1 评估标准

文本生成任务中的评估标准不同于普通 NLP 任务的准确率评估方法，评估方法有人工评估和自动评估的方式。文本生成的评估标准方面包括主题一致性、新颖性、文章多样性、流畅性。下面整理了一些通用评估标准。

BLEU[43]在 2002 提出为了评估机器翻译的准确率，根据 n-gram 的方法可以划分多个评价指标，$n\in\{1,2,3,4\}$，n-gram 中 n 一般是指同窗口连续单词的数量。$n=1$ 时就是评估词的准确率，而 $n=2,3,4$ 时用于衡量句子的流畅性。以下公式的 p_n 是 n-gram 的概率，BP 是惩罚因子(对译文太短惩罚)。

$$\begin{aligned}BLEU&=BP*\exp(\sum_{n=1}^{N}W_n\log p_n)\\ BP&=\begin{cases}1, & \text{if } c>r\\ e^{(1-r/c)}, & \text{if } c\leqslant r\end{cases}\end{aligned}\tag{19}$$

ROUGE[44]以召回率作为指标，ROUGE-L 指标衡量的是两段文本的最长公共序列，按照词的顺序匹配，表示被正确预测词的个数。其中 ROUGE-N 用来评估摘要，引入马尔可夫假设，将系统生成的文本与人工生成的参考文本对比，计算二者的重叠度以评估表达的意思是否正确，其中 N 也是 n-gram 的窗口大小。

$$R=\frac{\sum_{S\in\{Summaries\}}\sum_{gram_n}Count_{\mathrm{match}}(gram_n)}{\sum_{S\in\{Summaries\}}\sum_{gram_n}Count(gram_n)}\tag{20}$$

其中，$Count_{\mathrm{match}}$ 是同时出现在生成文本和参考文本的 n-gram 个数，$Count$ 表示参考文本的 n-gram 个数。

METROR[45]在 BLEU 基础上加上了召回率，在计算前先提取生成文本与参考文本相同的词，提取相同词形的词和同义词，然后设输入文本词数为 n，生成文本词数为 t，参考文本词数为 r，先计算准确率 $P=n/t$ 和召回率 $R=n/r$，再将其加权后计算得分：

$$\begin{aligned}F&=\frac{PR}{\alpha P+(1-\alpha)R}\\ pen&=\gamma(ch/n)^{\beta}\\ METROR&=(1-pen)F\end{aligned}\tag{21}$$

其中，pen 为惩罚因子，ch 为文本分块的数量，α,γ,β 都是超参数。

PERPLEXITY[46]方法在内容层面评估内容是否符合语法且与主题相关，在测试集上得到的分数越低越好，计算公式如下：

$$\begin{aligned}P(S)&=p(w_1,w_2,\cdots,w_n)^{-\frac{1}{n}}\\ \log(P(S))&=-\frac{1}{n}\sum_{i=1}^{n}p(w_i|w_1,w_2,\cdots,w_n)\end{aligned}\tag{22}$$

PERPLEXITY 对数形式可以看作是交叉熵，描述为真实分布与预测分布之间的距离，经过数学推算两者等价，由此认为生成的文本概率越大，与参考文本差距越小，语言模型越好，PERPLEXITY 值越低，模型更贴合语料库。

基于词嵌入的评估方法，通过分布式的词嵌入表达来衡量词之间的相似度，例如 Greedy Matching[47] 贪心匹配方法依次计算词向量与文本中各个词的词向量的余弦相似度，最大值为相似度。

$$G(r,\hat{r})=\frac{\sum_{w\in r}\max_{\hat{w}\in\hat{r}}\cos(e_w,e_{\hat{w}})}{|r|}$$
$$GM(r,\hat{r})=\frac{G(r,\hat{r})+G(\hat{r},r)}{2} \tag{23}$$

其中，r 和 $\hat{r}$ 分别为生成文本和参考文本，通过计算两次取平均值，得到最终两个文本的相似度。

还有其他根据词嵌入评估的方法，如 Embedding Average 和 Vector Extrema 等[48-49]，都是在向量层级上比较两端文本的相似程度。

除以上通用的评估指数外，Jin 等[28] 对其设计的模型提出独特的评估标准 DMet 来衡量文本的多样性和新颖性，通过计算单词重叠、序列内相似性惩罚和句子长度校正来度量生成文本的多样性和新颖性。

4.2 数据集

根据调研，将数据集按照各项生成任务分类，如表 3 所列。

表 3 文本生成任务中的各项数据集及地址

Table 3 Various datasets and addresses in text generation task

任务	数据集				包括数据集的文献	地址
	名称	#训练	#验证	#测试		
生成摘要	ROTOWIRE	7633	1635	1635	[19,50]	github.com/harvardnlp/boxscore-data
	WIKIBIO	582661	2000	2000	[17,52-53]	github.com/DavidGrangier/wikipedia-biography-dataset
对话任务	Douban	2816000	20000	20000	[5]	github.com/MarkWuNLP/MultiTurnResponseSelection
	Stanford	2425	302	304	[50]	nlp.stanford.edu/projects/kvret/kvret_dataset_public.zip
评论任务	Reddit	3384185	10000	10000	[32,36]	reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/
	Yelp	6908323	10000	10000	[28]	yelp.com/dataset

4.3 工具

在神经网络模型加载数据前，往往需要对语料库数据进行预处理，当处理某项任务时，往往需要用到预处理后的数据，以下是调研到的一些实用工具。

HanLP[1] 上有开源的 16 种预处理任务（多语言分词、词性标注、命名实体识别、关键词提取、自动摘要、短语提取、拼音转换、简繁转换、文本推荐、依存句法分析、文本分类、文本聚类、语义分析、Word2vec、新词发现、知识图谱创建工具）并提供了免费接口。语言云平台[2] 也提供了中文分词、词性标注等其他一些预处理功能。

结束语 在文本生成任务中，引入知识增强的方法已然是一种主流的加强生成模型的方法，调研发现通过结合多种形式的知识指导和促进文本生成的方法效果提高显著。知识的来源不局限于上述的网络结构、图结构、字典和表格，还有规则、推理等过程的知识来源难以融入进神经网络模型，这会是以后研究工作的难点。知识增强的文本生成任务仍然存在的两个难点，一是知识的获取，二是知识在模型中的融合。

参 考 文 献

[1] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[2] LIU J W, LIU Y, LUO X L. Research Progress in Deep Learning[J]. Computer Application Research, 2014, 31(7): 11.

[3] JI H, KE P, HUANG S, et al. Language generation with multi-hop reasoning on commonsense knowledge graph[C]// Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2020: 725-736.

[4] ZHOU H, YOUNG T, HUANG M, et al. Commonsense knowledge aware conversation generation with graph attention[C]// IJCAI. 2018: 4623-4629.

[5] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[6] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 140: 1-140: 67.

[7] YU W, ZHU C, LI Z, et al. A survey of knowledge-enhanced text generation[J]. arXiv: 2010.04389, 2020.

[8] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. MIT Press, 2016.

[9] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[10] TENENBAUM J B, DE SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.

[11] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]// Proceedings of the 25th International Conference on Machine Learning. 2008: 1096-1103.

[12] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4171-4186.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances inNneural Information Processing Systems. 2017: 5998-6008.

[14] KINGMA D P, WELLING M. Auto-Encoding Variational Bayes[J]. Stat, 2014, 1050: 1.

[15] BAHDANAU D, CHO K H, BENGIO Y. Neural machine trans-

1) https://www.hanlp.com/index.html

2) http://www.ltp-cloud.com/

lation by jointly learning to align and translate[C]// Proceedings of 3rd International Conference on Learning Representations (ICLR). 2015.

[16] FU Z,SHI B,LAM W,et al. Partially-Aligned Data-to-Text Generation with Distant Supervision[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing(EMNLP). 2020:9183-9193.

[17] LUONG M T,PHAM H,MANNING C D. Effective Approaches to Attention-based Neural Machine Translation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015:1412-1421.

[18] PUDUPPULLY R,DONG L,LAPATA M. Data-to-text generation with content selection and planning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:6908-6915.

[19] LIN Z,FENG M,SANTOS C N,et al. A structured self-attentive sentence embedding[C]// Proceedings of 3rd International Conference on Learning Representations(ICLR). 2017.

[20] ZHANG H,GOODFELLOW I,METAXAS D,et al. Self-attention generative adversarial networks[C]// International Conference on Machine Learning. PMLR,2019:7354-7363.

[21] LI W,XU J,HE Y,et al. Coherent Comments Generation for Chinese Articles with a Graph-to-Sequence Model[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:4843-4852.

[22] HAO J,WANG X,SHI S,et al. Multi-Granularity Self-Attention for Neural Machine Translation[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). 2019:887-897.

[23] KIPF T N,WELLING M. Semi-supervised classification with graph convolutional networks[C]// Proceedings of 3rd International Conference on Learning Representations(ICLR). 2017.

[24] MIHALCEA R,TARAU P. Textrank:Bringing order into text [C]// Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004:404-411.

[25] WU L,CHEN Y,SHEN K,et al. Graph Neural Networks for Natural Language Processing:A Survey[J]. arXiv:2106.06090, 2021.

[26] LIU Y,LIN Z,LIU F,et al. Generating paraphrase with topic as prior knowledge[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019:2381-2384.

[27] JIN K,ZHANG X,ZHANG J. Learning to Generate Diverse and Authentic Reviews via an Encoder-Decoder Model with Transformer and GRU[C]// 2019 IEEE International Conference on Big Data(Big Data). IEEE,2019:3180-3189.

[28] NARAYAN S,COHEN S B,LAPATA M. Don't Give Me the Details,Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018:1797-1807.

[29] 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. 2019:2002-2012.

[30] XING C,WU W,WU Y,et al. Topic aware neural response generation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2017.

[31] MOU L,SONG Y,YAN R,et al. Sequence to Backward and Forward Sequences:A Content-Introducing Approach to Generative Short-Text Conversation[C]// The 26th International Conference on Computational Linguistics:Technical Papers(COLING 2016). 2016:3349-3358.

[32] SERBAN I,KLINGER T,TESAURO G,et al. Multiresolution recurrent neural networks:An application to dialogue response generation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2017.

[33] LI H,ZHU J,ZHANG J,et al. Keywords-guided abstractive sentence summarization[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:8196-8203.

[34] BORDES A,USUNIER N,GARCIA-DURAN A,et al. Translating Embeddings for Modeling Multi-relational Data[C]// Neural Information Processing Systems(NIPS). 2013:1-9.

[35] ZHANG H,LIU Z,XIONG C,et al. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:2031-2043.

[36] ZHANG Y,DAI H,KOZAREVA Z,et al. Variational Reasoning for Question Answering With Knowledge Graph[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.

[37] SUN H,DHINGRA B,ZAHEER M,et al. Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018:4231-4242.

[38] SUN H,BEDRAX-WEISS T,COHEN W. PullNet:Open Domain Question Answering with Iterative Retrieval on Knowledge Bases and Text[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). 2019:2380-2390.

[39] SAXENA A,TRIPATHI A,TALUKDAR P. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:4498-4507.

[40] LUO F,DAI D,YANG P,et al. Learning to control the fine-grained sentiment for story ending generation[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:6020-6026.

[41] QIAO L,YAN J,MENG F,et al. A Sentiment-Controllable Topic-to-Essay Generator with Topic Knowledge Graph[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing:Findings. 2020:3336-3344.

[42] 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. 2002:311-318.

[43] DODDINGTON G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics[C]// Proceedings of the Second International Conference on Human Language Technology Research. 2002:138-145.

[44] LAVIE A, AGARWAL A. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments[C]// Proceedings of the Second Workshop on Statistical Machine Translation. 2007: 228-231.

[45] SERBAN I V, SORDONI A, BENGIOY, et al. Hierarchical neural network generative models for movie dialogues[J]. arXiv: 1507.04808, 2015.

[46] RUS V, LINTEAN M. An optimal assessment of natural language student input using word-to-word similarity metrics [C]// International Conference on Intelligent Tutoring Systems. Berlin: Springer, 2012: 675-676.

[47] WIETING J, BANSAL M, GIMPEL K, et al. Towards universal paraphrastic sentence embeddings[C]// Proceedings of 3rd International Conference on Learning Representations (ICLR 2016). 2016.

[48] FORGUES G, PINEAU J, LARCHEVÊQUE J M, et al. Bootstrapping dialog systems with word embeddings[C]// Nips, Modern Machine Learning and Natural Language Processing Workshop. 2014.

[49] XU P, HU Q. An End-to-end Approach for Handling Unknown Slot Values in Dialogue State Tracking[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers). 2018: 1448-1457.

[50] ISO H, UEHARA Y, ISHIGAKI T, et al. Learning to select, track, and generate for data-to-text[J]. Journal of Natural Language Processing, 2020, 27(3): 599-626.

[51] TRISEDYA B, QI J, ZHANG R. Sentence generation for entity description with content-plan attention[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 9057-9064.

[52] SHAHIDI H, LI M, LIN J. Two Birds, One Stone: A Simple, Unified Model for Text Generation from Structured and Unstructured Data[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 3864-3870.

[53] GU J, LU Z, LI H, et al. Incorporating Copying Mechanism in Sequence-to-Sequence Learning[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Long Papers). 2016: 1631-1640.

[54] SEE A, LIU P J, MANNING C D. Get To The Point: Summarization with Pointer-Generator Networks[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Long Papers). 2017: 1073-1083.

LIANG Minxuan, born in 1998, master. His main research interest is natural language processing.

WANG Shi, born in 1981, Ph.D, associate researcher, is a member of China Computer Federation. His main research interests include natural language processing semantic analysis and knowledge graph.