文章编号：2095-2163(2020)06-0101-05 中图分类号：TP391.1 文献标志码：A

基于全局编码信息的生成式自动文摘模型

宋治勋，赵铁军
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150000）

摘 要：目前主流的生成式自动文摘方法通常采用基于注意力机制的序列到序列模型，然而此方法只能使解码器关注于原文本中重要单词的语义信息，无法对每一维度的语义特征进一步筛选。经本文研究发现，编码端输出的隐状态语义信息存在噪声，解码器应该对各维度语义信息进行判断，选择对文摘生成更有益的语义特征。因此，本文提出了基于全局编码信息的生成式自动文摘模型，并构造一个融合全局编码信息的选择门控单元，二者相互协作，共同解决编码器输出信息存在噪声的问题。实验结果表明，本文提出的模型在 TTNews 数据集上获得优于前人的结果，通过组件消融分析验证各组件对文摘生成的积极作用。

关键词：生成式自动文摘；全局编码信息；选择门控单元；注意力机制

Abstractive Summarization Based on Global Features of Encoding

SONG Zhixun, ZHAO Tiejun
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150000, China)

【Abstract】 Nowadays sequence to sequence model based on attention mechanism is usually used in abstractive summarization. However, this method only makes the decoder focus on the semantic information of important words in original text, without further filtering the semantic information of words. In this paper, we find that the hidden state semantic information output by encoder is full of noise, and the decoder should judge the semantic information of each dimension and choose more influential semantic features for summary. In view of this, we propose an abstractive summarization method based on global features of encoding, and construct a selective gating unit using global encoding information to deal with noisy encoder output. The results show that our model achieves the best results on TTNews. The ablation study confirms the effectiveness of our model components.

【Key words】 abstractive summarization; global encoding; selective gating unit; attention mechanism

0 引 言

目前主流生成式模型借鉴神经机器翻译中基于编码器-解码器结构的序列到序列模型[1]。这类模型通常采用循环神经网络作为网络结构的基本单元，并利用注意力机制实现源端与目标端的词语对齐。本文认为这种做法存在一些问题，原文与摘要间不存在与机器翻译类似的显示词对齐关系，文摘任务的难点也不在于计算源端和目标端词对齐关系，而是在于解码器如何关注到编码端输出的重要语义信息。同时，编码器输出的隐状态向量中存在噪声，模型需要在信息传递过程中筛选语义向量各维度的重要特征，否则会导致解码器对源端核心内容认识不充分，造成文摘语法错误以及语义不统一等问题。

为解决上述问题，本文对目前主流基于注意力机制编码器-解码器生成式文摘模型作出以下两点改进：(1)引入全局编码上下文信息，并使全局编码信息既参与编码器注意力计算和选择门控单元计算，又参与解码器每个时间步注意力更新以及目标端词表概率计算过程，模型通过参数共享实现全局编码信息的梯度更新。(2)引入选择门控单元，在信息传递过程中过滤编码器输出的对摘要生成无益的语义信息，修正编码器每个时间步的语义表示。实验证明，本文提出的模型在 TTNews 数据集上分别取得 2.2 个百分点的提升。依据组件消融分析实验结构，证明本文引入的各组件对摘要生成起不同程度的提升作用。

1 相关工作

关于生成式文摘方法，早期研究者们关于基于规则生成的方法、基于语法树剪枝的方法、以及基于语言学规律等方法。随着深度学习技术的发展，很多学者开始展开对序列到序列的生成式文摘模型的研究。Rush 首次在序列到序列模型中应用注意力机制解决生成式文摘任务，该模型在 Gigaword 测试集取得当时最好的成绩。随后，Chopra 对该模型做了改进，将解码器替换为循环神经网络；Nallapati 则将模型修改为完全以循环神经网络为基本单元的序列到序列模型，同时在编码端引入人工特征；Gu 于

作者简介：宋治勋(1996-)，男，硕士研究生，主要研究方向：自然语言处理、深度学习。

收稿日期：2020-04-21

2016 年提出 CopyNet 模型,通过引入拷贝机制来模拟人类生成摘要时从原文选择词语的过程,能够比较好处理 UNK 词语生成的问题。Gulcehre 提出使用门控机制,控制解码器每一时间步是从原文复制词语还是输出目标端词表概率最大的词。

2 基于全局编码信息的生成式文摘

本文提出的模型以基于注意力机制的序列到序列模型架构为基础,引入同时参与模型编码端、解码端计算的全局编码信息。同时,本文提出融合全局编码信息的选择门控单元,用于修正编码器输出语义信息。本文提出模型的结构图,如图 1 所示。

2.1 问题定义

对于生成式文摘任务,模型接收输入序列 $x = (x_1, x_2, \cdots, x_n)$,其中 n 表示输入文档长度,x_i 表示输入文档中的第 i 个词语,$x_i \in V_s$,V_s 表示源端词表。模型接收序列 x 后,生成一段简短的文摘序列 $y = (y_1, y_2, \cdots, y_l)$,其中:$l \leqslant n$ 是文摘序列的长度,y_i 表示文摘序列中的第 i 个词语,$y_i \in V_T$,V_T 表示目标端词表。通常情况下,生成式文摘任务允许 $|y| \nsubseteq |x|$ 的情况出现,表示输出文摘序列的词语可以未在输入文档中出现过。

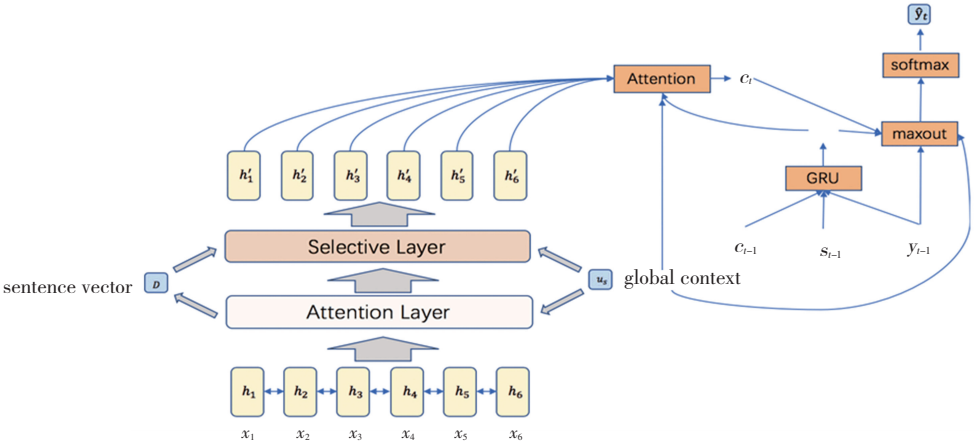

图 1 基于全局编码信息的生成式文摘模型结构图

Fig. 1 Overview of abstractive summarization model based on global features of encoding

2.2 编码器

本文提出的模型以基于注意力机制的序列到序列模型为基础,遵循编码器-解码器架构。编码器使用双向门控循环单元(BiGRU)实现,负责按序接收输入文档中每个单词的词向量 $w^S = (w_1^S, w_2^S, \cdots, w_n^S)$,$w_i^S$ 由输入文档中第 i 个词语 x_i 对词嵌入矩阵 $W^S \in \mathbb{R}^{|V_s| \times d_w}$ 查表得到。之后,编码器输出循环神经单元计算得到的每个单词的隐状态 $h = (h_1, h_2, \cdots, h_n)$,作为单词的上下文信息表示。门控循环单元 GRU 的计算公式(1)~公式(4)如下:

$$z_i = \sigma(W_z [w_i^S; h_{i-1}]), \tag{1}$$

$$r_i = \sigma(W_r [w_i^S; h_{i-1}]), \tag{2}$$

$$\widetilde{h}_i = tanh(W_h [w_i^S; r_i \odot h_{i-1}]), \tag{3}$$

$$h_i = (1 - z_i) \odot h_{i-1} + z_i \odot \widetilde{h}_i. \tag{4}$$

其中,W_z、W_h、W_h 均为权重矩阵,方便起见,公式中省略了偏置 b。由于这里使用的是双向门控循环单元,第 i 个时刻的隐状态 h_i 实际上由两个不同方向的隐状态向量级联得到 $h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$,$h_i \in \mathbb{R}^{2d_h}$,$d_h$ 是 GRU 的隐状态向量维度。

2.3 编码阶段

本文在编码端引入全局编码信息 $u_s \in \mathbb{R}^{2d_h}$,是模型可学习的向量参数,随整个训练集的学习过程不断更新。本文利用全局编码信息 u_s 与编码器 BiGRU 输出的隐状态 h,计算源端注意力分布,利用每个单词的注意力打分对隐状态 h 加权求和,得到输入文档的表示 D,计算公式(5)~公式(7)如下:

$$u_i = tanh(W_s h_i + b_s), \tag{5}$$

$$\alpha_i = \frac{\exp(u_i^{\mathrm{T}} u_s)}{\sum_i \exp(u_i^{\mathrm{T}} u_s)}, \tag{6}$$

$$D = \sum_i \alpha_i h_i. \tag{7}$$

其中,α_i 是使用 *softmax* 函数分配的输入序列概率权重,$D \in \mathbb{R}^{2d_h}$ 是整个输入序列的向量表示。本文采用公式(7)计算得到的向量 D 作为整个输入文档的表示,并传递给解码端用于初始化解码器隐状态 s_0。

2.4 融合全局编码信息的选择门控单元

本文提出融合全局编码信息的选择门控单元,

以控制编码端信息流动。具体来讲，选择门控单元接收编码器输出的隐状态 h、全局编码信息 u_s、输入序列的向量表示 D。对于每个单词 x_i，选择门控单元输出一个门控向量 g_i，并使用门控向量 g_i 和隐状态 h_i 计算调整后的编码器隐状态 h_i'。上述过程计算公式(8)和公式(9)如下：

$$g_i = \sigma(W_g [h_i; u_s; D] + b_g), \tag{8}$$

$$h_i' = g_i \odot h_i. \tag{9}$$

其中，W_g 是参数矩阵，b_g 是偏置向量。σ 是 *sigmoid* 激活函数，$\odot$ 表示点乘操作。

经过融合全局编码信息的选择门控单元，编码端输出更加精细的隐状态向量 $h' = (h_1', h_2', \cdots, h_n')$，之后将隐状态向量传递给解码器，用于生成相应的摘要词语。

2.5　解码器

本文使用单向门控循环单元(GRU)作为解码器，通过逐词生成的方式输出文摘序列直到输出标志句子结束的词语 EOS。

首先，解码器接收编码器输出的文档向量表示 D，初始化 GRU 的隐状态，公式(10)：

$$s_0 = tanh(W_d D + b_d). \tag{10}$$

在每个解码时刻 t，解码器同时接收上一时刻的文摘序列输入 y_{t-1}、上一时刻解码器的输出的隐状态 s_{t-1}、以及上一时刻的编码器-解码器上下文向量 c_{t-1}，经过循环单元计算，输出当前时刻的隐状态 s_t，公式(11)：

$$s_i = GRU(w_{t-1}^{\mathrm{T}}, s_{t-1}, c_{t-1}). \tag{11}$$

其中，w_{t-1}^T 由文摘序列中第 $t-1$ 个词语 y_{t-1} 对词嵌入矩阵 $W^T \in \mathbb{R}^{|V_T| \times d_w}$ 查表得到，d_w 是词向量维度大小。公式(11)中使用的编码器-解码器上下文向量 c 是通过 Luong 提出的级联注意力机制计算得到的，公式(12)~公式(14)：

$$e_{t,i} = v_a^T tanh(W_a s_{t-1} + U_a h_i'), \tag{12}$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_i \exp(e_{t,i})}, \tag{13}$$

$$c_t = \sum_i \alpha_{t,i} h_i'. \tag{14}$$

之后，采用 Goodfellow 提出的 maxout 网络，构造特征向量 r_t，使用窗口大小为 2 最大池化层提炼 r_t 重要语义信息，最后使用全连接层输出词表 V_T 大小的概率分布向量。上述过程的计算公式(15)~公式(17)：

$$r_t = W_r [w_{t-1}^{\mathrm{T}}; c_t; s_t], \tag{15}$$

$$m_t = [\max\{r_{t,2j-1}, r_{t,2j}\}]_{j=1,\cdots,d}, \tag{16}$$

$$p(\hat{y}_t \mid y_{<t}) = logsoftmax(W_o m_t). \tag{17}$$

其中，W_r、W_o 是权重矩阵，特征向量 $r_t \in \mathbb{R}^{2d}$，公式(16)比较每两个相邻特征的最大值，得到更加精细的特征向量 $m_t \in \mathbb{R}^{2d}$。y 表示模型训练阶段输入的真实文摘词语，$\hat{y}$ 表示模型预测的文摘词语。

2.6　解码阶段

本文认为文摘模型需要在解码端融入更多编码端的信息。这是因为在推理阶段，解码器仅仅依靠原文与已经输出的摘要序列计算下一时刻的摘要词语，而已输出的摘要序列可能会存在错误，解码器需要融入更多的编码端信息以指导下一时刻的摘要词汇生成。

基于上述考虑，本文将全局编码信息 u_s，进一步应用于模型解码阶段。首先，在计算编码器-解码器上下文向量 c_t 时，引入全局编码信息，将公式(12)修改为公式(18)：

$$e_{t,i} = v_a^T tanh(W_a s_{t-1} + U_a h_i' + E_a u_s). \tag{18}$$

此外，本文在计算词表 V_T 的概率分布向量是也融入全局编码信息 u_s，重新构造信息量更为全面的特征向量 r_t，将公式(15)修改为公式(19)：

$$r_t = W_r [w_{t-1}^T; c_t; s_t; u_s]. \tag{19}$$

如上所述，本文将全局编码信息应用于编码器-解码器上下文向量计算，以及词表概率分布向量计算，除了能够起到丰富特征信息量、融入更多编码端信息以指导文摘生成的作用。

2.7　训练

给定模型的可学习参数 θ 和输入文本序列 x，文摘模型输出预测的文摘序列 $\hat{y}$。模型的训练目标如公式(20)：

$$J(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T} p(y_t^{(n)} \mid \hat{y}_{<t}^{(n)}, x^{(n)}, \theta). \tag{20}$$

其中，$y_t^{(n)}$ 表示第 n 个训练实例第 t 时刻的真实文摘词语，$\hat{y}_{<t}^{(n)}$ 表示第 n 个训练实例前 $t-1$ 时刻的预测文摘序列，$x^{(n)}$ 表示第 n 个训练实例的输入文本序列。

3　实验设计与分析

3.1　数据集

TTNews 是 NLPCC2017 单文档文摘任务发布的评测数据集，源于今日头条平台上的真实数据，共包含 50 000 条训练集，2 000 条测试集。TTNews 数据集的平均原文长度为 669.0，平均摘要长度为 35.8，原文词表大小为 28 755，摘要词表大小为 9 998。

3.2 数据预处理

在数据预处理阶段,考虑到中文分词错误造成的词典误差,本任务采用 subword 方式重构中文文摘数据。相较于传统的分词方式,该方法可以减少分词错误带来的影响,也可以从一定程度上缩小中文词表规模。

3.3 实验参数设置

本文使用 PyTorch 框架在 NVIDIA V100 GPU 上进行实验。词向量维度大小为 512,编码器隐状态维度为 256,解码器隐状态维度为 512,批处理大小为 32,Dropout 丢弃旅为 0.5。本文使用 Adam 优化器训练,初始学习率为 1e-3,在模型训练阶段根据模型性能采取学习率衰减策略,衰减比例为 0.5。

3.4 基线模型

NLP_ONE 是 NLPCC2017 单文档文摘任务中获得第一名成绩的模型;PGN 利用指针网络判断每一解码时刻是否从原文中选择词语,利用覆盖度损失减少输出文摘中的重复词语出现频率;SEASS 应用选择门机制过滤编码器输出信息。此外,本文还采用 LEAD 和 ORACLE 两个具有代表性的启发式策略作为以上 3 种数据集的基线模型。

3.5 实验结果

本文用 ROUGE-1、ROUGE-2、ROUGE-L 的 F1 指标作为评价指标。本模型在 TTNews 数据集上的实验结果如表 1 所示。

表 1 TTNews 数据集实验结果

Tab. 1 Evaluation results on TTNews data set

方法	R-1	R-2	R-L
LEAD	33.3	20.0	27.0
ORACLE	46.8	32.6	40.1
NLP_ONE	57.4	46.3	53.5
PGN	59.4	48.0	55.2
SEASS	60.2	48.9	55.6
Seq2Seq (Our impl.)	58.8	47.6	54.3
本文模型	**60.7**	**49.8**	**56.2**

实验结果表明,本文提出的模型在各项 Rouge 指标上均超过所有基线模型,以 ROUGE-2 F1 值为例,本模型在 TTNews 取得 2.2 个百分点的提升。实验结果中 LEAD 和 ORACLE 方法的表现较差,这表明数据集中的摘要与原文句子重合度较低,是高度人工生成的,利用该数据集验证本模型在生成式文摘任务的效果具有一定说服力。

3.6 消融分析

为分析本模型各组件对最终性能的贡献,本节以 TTNews 数据集为例,采用消融分析,逐步验证各组件对模型最终结果的影响程度。本模型的消融分析结果如表 2 所示。

表 2 消融分析实验结果

Tab. 2 Results of ablation studies

方法	R-1	R-2	R-L
本文模型	**60.7**	**49.8**	**56.2**
- 选择门控单元	59.3	48.9	55.1
- 全局编码信息	58.8	47.6	54.3

通过分析消融分析结果可以发现,每移除一个模型组件,模型的 3 个指标均以会不同程度下降。当本模型仅使用融合编码信息的选择门控单元时,性能优于同样使用选择门控单元的 SEASS 模型,这证明全局编码信息在编码端发挥重要作用;当移除选择门控单元后,此时模型仅应用全局编码信息计算编码端注意力,ROUGE 指标略低于 SEASS 模型,但仍优于 PGN 模型。当彻底不使用全局编码信息时,本文的模型指标略低于 SEASS 和 PGN 模型,但仍然优于 NLP_ONE 模型。

4 结束语

本文论证目前主流的基于注意力机制的序列到序列模型存在的不足,仅依靠注意力机制只能选择重要的词语对应的隐状态信息,但是无法对隐状态各个维度信息进一步把控。如果模型将存在噪声的语义信息全部接收,会导致模型输出的文摘无法充分关注到原文的核心内容。为解决上述问题,本文提出应用全局编码信息的模型方案,将全局编码信息应用于编码端的信息重构、选择门控单元的计算、解码端注意力上下文的计算,以及全连接网络中特征向量的融合。经实验证明,本文提出的模型在 TTNews 数据集上取得最好的结果,通模型不同组件的消融分析,证明本文提出的优化方案对生成式文摘模型的性能起积极作用。

参考文献

[1] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate. arXiv preprint, 2367 arXiv: 1409.0473, 2014.

[2] ZHOU Q, YANG N, WEI F, et al. Selective encoding for abstractive sentence summarization[J].arXiv preprint arXiv:1704.07073, 2017.

[3] ZAJIC D, DORR B, SCHWARTZ R. Bbn/umd at duc-2004: Topiary[C]//Proceedings of the HLT-NAACL 2004 Document Understanding Workshop, Boston. 2004: 112-119.

[4] KNIGHT K, MARCU D. Summarization beyond sentence extraction: A probabilistic approach to sentence compression[J]. Artificial Intelligence, 2002, 139(1): 91-107.

(下转第 108 页)