基于词性软模板注意力机制的短文本自动摘要方法

张亚飞[1,2] 左一溪[1,2] 余正涛[1,2] 郭军军[1,2] 高盛祥[1,2]

摘 要 在短文本摘要任务中,带有直观主谓宾结构的摘要句语义完整性较强,但词性组合对该结构具有约束作用.为此文中提出基于词性软模板注意力机制的短文本自动摘要方法.首先对文本进行词性标记,将标记的词性序列视为文本的词性软模板,指导方法构造摘要句的结构规范,在编码端实现词性软模板的表征.再引入词性软模板注意力机制,增强对文中核心词性(如名词、动词等)的关注.最后在解码端联合词性软模板注意力与传统注意力,产生摘要句.在短文本摘要数据集上的实验验证文中方法的有效性.

关键词 文本摘要,词性软模板,词性注意力机制,自然语言生成

引用格式 张亚飞,左一溪,余正涛,郭军军,高盛祥.基于词性软模板注意力机制的短文本自动摘要方法.模式识别与人工智能,2020,33(6):551-558.

DOI 10.16451/j.cnki.issn1003-6059.202006008 **中图法分类号** TP 391

Automatic Short Text Summarization Based on Part-of-Speech Soft Template Attention Mechanism

ZHANG Yafei[1,2], ZUO Yixi[1,2], YU Zhengtao[1,2], GUO Junjun[1,2], GAO Shengxiang[1,2]

ABSTRACT Semantic integrity of the summary with intuitive subject-predicate-object structure is strong in the short text summarization task. However, part-of-speech combinations impose constraints on the structure. Aiming at this problem, an automatic short text summarization method based on part-of-speech soft template attention mechanism is proposed. Firstly, text is tagged with part-of-speech tags, and the tagged part-of-speech sequence is regarded as part-of-speech soft template of the text to guide the method to construct the structural specifications of a summary. Part-of-speech soft template is characterized at the encoder. Then, the part-of-speech soft template attention mechanism is introduced to enhance the attention of core part-of-speech information in text, such as nouns and verbs. Finally, the part-of-speech soft template attention and traditional attention are combined to generate a summary at the decoder. Experimental results verify the effectiveness of the proposed method on short text summarization datasets.

Key Words Text Summarization, Part-of-Speech Soft Template, Part-of-Speech Attention Mechanism, Natural Language Generation

收稿日期: 2020-03-14; **录用日期:** 2020-06-12

Manuscript received March 14, 2020;
accepted June 12, 2020

国家重点研发计划(No.2018YFC0830105,2018YFC0830101,2018YFC0830100)、**国家自然科学基金项目**(No.61762056,61866020,61761026,61972186)、**云南省自然科学基金项目**(No.2018FB104)**资助**

Supported by National Key Research and Development Program of China(No.2018YFC0830105,2018YFC0830101,2018YFC0830100), National Natural Science Foundation of China(No.61762056,61866020,61761026,61972186), Natural Science Foundation of Yunnan Province(No.2018FB104)

本文责任编委 马少平

Recommended by Associate Editor MA Shaoping

1.**昆明理工大学 信息工程与自动化学院 昆明** 650504

2.**昆明理工大学 云南省人工智能重点实验室 昆明** 650504

1.Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504

2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming 650504

Citation ZHANG Y F, ZUO Y X, YU Z T, GUO J J, GAO S X. Automatic Short Text Summarization Based on Part-of-Speech Soft Template Attention Mechanism. Pattern Recognition and Artificial Intelligence, 2020, 33(6): 551-558.

文本摘要任务旨在通过提炼原文核心信息,生成一段高度概括原文内容的摘要句,帮助人们提高日常浏览和获取知识的效率.

文本摘要从实现方法上主要分为如下两种: 1)抽取式文摘.直接从原文本中选出若干重要句子,拼接组合这些句子,形成摘要句.2)生成式文摘.主要利用序列到序列的深度学习模型进行文本语义理解,再通过语言生成模型、信息压缩等处理手段生成最终的摘要句[1],生成的摘要可读性、连贯性更强.

近年来,带有注意力机制的序列到序列模型(Sequence to Sequence, seq2seq)在生成式摘要任务上受到广泛关注[2],注意力用于提取原文重要的上下文信息[3].序列模型能将短文本进行较好的语义编码,所以目前相关研究主要集中在短文本摘要生成任务上.该任务通常以文档首句作为模型输入,以一个较短句子作为输出.

在短文本摘要任务中,基于模板的方法为传统方法[4].近年来神经网络模型应用在该任务上,取得一定效果,循环神经网络(Recurrent Neural Network, RNN)为常用的基础模型[5-6].Hochreiter等[7]在RNN上利用长短期记忆网络(Long Short Term Memory, LSTM)构造编码器与解码器. Nallapati等[1]尝试将外部知识融入模型,提升模型性能.

为了解决生成文摘中未登录词问题,Gu等[8]使用拷贝机制直接复制原文本中某些重要词汇,并输出至生成的摘要序列中.See等[9]使用覆盖机制,避免注意力多次在相同位置赋予高权值,解决文摘句中可能出现重复单词的问题.Cao等[10]将主题模型整合至自动摘要模型中,用于提高摘要对原文内容的吻合度.Li等[11]利用提出的潜在结构模块,将句子的潜在结构信息融入生成式摘要模型中,提高生成句子的质量.

为了解决生成的摘要可能会出现重复或无语义的问题,Lin等[12]提出基于原文本上下文信息的全局编码框架,用于过滤编码器到解码器的信息流. Gao等[13]提出读者评论感知的方法,解决序列到序列模型对文档主题建模不佳的问题.Cao等[14]受传统模板方法构建摘要的启发,将软模板思想融入RNN,利用软模板指导摘要生成.

从上述研究工作可看出,基于深度学习的文本摘要任务取得较优性能,但大部分模型的注意力仅限于考虑整个原文内容,忽略文本背后重要结构信息的影响,而词性组合对句子结构具有约束作用.未来研究应增强模型对句子结构的学习,有效结合词性信息与注意力机制,使模型学习合理的词性组合方式,有利于文摘系统生成结构清晰、语义完整的摘要.

鉴于此种情况,本文改进当前通用的基于编码器-解码器的生成式自动文摘模型,提出基于词性软模板注意力机制(Part-of-Speech Soft Template Attention Mechanism, POSTemp_Att)的短文本自动摘要方法.将原文对应的词性软模板(Part-of-Speech Soft Template, POSTemp)进行编码,得到词性表征向量.提出词性软模板注意力机制,指导模型学习合理的词性组合方式,辅助摘要生成.

1 基于词性软模板注意力机制的短文本自动摘要方法

本文改进现有带有注意力机制的编码器-解码器模型,提出基于词性软模板注意力机制的短文本自动摘要方法.方法整体结构如图1所示.

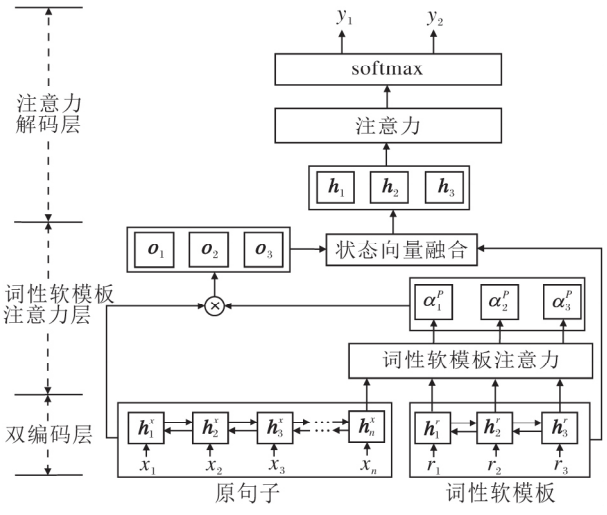


图1 本文方法结构图

Fig.1 Structure of the proposed method

首先将原句子及其对应的词性软模板分别在双向LSTM中进行编码.再引入词性软模板注意力机制,将模型学习的核心词性信息融入原文本隐层向

量中,得到融合词性的原文上下文语义向量.最后将融合词性的原文上下文语义向量与词性软模板隐藏层向量进行融合,利用带有传统注意力机制的单向 LSTM 解码文摘.

1.1　句子语义和词性软模板双重编码

在对文本的词向量进行语义编码时,LSTM 可以较好地学习到词间依赖关系,但 LSTM 单向处理的方式无法综合上下文语义信息,影响预测效果.双向 LSTM 对 LSTM 进行改进,使其具有前向和后向两个状态,充分掌握文本上下文语义信息.因此使用双向 LSTM 对文本句进行语义编码.

本文把原文对应的词性标注序列称为词性软模板,完整的词性组合序列能为模型学习核心的词性组合方式提供参考,这种参照仿写的方式更接近人类学习写作的过程.同样使用双向 LSTM 对原文的词性软模板进行编码.

1.2　基于词性软模板注意力的解码

在生成摘要的过程中,注意力机制可帮助模型选出原文关键信息,但传统注意力机制更多关注原文本身语义,关注点较单一.因此本文在序列到序列模型中加入词性软模板注意力机制,引导模型学习文本更深层次的结构特点.注意力机制的关键在于计算注意力系数,词性软模板注意力系数

$$\alpha_i^p = \frac{\exp(\eta_0(\boldsymbol{h}_n^x, \boldsymbol{h}_i^r))}{\sum_{j=1}^{n}\exp(\eta_0(\boldsymbol{h}_n^x, \boldsymbol{h}_j^r))},$$

其中,x 表示原文本,r 表示词性软模板,利用涵盖原文全局信息的隐向量 $\boldsymbol{h}_n^x$ 和每个时间步 i 下的词性隐向量 $\boldsymbol{h}_i^r$ 计算词性软模板注意力系数,η_0 表示多层感知器,使用 η_0 作为激活函数.

进一步,将计算的词性软模板注意力系数融入原文隐向量中,得到融合词性信息的原文上下文语义向量:

$$\boldsymbol{o}_i = \sum_{j=1}^{n}\alpha_i^p \boldsymbol{h}_j^x.$$

获得融合词性的原文上下文语义向量后,继续将该向量与词性向量进行融合,最终在带有传统注意力机制的解码端对输出向量进行解码,得到输出词的条件概率.状态向量融合后,用于解码的向量不仅包含原文本的语义信息,还包括模型学习得到的关键词性信息.综合这两种信息,可输出句子结构明确的优质摘要.

在融合带有词性信息的原文上下文语义向量 $\boldsymbol{o}_i$ 和词性隐向量 $\boldsymbol{h}_i^r$ 过程中,分别采用点乘和线性相加的方式.使用点乘的方式进行融合: $\boldsymbol{h}_i = \boldsymbol{o}_i \otimes \boldsymbol{h}_i^r$.使用线性相加的方式进行融合: $\boldsymbol{h}_i = \boldsymbol{o}_i \oplus \boldsymbol{h}_i^r$.

向量融合后得到解码端的输入 $\boldsymbol{h}_i$,$\boldsymbol{h}_i$ 表示融合词和词性的高层语义表达.模型解码过程如下:

$$P_{\text{vocab}} = \text{softmax}(g([\boldsymbol{c}_t; \boldsymbol{s}_t])),$$

$$\boldsymbol{s}_t = LSTM(y_{t-1}, \boldsymbol{s}_{t-1}, C_{t-1}),\ \boldsymbol{c}_t = \sum_{i=1}^{n}\alpha_{t,i}\boldsymbol{h}_i,$$

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{n}\exp(e_{t,j})},\ e_{t,i} = \boldsymbol{s}_{t-1}^{\mathrm{T}}\boldsymbol{W}_a\boldsymbol{h}_i,$$

其中,$\boldsymbol{s}_t$ 表示解码器的隐状态,$\boldsymbol{W}_a$ 表示神经网络模型学习的权重矩阵,通过 $\boldsymbol{s}_t$、$\boldsymbol{h}_i$ 计算得到 i 时刻的注意力系数值 $\alpha_{t,i}$,C 表示 LSTM 中的细胞状态,$g(\cdot)$ 表示非线性函数.利用上下文向量 $\boldsymbol{c}_t$ 解码,经过序列到序列模型的处理,得到预测单词 y.

2　实验及结果分析

本文在公共数据集 Gigaword[15]、LCSTS[16] 和自己收集的司法领域数据集上进行实验.3 个数据集的详细信息如表 1 所示.

表 1　实验数据集统计信息

Table 1　Statistics of experimental datasets

名称		文本数量	总文本数量	语言
Gigaword	训练集	3803957	3995559	英文
	验证集	189651		
	测试集	1951		
LCSTS	训练集	2400591	2410001	中文
	验证集	8685		
	测试集	725		
司法领域	训练集	121368	127368	中文
	验证集	5000		
	测试集	1000		

在 LCSTS 数据集上,使用第 I 部分作为训练集,将第 II 部分和第 III 部分中人工评分 3 分以上的子集作为验证集和测试集.

在收集司法领域数据集时,从新浪微博的新闻号上爬取约 468 000 条包括微博正文及其标题的新闻文本,课题组人员人工筛选司法领域文本进行汇总,将正文对应的标题作为参考摘要,过滤正文长度小于 10 的文本,删除文本中含有的 emoji 等特殊符号.经统计,微博正文中训练集的平均长度为 129.3,验证集的平均长度为 120.0,测试集的平均长度为 117.7.参考摘要中训练集的平均长度为 23.7,验证集的平均长度为 23.5,测试集的平均长度为 23.0.

由于不同工具对中英语料处理各有优势,最终使用词性标注工具 NLTK 对 Gigaword 数据集上英文

语料进行词性标注,用 pyhanlp 对 LCSTS 数据集上和自己收集的司法领域数据集的中文语料进行分词与词性标注.

本文采用基于召回率统计的 ROUGE 评价方法[17]评估模型性能.ROUGE 主要用于计算模型产生的文摘与标准文摘之间的一元词、二元词及最长公共子串等的重叠率.通过分析 ROUGE 评测标准中的 ROUGE-1、ROUGE-2、ROUGE-L 的 F 值评价实验结果.

本文使用 PyTorch 深度学习框架编写模型,在 NVIDIA Tesla K40m GPU 上进行实验.原文本词典大小限制为 50 000.词嵌入向量和 LSTM 的隐藏层向量维度都为 512.考虑到原文对应词性的词典规模太小,将词性的词向量维度设为 30.编码端与解码端的 LSTM 都采用三层结构.在训练阶段,使用带默认参数的 Adam(Adaptive Moment Estimation)优化器[18],$\alpha = 0.001$,$\beta_1 = 0.9$,$\beta_2 = 0.999$,$\varepsilon = 1 \times 10^{-8}$,学习率在每轮的训练过程中折半进行衰减.批处理大小设置为 64,随机失活(Dropout)为 0.3.

在 Gigaword 数据集上进行实验时,选取生成式摘要任务的 7 个基线模型与本文模型进行对比.对比时直接使用如下模型在原论文中的实验结果.

1) ABS(Attention-Based Summarization)[5].使用带有注意力机制的卷积神经网络(Convolutional Neural Network,CNN)编码器和神经网络语言模型(Neural Network Language Model,NNLM)解码器生成摘要.

2) ABS+(ABS with Additional Features)[5].在 ABS 基础上加入文本的其它特征,用于强化模型学习.

3) Luong-NMT(Neural Machine Translation Model of Luong for Summarization)[6].基于 ABS+,使用条件循环神经网络进行解码.

4) Feats2s(Sequence to Sequence Model with Feature-Rich Encoder)[1].在基于 RNN 的 seq2seq 上融合文本其它特征,提升模型编码端的表达能力.

5) SEASS(Selective Encoding for Abstractive Sentence Summarization)[19].在序列到序列模型的编码端实现选择性门控网络,过滤无关信息.

6) FTSum(Fact Aware Neural Abstractive Summarization)[10].在基于 CNN 的 seq2seq 框架上融入主题模型,使用开放信息抽取和依存分析技术将原文中提取的事实描述内容加入摘要句中.

7) Re^{3}Sum(Retrieve,Rerank and Rewrite Summarization)[14].把已有的摘要句视为软模板,将软模板融入模型,指导摘要生成.

在 LCSTS 数据集上进行实验时,选取该语料论文中提供的 2 个基线模型与本文模型进行对比.对比时直接使用如下模型在原论文中的实验结果.

1) RNN[16].使用 RNN 作为编码器,将最终的隐状态输入解码器,解码期间不使用上下文.

2) RNN-context(RNN with context)[16].使用 RNN 作为编码器,将编码器的所有隐状态组合输入到解码器,解码过程中使用上下文信息.

为了分析本文的词性软模板注意力对模型解码的约束效果,将词性软模板注意力权重和解码阶段的传统注意力权重分别进行可视化呈现,两种注意力权重的可视化分别如图2和图3所示,图中颜色越深表示注意力值越大.

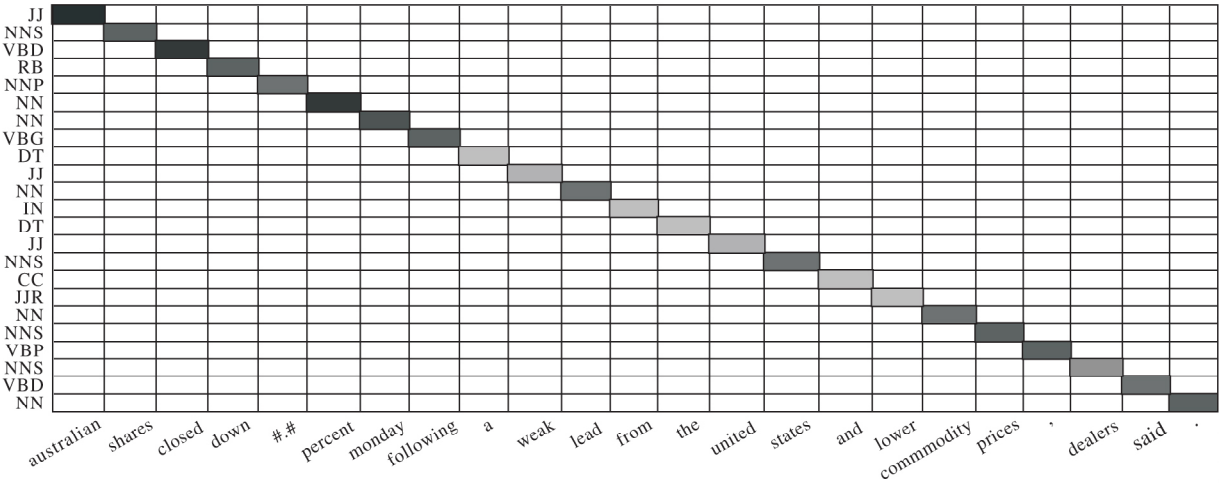


图2 词性软模板注意力权重可视化

Fig.2 Attention weight visualization of part-of-speech soft template

图 2 的行和列分别表示原句子及其对应的词性软模板.图 2 表明词性软模板注意力机制已引导模型对句中的词性进行不同注意力的分配,并且对名词和动词关注较多.

而句子的主谓宾结构大多由名词和动词组成,说明模型抓住句子结构的核心部分.图中注意力大多分配在句子的前半段,说明该注意力机制认为句子的前半部分重要程度更高.

图 3 的行和列分别表示原句子及其对应的参考摘要.结合图 2 和图 3 的结果可见,词性软模板注意力与解码阶段的注意力基本吻合,说明在摘要生成过程中模型增强对词性注意力较大的词的解码,词性软模板注意力对模型的解码具有一定约束力.这表明引入的词性软模板注意力机制是有效的.

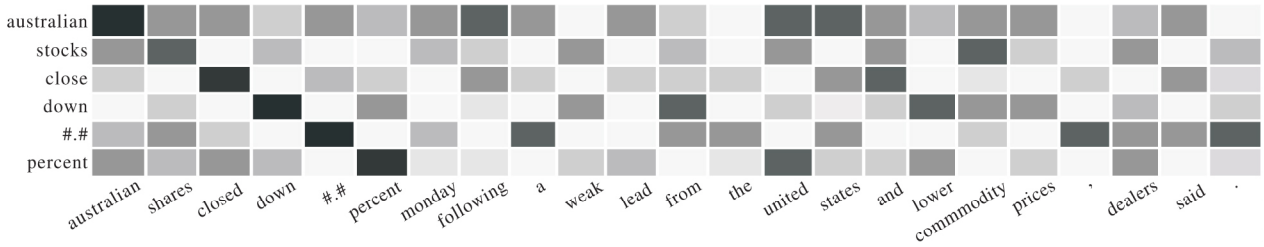


图 3　解码的注意力权重可视化

Fig.3　Attention weight visualization of decoder

在带有传统注意力机制的 RNN 模型 seq2seq(base)上对比各模型性能,实验结果如表 2 所示.

表 2　各模型在 Gigaword 数据集上的实验结果

Table 2　Experimental results of different models on Gigaword dataset

%

模型	ROUGE-1	ROUGE-2	ROUGE-L
ABS	29.55	11.32	26.42
ABS+	29.76	11.88	26.96
Luong-NMT	33.10	14.45	30.71
Feats2s	32.67	15.59	30.64
SEASS	36.15	17.54	33.63
FTSum	37.27	17.65	34.24
Re^3Sum	37.04	**19.03**	**34.46**
seq2seq(base)	33.30	16.21	30.08
POSTemp_Att	**37.39**	17.56	34.16

由表 2 可知,POSTemp_Att 在 ROUGE-1 分数上达到最优结果.相比 seq2seq(base),POSTemp_ Att 在 ROUGE-1、ROUGE-2、ROUGE-L 上分别提高 4.09%、1.35%、4.08%,说明本文模型是有效的.但 POSTemp_Att 的 ROUGE-2、ROUGE-L 分数略低于其它系统,说明该系统生成的摘要流畅程度有待加强,模型还需要进一步改进.

为了分析模型中不同组件对模型的贡献程度,在基础模型 seq2seq(base)上依次添加组件进行实验.首先在基础模型上加入词性软模板(POSTemp),解码前针对原文上下文向量与词性软模板上下文向量的融合默认采用线性相加的方式.在上一步基础上加入词性软模板注意力机制,解码前对向量的融合方式稍作改变,$POSTemp\_Att_*$、$POSTemp\_Att_+$ 分别表示经词性软模板注意力作用后,带有词性信息的原文上下文向量与词性软模板上下文向量在融合上采用点乘和线性相加的方式.

添加组件后的模型性能如表 3 所示.

表 3　具有不同组件模型的性能对比

Table 3　Performance comparison of models with different components

%

模型	ROUGE-1	ROUGE-2	ROUGE-L
seq2seq(base)	33.30	16.21	30.08
seq2seq(base) + POSTemp	34.91	16.67	32.27
seq2seq(base) + $POSTemp\_Att_*$	35.27	15.86	32.17
seq2seq(base) + $POSTemp\_Att_+$	37.39	17.56	34.16

对比表 3 中的实验结果发现,模型中不同组件的贡献有所不同.在基础模型上加入 POSTemp 后,相比 seq2seq(base),模型在 ROUGE-1、ROUGE-2、ROUGE-L 上分别提高 1.61%、0.46%、2.19%,说明词性软模板对辅助摘要的生成是有效的,它在模型生成摘要过程中发挥一定的参考作用.基础模型加入 $POSTemp\_Att_*$ 后,相比 POSTemp,模型在 ROUGE-1 上提高 0.36%,而在 ROUGE-2、ROUGE-L 上分别降低 0.

81%、0.1%.加入 POSTemp_Att+ 后,相比 POSTemp,模型在 ROUGE-1、ROUGE-2、ROUGE-L 上分别提高 2.48%、0.89%、1.89%.以上说明提出的词性软模板注意力机制是有效的,该机制可以帮助模型捕获合理的词性组合方式,生成结构更优的摘要句.在解码前向量的融合上,采用线性相加的方法优于点乘的方法.综合分析两种向量融合方式对模型效果的影响发现,可能点乘这种非线性方法会造成一定的语义混乱,而相加的方式属于线性方法,在本文模型中线性方法的表达能力优于非线性方法.

为了验证本文模型在中文语料上的有效性,在 LCSTS 数据集上进行实验.在 LCSTS 数据集的处理上,原文献已证实经字符级别分词处理后的实验效果优于基于词语级别的分词处理效果,因为字符分词生成的词典远小于词语分词生成的词典,后者更容易产生大量未登录词问题.由于本文模型需要对文本进行词性标注,所以只选择对语料进行词语级别的分词处理.

各模型实验结果如表 4 所示.

表 4 各模型在 LCSTS 数据集上的实验结果

Table 4 Experimental results of different models on LCSTS dataset

%

模型	ROUGE-1	ROUGE-2	ROUGE-L
RNN	17.7	8.5	15.8
RNN-context	26.8	16.1	24.1
seq2seq(base)_p	27.4	11.1	24.5
seq2seq(base)	30.2	13.6	27.3
seq2seq(base) + POSTemp	31.8	15.1	28.8
seq2seq(base) + POSTemp_Att*	32.1	15.5	29.1
seq2seq(base) + POSTemp_Att+	32.7	16.0	29.7

表 4 中 RNN 和 RNN-context 的评测值引自原论文中经词语级别分词处理后的实验结果.seq2seq(base)_p 和 seq2seq(base) 分别表示在保留和去除原语料中特殊字符及标点符号基础上的实验结果.实验结果表明,在大规模中文语料中过滤特殊字符和标点符号有助于提升生成摘要的质量.

由表 4 可知,在 seq2seq(base) 上加入 POSTemp 部分后,相比 seq2seq(base),模型在 ROUGE-1、ROUGE-2、ROUGE-L 上分别提高 1.6%、1.5%、1.5%.加入 POSTemp_Att+ 后,相比 seq2seq(base),模型在 ROUGE-1、ROUGE-2、ROUGE-L 上分别提高 2.5%、2.4%、2.4%,说明本文模型同样适用于中文语料.

为了验证本文模型在小规模中文语料上的有效性,在自收集的中文短文本数据集上进行实验,实验结果如表 5 所示.

表 5 各模型在司法领域数据集上的实验结果

Table 5 Experimental results of different models on judicial field dataset

模型	ROUGE-1	ROUGE-2	ROUGE-L
seq2seq(base)	39.19	21.05	35.77
seq2seq(base) + POSTemp	41.41	23.43	38.12
seq2seq(base) + POSTemp_Att*	41.77	24.00	38.56
seq2seq(base) + POSTemp_Att+	43.27	27.22	40.55

由表 5 可知,在 seq2seq(base) 上加入 POSTemp 部分后,相比 seq2seq(base),模型在 ROUGE-1、ROUGE-2、ROUGE-L 上分别提高 2.22%、2.38%、2.35%.加入 POSTemp_Att+ 后,相比 seq2seq(base),模型在 ROUGE-1、ROUGE-2、ROUGE-L 上分别提高 4.08%、6.17%、4.78%,说明本文模型同样适用于小规模中文语料.

综合对比 3 个数据集上的实验结果可知,语料规模大小和语料处理方式如文本分词与词性标注方法对模型效果具有一定影响.

从序列到序列模型中选取几个模型生成的摘要进行分析.

在 Gigaword 数据集上生成的摘要示例如表 6 所示.

表 6 各模型在 Gigaword 数据集上生成的摘要示例

Table 6 Examples of generative summaries of different models on Gigaword dataset

原文	india's children are getting increasingly overweight and unhealthy and the government is asking schools to ban junk food, officials said thursday.
参考摘要	indian government asks schools to ban junk food
Re3Sum	indian schools to ban junk food
seq2seq(base)	india asks schools to ban food
POSTemp_Att+	indian government urged to ban junk food

表 6 包含 3 个模型生成的英文摘要示例.POS-Temp_Att$_+$ 生成的摘要较好地抓住原文主旨,句子的主谓宾结构也较直观,并且使用原文中未出现的 urged to 搭配替换参考摘要中的 asks to,说明生成式方法输出的文摘在表达方式上更出彩.Re^{3}Sum 和 seq2seq(base) 生成的摘要未能完全抓住原文的主旨,甚至略微曲解原文的表述.

在 LCSTS 数据集上生成的摘要示例如表 7 所示.表 7 包含 4 个模型生成的中文摘要示例.RNN + Word 和 RNN-context + Word 生成的摘要部分可读性较差,seq2seq(base) 和 POSTemp_Att$_+$ 生成的摘要可读性较好,并且都较好地抓住原文主旨. POSTemp_ Att$_+$ 摘要句中“连续”和“蝉联”同时出现,虽有语病,但不影响阅读,模型生成的“蝉联”一词未出现在原文中,该词的生成使摘要句表述更出色.

表 7　各模型在 LCSTS 数据集上生成的摘要示例

Table 7　Examples of generative summaries of different models on LCSTS dataset

方法	9 月 3 日,总部位于日内瓦的世界经济论坛发布了《2014 - 2015 年全球竞争力报告》,瑞士连续六年位居榜首,成为全球最具竞争力的国家,新加坡和美国分列第二位和第三位.中国排名第 28 位,在金砖国家中排名最高.
参考摘要	全球竞争力排行榜中国居 28 位居金砖国家首位
RNN+Word	2014 年全球竞争力排名: 瑞士第一北京第第第第名单第第第名单第 68 位世界第第第名单第 68 位
RNN-context+Word	2014 - 2015 年全球竞争力报告: 瑞士连续 6 年居榜首中国居 28 位(首 /3——— 访榜首) 中国排名第 28 位
seq2seq(base)	全球竞争力报告中国排名第 28 位第 28 位
POSTemp_Att$_+$	全球竞争力报告瑞士连续六年蝉联中国第 28 位

在司法领域数据集上生成的摘要示例如表 8 所示.表 8 包含两个模型生成的中文摘要示例.两个模型生成的摘要都较好地抓住原文主旨,但 Seq2seq (base) 生成的文摘“境内”一词显得多余,而 POSTemp_ Att$_+$ 将“安徽境内”精简为“安徽”,并且将“油罐车追尾”重述为“两车相撞”,提升文摘句的可读性.

表 8　各模型在司法领域数据集上生成的摘要示例

Table 8　Examples of generative summaries of different models on judicial field dataset

方法	6 月 28 日凌晨 2 时许,S102 线安徽颍上县境内 100KM + 700M 处发生一起两货车追尾相撞事故,其中一辆油罐车汽油泄露起火,引燃车辆及路边民房.火灾已造成现场 2 人、民房内 6 人共 8 人死亡,4 人受伤.目前,该事故正在调查处理中.
参考摘要	安徽油罐车追尾引燃民房致 8 死 4 伤
seq2seq(base)	安徽境内两车相撞引燃民房已致 8 死 4 伤
POSTemp_Att$_+$	安徽两车相撞引燃民房致 8 死 4 伤

通过分析模型生成的文摘示例可以看出,本文模型在中英文语料上生成的文摘都抓住原文的主旨,语义都较完整.

3　结 束 语

本文针对当前通用的基于编码器-解码器的模型进行改进,提出基于词性软模板注意力机制的短文本自动摘要方法.将原文的词性软模板进行编码,得到词性表征向量,提出词性软模板注意力机制指导模型学习核心的词性信息.实验表明,本文模型在短文本摘要任务上取得较优结果.今后考虑将文本的其它语义结构信息融入模型,探索以不同的模板方式辅助短文本摘要的生成.

参 考 文 献

[1] NALLAPATI R, ZHOU B W, DOS SANTOS C N, *et al*. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Be-yond // Proc of the 20th SIGNLL Conference on Computational Na-tural Language Learning. Stroudsburg, USA: ACL, 2016: 280-290.

[2] BAHDANAU D, CHO K, BENGIO Y. Neural Machine Translation by Jointly Learning to Align and Translate [C/OL]. [2020-03-13]. https://arxiv.org/pdf/1409.0473.pdf.

[3] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention Is All You Need [C/OL]. [2020-03-13]. https://arxiv.org/pdf/1706.03762.pdf.

[4] ZHOU L, HOVY E. Template-Filtered Headline Summarization [C/

OL]. [2020-03-13]. https://www.aclweb.org/anthology/W04-1010.pdf.

[5] RUSH A M, CHOPRA S, WESTON J. A Neural Attention Model for Abstractive Sentence Summarization [C/OL]. [2020-03-13]. https://www.aclweb.org/anthology/W04-1010.pdf.

[6] CHOPRA S, AULI M, RUSH A M. Abstractive Sentence Summarization with Attentive Recurrent Neural Networks // Proc of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies. Stroudsburg, USA: ACL, 2016: 93-98.

[7] HOCHREITER S, SCHMIDHUBER J. Long Short-Term Memory. Neural Computation, 1997, 9(8): 1735-1780.

[8] GU J T, LU Z D, LI H, *et al*. Incorporating Copying Mechanism in Sequence-to-Sequence Learning // Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2016: 1631-1640.

[9] SEE A, LIU P J, MANNING C D. Get to the Point: Summarization with Pointer-Generator Networks // Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2017: 1073-1083.

[10] CAO Z Q, WEI F R, LI W J, *et al*. Faithful to the Original: Fact Aware Neural Abstractive Summarization [C/OL]. [2020-03-13]. https://arxiv.org/pdf/1711.04434.pdf.

[11] LI P J, LAM W, BING L D, *et al*. Deep Recurrent Generative Decoder for Abstractive Text Summarization // Proc of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACL, 2017: 2091-2100.

[12] LIN J Y, SUN X, MA S M, *et al*. Global Encoding for Abstractive Summarization // Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2018: 163-169.

[13] GAO S, CHEN X Y, LI P J, *et al*. Abstractive Text Summarization by Incorporating Reader Comments // Proc of the 33rd AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2019: 6399-6406.

[14] CAO Z Q, LI W J, LI S J, *et al*. Retrieve, Rerank and Rewrite: Soft Template Based Neural Summarization // Proc of the 56th Annual Meeting of the Association for Computational Linguistics(Long Papers). Stroudsburg, USA: ACL, 2018, I: 152-161.

[15] NAPOLES C, GORMLEY M, VAN DURME B. Annotated Gigaword // Proc of the Joint Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction. Stroudsburg, USA: ACL, 2012: 95-100.

[16] HU B T, CHEN Q C, ZHU F Z. LCSTS: A Large Scale Chinese Short Text Summarization Dataset // Proc of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: ACL, 2015: 1967-1972.

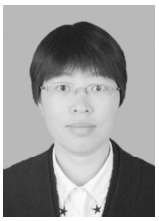
[17] LIN C Y, HOVY E. Automatic Evaluation of Summaries Using *N*-gram Co-occurrence Statistics // Proc of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Stroudsburg, USA: ACL, 2003: 71-78.

[18] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization [C/OL]. [2020-03-13]. https://arxiv.org/pdf/1412.6980.pdf.

[19] LUONG M T, PHAM H, MANNING C D. Effective Approaches to Attention-Based Neural Machine Translation [C/OL]. [2020-03-13]. https://arxiv.org/pdf/1508.04025.pdf.

[20] ZHOU Q Y, YANG N, WEI F R, *et al*. Selective Encoding for Abstractive Sentence Summarization // Proc of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2017: 1095-1104.

作者简介



张亚飞,博士,讲师,主要研究方向为自然语言处理、模式识别. E-mail: zyfeimail@163.com.

(**ZHANG Yafei**, Ph.D., lecturer. Her research interests include natural language processing and pattern recognition.)

左一溪,硕士研究生,主要研究方向为自然语言处理. E-mail: zuoyixi286@163.com.

(**ZUO Yixi**, master student. Her research interests include natural language proce-ssing.)

余正涛(通讯作者),博士,教授,主要研究方向为自然语言处理、信息检索、机器翻译. E-mail: ztyu@hotmail.com.

(**YU Zhengtao** (Corresponding author), Ph.D., professor. His research interests include natural language processing, information retrieval and machine translation.)

郭军军,博士,讲师,主要研究方向为自然语言处理. E-mail: guojjgb@163.com.

(**GUO Junjun**, Ph.D., lecturer. His research interests include natural language proce-ssing.)

高盛祥,博士,副教授,主要研究方向为自然语言处理. E-mail: gaoshengxiang.yn@foxmail.com.

(**GAO Shengxiang**, Ph.D., associate professor. Her research interests include natural language processing.)