Journal of Computer Applications
计算机应用,2021,41(1):127 - 132
ISSN 1001-9081
CODEN JYIIDU
2021-01-10
http://www.joca.cn

文章编号:1001-9081(2021)01-0127-06
DOI:10.11772/j.issn.1001-9081.2020060920

基于BERT-PGN模型的中文新闻文本自动摘要生成

谭金源[1],刁宇峰[1],祁瑞华[2],林鸿飞[1*]
(1.大连理工大学 计算机科学与技术学院,辽宁 大连 116024; 2.大连外国语大学 语言智能研究中心,辽宁 大连 116024)
(*通信作者电子邮箱hflin@dlut.edu.cn)

摘 要:针对文本自动摘要任务中生成式摘要模型对句子的上下文理解不够充分、生成内容重复的问题,基于BERT和指针生成网络(PGN),提出了一种面向中文新闻文本的生成式摘要模型——BERT-指针生成网络(BERT-PGN)。首先,利用BERT预训练语言模型结合多维语义特征获取词向量,从而得到更细粒度的文本上下文表示;然后,通过PGN模型,从词表或原文中抽取单词组成摘要;最后,结合coverage机制来减少重复内容的生成并获取最终的摘要结果。在2017年CCF国际自然语言处理与中文计算会议(NLPCC2017)单文档中文新闻摘要评测数据集上的实验结果表明,与PGN、伴随注意力机制的长短时记忆神经网络(LSTM-attention)等模型相比,结合多维语义特征的BERT-PGN模型对摘要原文的理解更加充分,生成的摘要内容更加丰富,全面且有效地减少重复、冗余内容的生成,Rouge-2和Rouge-4指标分别提升了1.5%和1.2%。

关键词:生成式摘要模型;预训练语言模型;多维语义特征;指针生成网络;coverage机制
中图分类号:TP391.1 **文献标志码**:A

Automatic summary generation of Chinese news text based on BERT-PGN model

TAN Jinyuan[1], DIAO Yufeng[1], QI Ruihua[2], LIN Hongfei[1*]
(1. *School of Computer Science and Technology, Dalian University of Technology, Dalian Liaoning* 116024, *China*;
2. *Language Intelligence Research Center, Dalian University of Foreign Languages, Dalian Liaoning* 116024, *China*)

Abstract: Aiming at the problem that the abstractive summarization model in text automatic summarization task does not fully understand the context of sentence and generates duplicate contents, based on BERT (Bidirectional Encoder Representations from Transformers) and Pointer Generator Network (PGN), an abstractive summarization model for Chinese news text was proposed, namely Bidirectional Encoder Representations from Transformers-Pointer Generator Network (BERT-PGN). Firstly, combining with multi-dimensional semantic features, the BERT pre-trained language model was used to obtain the word vectors, thereby obtaining a more fine-grained text context representation. Then, through PGN model, the words were extracted from the vocabulary or the original text to form a summary. Finally, the coverage mechanism was combined to reduce the generation of duplicate contents and obtain the final summarization result. Experimental results on the single document Chinese news summary evaluation dataset of the 2017 CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC2017) show that, compared with models such as PGN and Long Short-Term Memory with attention mechanism (LSTM-attention), the BERT-PGN model combined with multi-dimensional semantic features has a better understanding of the original text of the summary, has the generated summary content richer and more comprehensive with the generation of duplicate and redundant contents effectively reduced, and has Rouge-2 and Rouge-4 indicators increased by 1.5% and 1.2% respectively.

Key words: abstractive summarization model; pre-trained language model; multi-dimensional semantic feature; Pointer Generator Network (PGN); coverage mechanism

0 引言

随着近些年互联网产业的飞速发展,大量的新闻网站、新闻手机软件出现在日常生活中,越来越多的用户通过新闻网站、手机软件快速获取最新资讯。根据中国互联网络信息中心(China Internet Network Information Center, CNNIC)第42次发展统计报告,到2018年6月,中国的移动电话用户规模达到7.88亿,网民接入互联网的比例也在增加,通过手机达到98.3%[1]。网友人数增多、新闻媒体网络平台使用率不断提升,网友们使用今日头条等新闻媒体的频率也不断提升。

为了适应当下快节奏的生活,网友需要阅读最少的新闻字数,获取新闻文章的关键内容。网友们可以通过文本自动

收稿日期:2020-05-31;**修回日期**:2020-07-07;**录用日期**:2020-07-08。
基金项目:国家重点研发计划项目(2019YFC1200302);国家自然科学基金重点项目(61632011)。
作者简介:谭金源(1997—),男,辽宁大连人,硕士研究生,主要研究方向:自然语言处理; 刁宇峰(1987—),女,辽宁沈阳人,博士研究生,主要研究方向:自然语言处理; 祁瑞华(1974—),女,湖北襄樊人,教授,博士,主要研究方向:自然语言处理; 林鸿飞(1962—),男,辽宁大连人,教授,博士,主要研究方向:自然语言处理。

摘要技术，概括出新闻的主要内容，节省阅读时间，提升信息使用效率。因此，本文提出的面向新闻的文本自动摘要模型具有重要意义。

国内外学者针对文本自动摘要已经做了大量的研究。文本自动摘要是20世纪50年代出现的一种用计算机完成的文本摘要技术，帮助人们从信息海洋中解放，提高信息的使用效率[2]。自2001年美国国家标准技术研究所举办文档理解会议以来，文本自动摘要研究得到了越来越多的关注[3]。

本文受文献[4]启发，针对网友阅读理解新闻时需要花费大量时间的问题，基于BERT（Bidirectional Encoder Representations from Transformers）和指针生成网络（Pointer Generator Network，PGN），提出了一种面向中文新闻文本的自动摘要模型——BERT-指针生成网络（Bidirectional Encoder Representations from Transformers-Pointer Generator Network，BERT-PGN），能够有效节省时间，提高信息使用效率。该模型首先利用BERT预训练语言模型获取新闻文本的词向量，结合多维语义特征对新闻中的词所在的句子进行打分，其结果作为输入序列输入到指针生成网络中进行训练，得到新闻摘要的结果。

本文主要贡献如下。

1）本文提出了一种面向新闻文本进行自动摘要的模型——BERT-PGN，分为两个阶段实现：基于预训练模型及多维语义特征的词向量获取阶段以及基于指针生成网络模型的句子生成阶段。

2）实验结果表明，该模型在2017年CCF国际自然语言处理与中文计算会议（the 2017 CCF International Conference on Natural Language Processing and Chinese Computing，NLPCC2017）单文档中文新闻摘要评测数据集上取得了很好的效果，Rouge-2和Rouge-4指标分别提升1.5%和1.2%。

1 相关工作

自动文本摘要有两种主流方式，即抽取式摘要和生成式摘要[5]。在对文本进行语义挖掘的研究中，许多经典的分类、聚类算法被先后提出[6]。最早的摘要工作主要是利用基于词频和句子位置的基于统计的技术[7]。1958年，Luhn[8]提出了第一个自动文本摘要系统。近十几年来，随着机器学习（Machine Learning，ML）以及自然语言处理（Natural Language Processing，NLP）的快速发展，许多准确高效的文本摘要算法被提出[9]。互联网作为商业媒介快速发展，导致用户吸收了太多信息。为了解决这种信息过载，文本自动摘要起到了关键作用。文本自动摘要可以在屏蔽大量干扰文本的同时，让用户更加快捷地获取关键信息，适应当下快节奏的生活[10]。

抽取式摘要方法是将一篇文章分成小单元，然后将其中的一些作为这篇文章的摘要进行提取。Liu等[11]提出了一个抽取式文本摘要的对抗过程，使用生成对抗网络（Generative Adversarial Network，GAN）模型获得了具有竞争力的Rouge分数，该方法可以生成更多抽象、可读和多样化的文本摘要；Al-Sabahi等[12]使用分层结构的自注意力机制模型（Hierarchical Structured Self-Attentive Model，HSSAM），反映文档的层次结构，进而获得更好的特征表示，解决因占用内存过大模型无法充分建模等问题；Slamet等[13]提出了一种向量空间模型（Vector Space Model，VSM），利用VSM进行单词相似性测试，对文本自动摘要的结果进行测评，比较文本摘要实现的效果；Alguliyev等[14]发现，与传统文本自动摘要方法相比，基于聚类、优化和进化算法的文本自动摘要研究最近表现出了良好的效果。但抽取式摘要并未考虑文本的篇章结构信息，缺少对文本中关键字、词的理解，生成的摘要可读性、连续性较差。

生成式摘要方法是一种利用更先进自然语言处理算法的摘要方法，对文章中的句子进行转述、替换等生成文章摘要，而不使用其中任何现有的句子或短语。随着近些年深度学习的快速发展，越来越多的深度学习方法被利用到文本摘要中。Cho等[15]和Sutskever等[16]最早提出了由编码器和解码器构成的seq2seq（sequence-to-sequence）模型；Tan等[17]提出了基于图的注意力机制神经模型，在文本自动摘要的任务中取得了很好的效果；Siddiqui等[18]在谷歌大脑团队提出的序列到序列模型的基础上进行改进，使用局部注意力机制代替全局注意力机制，在解决生成重复的问题上取得了很好的效果；Celikyilmaz等[19]针对生成长文档的摘要，提出了一种基于编码器-解码器体系结构的深层通信代理算法；Khan等[20]提出了一种基于语义角色标记的框架，使用深度学习的方法从语义角色理解的角度实现多文档摘要任务；江跃华等[21]提出了一种基于seq2seq结构和注意力机制并融合了词汇特征的生成式摘要算法，能在摘要生成过程中利用词汇特征识别更多重点词汇内容，进一步提高摘要生成质量。

现阶段大多数的文本自动摘要方法主要是利用机器学习或深度学习模型自动提取特征，利用模型进行摘要句子的选取及压缩。但自动提取的特征和摘要文本会存在不充分、不贴近的情况，不能很好地刻画摘要文本。本文提出的BERT-PGN模型基于BERT预训练语言模型及多维语义特征，针对中文新闻文本，从更多维度进行特征抽取，深度刻画摘要文本，能够得到更贴近主题的摘要内容。

2 BERT-PGN模型

本文提出的BERT-PGN模型主要分成两个阶段实现，即基于预训练模型及多维语义的词向量获取阶段以及基于指针生成网络模型的句子生成阶段，如图1所示。该模型第一阶段利用预训练语言模型BERT获取新闻文章的词向量，同时利用多维语义特征对新闻中的句子进行打分，将二者进行简单拼接生成输入序列；第二阶段将得到的输入序列输入到指针生成网络模型中，使用coverage机制减少生成重复文字，同时保留生成新文字的能力，得到新闻摘要。

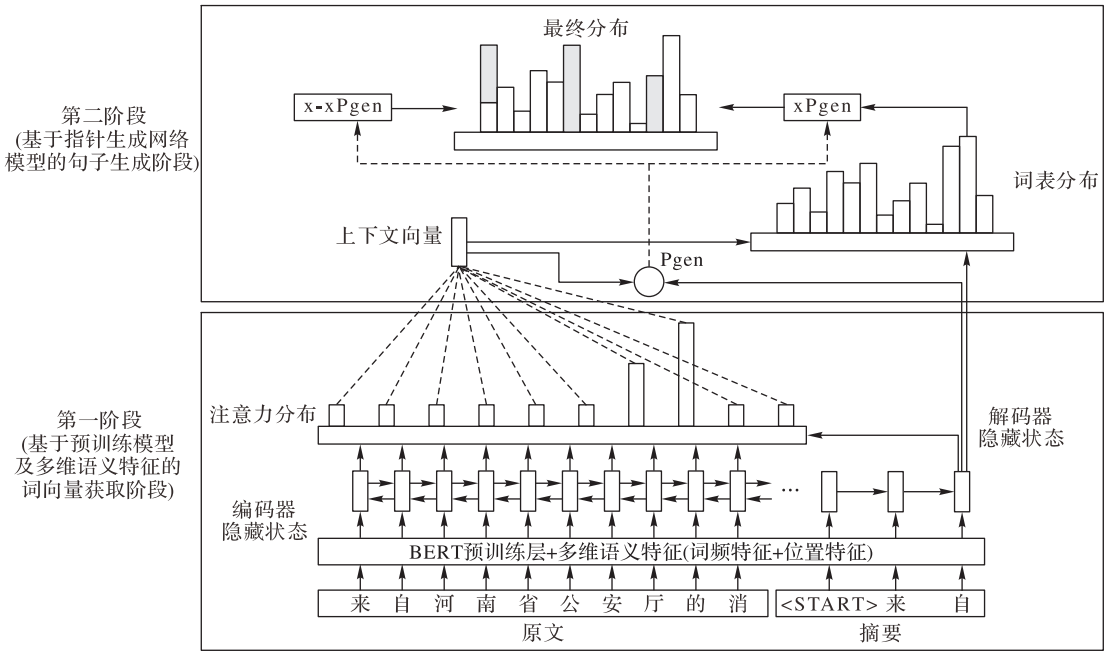

图1　BERT-PGN模型

Fig. 1　BERT-PGN model

2.1　基于预训练模型及多维语义特征的词向量获取阶段

2.1.1　BERT预训练语言模型

语言模型是自然语言处理领域一个比较重要的概念,利用语言模型对客观事实进行描述后,能够得到可以利用计算机处理的语言表示。语言模型用来计算任意语言序列$a_1,a_2,\cdots,a_n$出现的概率$p(a_1,a_2,\cdots,a_n)$,即:

$$p(a_1,a_2,\cdots,a_n)=\prod_{i=1}^{n}p(a_i|a_1,a_2,\cdots,a_{i-1}) \tag{1}$$

通过传统的神经网络语言模型获取的词向量是单一固定的,存在无法表示字的多义性等问题。预训练语言模型很好地解决了这一问题,能够结合字的上下文内容来表示字。BERT采用双向Transformer作为编码器进行特征抽取,能够获取到更多的上下文信息,极大程度地提升了语言模型抽取特征的能力。Transformer编码单元包含自注意力机制和前馈神经网络两部分。自注意力机制的输入部分是由来自同一个字的三个不同向量构成的,分别***Query***向量($\boldsymbol{Q}$),***Key***向量($\boldsymbol{K}$)和***Value***向量($\boldsymbol{V}$)。通过***Query***向量和***Key***向量相乘来表示输入部分字向量之间的相似度,记做$[\boldsymbol{QK}]^{\mathrm{T}}$,并通过$d_k$进行缩放,保证得到的结果大小适中。最后经过softmax进行归一化操作,得到概率分布,进而得到句子中所有词向量的权重求和表示。这样得到的词向量结合了上下文信息,表示更准确,计算方法如下:

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

BERT预训练模型使用了"MultiHead"模式,即使用了多个注意力机制获取句子的上下文语义信息,称为多头注意力机制。BERT预训练语言模型能够使词向量获取更多的上下文信息,更好地表示原文内容。

2.1.2　多维语义特征

针对中文新闻重点内容集中在新闻开头、关键词出现频率高等特点,本文引入了传统特征以及主题特征对中文新闻文本中的句子进行细粒度的描述,提升对文本中句子的上下文语义表述性能。

1)传统特征。

本文所选择的传统特征主要为句子层次的两种特征:句子中的词频以及在文章中的位置。

词频特征是反映新闻文章中最重要信息的一种统计特征,也是最简单、最直接的一种统计特征。新闻文章中出现词的词频可以利用式(3)进行计算:

$$tf(word)=\frac{word}{\sum_{j=1}^{n}word_j} \tag{3}$$

其中,$word_j$代表文章中第j个词出现的次数。

在本文中,选择文章中的句子作为最终的打分基本单位。句子是词的集合,如果句子包含的词语中,有在新闻文章中频繁出现的高频词,则认为这个句子在文章中更加重要。新闻文章中第i个句子的词频特征打分公式如下:

$$TF_i=\sum_{word\in sen_i}tf(word) \tag{4}$$

其中:TF_i表示第i个句子中包含的词的词频之和,sen_i代表第i个句子中包含的所有词。

位置特征同样是反映新闻文章中重要信息的一种统计特征。一篇新闻文章是由多个句子组成的,句子所在的位置不同,其代表的重要性也不同,例如文章中的第一个句子大多是新闻文章中最重要的一句话。新闻文章中第i个句子的位置特征打分公式如下:

$$Pos_i=\frac{n-p_i+1}{n} \tag{5}$$

其中:Pos_i代表第i个句子的位置得分,p_i代表第i个句子在新闻文章中的位置,n代表文章中的句子总个数。

2)主题特征。

本文选取的主题特征也可表述为标题特征。新闻文章中的标题具有很高的参考价值,很大程度上可以代表文章中的主题。因此,如果文章中的句子与新闻文章的标题有较高的相似度,那么这个句子更容易被选择为文章摘要中的句子。本文使用余弦相似度计算新闻文章中第i个句子的主题特征得分,打分公式如下:

$$Sim_i = \frac{\boldsymbol{s} \cdot \boldsymbol{t}}{\sqrt{\sum_{j=1}^{n} s_j^2} \cdot \sqrt{\sum_{j=1}^{n} t_j^2}} \tag{6}$$

其中：Sim_i表示第i个句子与新闻文章标题的相似度，$\boldsymbol{s}$和$\boldsymbol{t}$分别代表标题和新闻文章中句子的向量化表示。

2.2 基于指针生成网络模型的句子生成阶段

指针生成网络模型结合了指针网络(Pointer Network，PN)和基于注意力机制的序列到序列模型，允许通过指针直接指向生成的单词，也可以从固定的词汇表中生成单词。文本中的文字w_i依次传入BERT-多维语义特征编码器、双向长短时记忆神经网络(Bidirectional Long Short-Term Memory，Bi-LSTM)编码器，生成隐层状态序列$\boldsymbol{h}_i$。在t时刻，长短时记忆(Long Short-Term Memory，LSTM)神经网络解码器接收上一时刻生成的词向量，得到解码状态序列$\boldsymbol{s}_t$。

注意力分布$\boldsymbol{a}^t$用来确定t时刻输出序列字符时，输入序列中需要关注的字符。计算公式如下：

$$e_i^t = \boldsymbol{v}^{\mathrm{T}} \tanh(\boldsymbol{W}_h \boldsymbol{h}_i + \boldsymbol{W}_s \boldsymbol{s}_t + \boldsymbol{b}_{\mathrm{attn}}) \tag{7}$$

$$a_i^t = \mathrm{softmax}(e_i^t) \tag{8}$$

其中，$\boldsymbol{v}$、$\boldsymbol{W}_h$、$\boldsymbol{W}_s$、$\boldsymbol{b}_{\mathrm{attn}}$是通过训练得到的参数。利用注意力分布对编码器隐层状态加权平均，生成上下文向量$\boldsymbol{h}_t^*$。

$$\boldsymbol{h}_t^* = \sum_i a_i^t \boldsymbol{h}_i \tag{9}$$

将上下文向量$\boldsymbol{h}_t^*$与解码状态序列$\boldsymbol{s}_t$串联，通过两个线性映射，生成当前预测在词典上的分布P_{vocab}，计算公式如下：

$$P_{\mathrm{vocab}} = \mathrm{softmax}(\boldsymbol{V}'(\boldsymbol{V}[\boldsymbol{s}_t, \boldsymbol{h}_t^*] + \boldsymbol{b}) + \boldsymbol{b}') \tag{10}$$

其中，$\boldsymbol{V}'$、$\boldsymbol{V}$、$\boldsymbol{b}$、$\boldsymbol{b}'$是通过训练得到的参数。

模型利用生成概率P_{gen}来确定复制单词还是生成单词，计算公式如下：

$$P_{\mathrm{gen}} = \sigma(\boldsymbol{w}_h^{\mathrm{T}} * \boldsymbol{h}_i^* + \boldsymbol{w}_s^{\mathrm{T}} \boldsymbol{s}_t + \boldsymbol{w}_x^{\mathrm{T}} \boldsymbol{x}_t + b_{\mathrm{ptr}}) \tag{11}$$

其中，$\boldsymbol{w}_h$、$\boldsymbol{w}_s$、$\boldsymbol{w}_x$、b_{ptr}是通过训练得到的参数，σ是sigmoid函数，$\boldsymbol{x}_t$是解码输入序列。将$\boldsymbol{a}^t$作为模型输出，得到生成单词w的概率分布：

$$P(w) = P_{\mathrm{gen}} P_{\mathrm{vocab}}(w) + (1 - P_{\mathrm{gen}}) \sum_{i: w_i = w} a_i^t \tag{12}$$

为了解决生成词语重复的问题，本文引入了coverage机制。通过coverage机制对指针生成网络模型进行改进，能够有效减少生成摘要中的重复。引入coverage向量$\boldsymbol{c}^t$跟踪已经生成的单词，并对已经生成的单词施加一定的惩罚，尽量减少生成重复。coverage向量$\boldsymbol{c}^t$计算方式如下：

$$\boldsymbol{c}^t = \sum_{t'=0}^{t-1} \boldsymbol{a}^{t'} \tag{13}$$

通俗来说，$\boldsymbol{c}^t$表示目前为止单词从注意力机制中获得的覆盖程度。使用coverage向量$\boldsymbol{c}^t$影响注意力分布，重新得到注意力分布$\boldsymbol{a}^t$，计算公式如下：

$$e_i^t = \boldsymbol{v}^{\mathrm{T}} \tanh(\boldsymbol{W}_h \boldsymbol{h}_i + \boldsymbol{W}_s \boldsymbol{s}_t + \boldsymbol{W}_c c_i^t + \boldsymbol{b}_{\mathrm{attn}}) \tag{14}$$

其中$\boldsymbol{W}_c$是通过训练得到的参数。

3 实验与分析

3.1 实验数据

本文的实验部分使用的数据是由2017年CCF国际自然语言处理与中文计算会议(NLPCC2017)提供，来自于NLPCC2017中文单文档新闻摘要评测数据集，包含训练集新闻文本49 500篇，测试集新闻文本500篇。该任务中要求生成的摘要长度不超过60个字符。

3.2 评价指标

Rouge是文本自动摘要领域摘要评价技术的通用指标之一，通过统计模型生成的摘要与人工摘要之间重叠的基本单元，评判模型生成摘要的质量。本文参考NLPCC2017中文单文档新闻摘要评测任务，使用Rouge-2、Rouge-4和Rouge-SU4作为评价指标，对摘要结果进行评价。

3.3 对比实验

本文实验部分选取8种基本模型：NLPCC2017单文档新闻摘要评测任务结果较好团队(ccnuSYS、LEAD、NLP@WUST、NLP_ONE)提出的模型[22]、PGN(without coverage mechanism)[23]、PGN[23]、主题关键词信息融合模型[24]以及BERT-PGN(without semantic features)。对人工提取的主题特征、传统特征进行特征的有效性验证，验证本文提出方法的有效性。

1)ccnuSYS[22]：使用基于注意力机制的LSTM编码器-解码器结构模型生成摘要。

2)LEAD[22]：从原文选取前60个字作为文本摘要。

3)NLP@WUST[22]：使用特征工程的方法进行句子抽取，并利用句子压缩算法对抽取的句子进行压缩。

4)NLP_ONE[22]：NLPCC2017单文档新闻摘要评测任务第一名的算法，包含输入、输出序列的注意力机制。

5)PGN(without coverage mechanism)[23]：ACL2017中提出的一种生成模型，使用指针网络和基于注意力机制的序列到序列模型生成摘要，不使用coverage机制。

6)PGN(coverage mechanism)[23]：改进的指针生成网络模型，利用coverage机制解决生成重复词和未登录词的问题。

7)主题关键词融合模型[24]：一种结合主题关键词信息的多注意力机制模型。

8)BERT-PGN(without semantic features)：本文提出的一种基于BERT和指针生成网络的模型，利用coverage机制减少生成重复内容。

9)BERT-PGN(semantic features)：在BERT-PGN(without semantic features)模型上进行优化得到的模型，结合多维语义特征获取细粒度的文本上下文表示。

3.4 实验环境及参数设置

本文实验使用单个GTX-1080Ti(GPU)进行训练。本实验获取文本词向量使用BERT-base预训练模型。BERT-base模型共12层，隐层768维。设置最大序列长度为128，*train_batch_size*为16，*learning_rate*为5E-5。

指针生成网络模型设置*batch_size*为8，隐层256维，设置字典大小为50k。训练过程共进行700k次迭代，训练总时长约为7 d5 h(合计173 h)。

3.5 实验结果与分析

3.5.1 总体摘要结果对比实验

本文重新运行了部分baseline模型，将获取的结果与本文提出的模型结果做对比，实验结果如表1。

表1 总体摘要结果对比

Tab. 1 Results comparison of overall summarization

模型	Rouge-2	Rouge-4	Rouge-SU4
ccnuSYS*	14.98	7.66	13.89
LEAD*	20.88	11.69	19.14
NLP@WUST*	22.53	10.39	20.81
NLP_ONE*	22.87	12.78	21.18
PGN(without coverage mechanism)	23.15	12.33	21.40
PGN(coverage mechanism)	23.57	12.57	21.56
主题关键词融合模型	24.02	12.89	22.35
BERT-PGN(without semantic features)	24.32	13.09	22.68
BERT-PGN(semantic features)	**25.53**	**14.06**	**23.74**

注：带*方法的结果引自文献[22]。

从表1可以看出,本文提出的模型性能相较于PGN、NLP_ONE等模型有了显著的提升,在Rouge-2、Rouge-4以及Rouge-SU4的评价指标中有着明显的优势,Rouge指标提升了1.2~1.5个百分点。

由BERT-PGN(semantic features)模型与PGN、BERT-PGN(without semantic features)模型进行对比,可以看出使用BERT预训练模型并结合有效的多维人工特征,能够显著提升模型效果。使用BERT预训练模型并结合人工抽取的特征得到的句子上下文表示,对文本中句子的语义理解更加深刻、准确,在文本自动摘要任务中能够有效提升性能。

根据表2不同模型生成摘要的内容可以发现,本文提出的BERT-PGN模型相较于其他模型,在中文新闻文本的自动摘要任务中生成的摘要内容更丰富、更全面、更贴近标准摘要,说明该模型对全文的理解更加充分,能够结合文中句子的上下文充分理解句子、词语的含义,对文中的句子、词语进行更细致的刻画。

表2　摘要结果示例

Tab. 2　Summarization result examples

模型	摘要内容
标准摘要	河南出台举报暴恐线索奖励办法,最高奖5万元;企业商户未及时向公安机关报告,可依法追究责任
LEAD	来自河南省公安厅的消息,为做到预知预警,防患未然,有效维护公共安全,该厅制定出台《举报暴力恐怖违法犯罪线索奖励办法》
NLP@WUST	来自河南省公安厅的消息,该厅制定出台《举报暴力恐怖违法犯罪线索奖励办法》,鼓励公民和组织积极举报暴恐犯罪线索
主题关键词融合模型	河南省公安厅出台《举报暴力恐怖违法犯罪线索奖励办法》,有关企业、商户、从业人员发现违法犯罪可疑人员及时向公安机关报告
BERT-PGN	**河南省公安厅制定出台《举报暴力恐怖违法犯罪线索奖励办法》,有关企业、商户、从业人员未及时向公安机关报告,依法追究责任**

3.5.2　多维语义特征对比实验

多维特征选取的部分,本文针对新闻文本"主要内容集中在开头部分"的特点,选取传统特征、主题特征中的词频特征、位置特征以及标题特征,分别表示为TF、Pos以及Main。

由表3可以看出,同一模型结合人工提取的词频特征和位置特征效果最好,Rouge-2指标最多提升了1.2个百分点,Rouge-4指标最多提升了1.0个百分点。

表3　特征组合结果对比　　单位:%

Tab. 3　Feature combination result comparison　　unit:%

特征选择	Rouge-2	Rouge-4	Rouge-SU4
None	24.32	13.09	22.68
Main	24.74	13.47	22.85
Pos	25.04	13.81	23.23
Pos+Main	25.15	13.86	23.30
TF	25.35	13.93	23.46
TF+Main	25.37	13.92	23.51
TF+Pos+Main	25.44	13.96	23.58
TF+Pos	**25.53**	**14.06**	**23.74**

本文选取的主题特征Main能够在一定程度上提升模型的Rouge指标。从Pos和Pos+Main、TF和TF+Main的特征组合结果对比可以得知,主题特征结合词频特征时提升明显,结合位置特征时基本没有提升。句子在新闻中的位置靠前时,与标题的相似度也更高,说明两种人工特征在衡量句子在新闻中的重要性时起到了相似的作用。通过对比TF+Main和TF+Pos两种特征组合的结果可以得知,词频信息结合位置信息相较于结合主题信息效果更好,能够充分表达句子在新闻文章中的重要性。因此,本文选择使用词频特征以及位置特征的特征组合作为多维特征。

新闻文章中多次出现的关键词,是反映新闻文章中最重要信息的一种统计特征,进行词频统计的意义在于找出文章表达的重点;此外,句子出现的位置也是反映句子重要程度的关键,出现的位置越靠前,说明该句子在文章中起到的作用越大。因此,词频、位置特征是自动摘要模型提升的关键。

3.5.3　coverage机制实验分析

本文使用的模型使用了coverage机制,试图解决生成重复内容的问题。通过计算生成摘要中1-gram、2-gram、3-gram以及4-gram所占比例,定量分析引入coverage机制解决生成内容重复问题的效果。

由表4可以看出,本文提出的BERT-PGN模型相较于NLP_ONE能够有效减少生成内容的重复,在解决重复的方面效果明显,在3-gram、4-gram的摘要结果定量分析中,接近标准摘要的效果。

表4　coverage机制验证　　单位:%

Tab. 4　Verification of coverage mechanism　　unit:%

数据片段	NLP_ONE	BERT-PGN	参考模型
1-gram	26	19	17
2-gram	18	6	3
3-gram	16	2	1
4-gram	14	1	0

4　结语

本文提出了一种面向中文新闻文本的BERT-PGN模型,结合BERT预处理模型及多维语义特征获取词向量,利用指针生成网络模型结合coverage机制减少生成重复内容。经实验表明,BERT-PGN模型在中文新闻摘要任务中,生成的摘要结果更接近标准摘要,包含更多原文的关键信息,能有效解决生成内容重复的问题。

下一步将尝试挖掘更多要素,例如:面向新闻文本的有效人工特征等,提升摘要结果;简化模型,缩短模型训练时间;提升生成摘要内容的完整性、流畅性;构建新闻领域的外部数据,帮助模型结合句子上下文充分理解句子含义。

参考文献(References)

[1] LING H. Innovative exploration on empowering students to study and propagandize the new thought under the Internet thinking taking Xi Jinping thought on the socialism with Chinese characteristics in a new era as an example[J]. PEOPLE: International Journal of Social Sciences, 2018, 4(3):1395-1408.

[2] 刘家益,邹益民. 近70年文本自动摘要研究综述[J]. 情报科

学，2017，35(7)：154-161.（LIU J Y，ZOU Y M. A review of automatic text summarization research in recent 70 years [J]. Information Science，2017，35(7)：154-161.）

[3] 吴云，杨长春，梅佳俊，等. 词句协同自动摘要提取方法[J]. 计算机工程与设计，2018，39(9)：2776-2779，2810.（WU Y，YANG C C，MEI J J，et al. Method of automatic summarization algorithm based on word-sentence co-ranking [J]. Computer Engineering and Design，2018，39(9)：2776-2779，2810.）

[4] WEI R，HUANG H，GAO Y. Sharing pre-trained BERT decoder for a hybrid summarization [C]// Proceedings of the 2019 China National Conference on Chinese Computational Linguistics，LNCS 11856. Cham：Springer，2019：169-180.

[5] 石磊，阮选敏，魏瑞斌，等. 基于序列到序列模型的生成式文本摘要研究综述[J]. 情报学报，2019，38(10)：1102-1116.（SHI L，RUAN X M，WEI R B，et al. Abstractive summarization based on sequence to sequence models：a review[J]. Journal of the China Society for Scientific and Technical Information，2019，38(10)：1102-1116.）

[6] 戴天，吴渝，雷大江. 利用组合模型生成微博热点话题事件摘要[J]. 计算机应用研究，2016，33(7)：2026-2029，2038.（DAI T，WU Y，LEI D J. Hot topic summarization on microblog generated by model combination [J]. Application Research of Computers，2016，33(7)：2026-2029，2038.）

[7] IBOI H，CHUA S，RANAIVO-MALANÇON B，et al. Performance of opinion summarization towards extractive summarization [J]. Journal of Telecommunication，Electronic and Computer Engineering，2017，9(2-10)：57-64.

[8] LUHN H P. The automatic creation of literature abstracts[J]. IBM Journal of Research and Development，1958，2(2)：159-165.

[9] 侯圣峦，张书涵，费超群. 文本摘要常用数据集和方法研究综述[J]. 中文信息学报，2019，33(5)：1-16.（HOU S L，ZHANG S H，FEI C Q. A survey to text summarization：popular datasets and methods[J]. Journal of Chinese Information Processing，2019，33(5)：1-16.）

[10] LERMAN K，BLAIR-GOLDENSOHN S，MCDONALD R. Sentiment summarization：evaluating and learning user preferences [C]// Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg，PA：Association for Computational Linguistics，2009：514-522.

[11] LIU L，LU Y，YANG M，et al. Generative adversarial network for abstractive text summarization[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto，CA：AAAI Press，2018：8109-8110.

[12] AL-SABAHI K，ZHANG Z，NADHER M. A Hierarchical Structured Self-Attentive model for extractive document Summarization（HSSAS）[J]. IEEE Access，2018，6：24205-24212.

[13] SLAMET C，ATMADIA A R，MAYLAWATI D S，et al. Automated text summarization for Indonesian article using vector space model [J]. IOP Conference Series：Materials Science and Engineering，2018，288：No. 012037.

[14] ALGULIYEV R M，ALIGULIYEV R M，ISAZADE N R，et al. COSUM：text summarization based on clustering and optimization [J]. Expert Systems，2019，36(1)：e12340.

[15] CHO K，VAN MERRIËNBOER B，GULCEHRE C，et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg，PA：Association for Computational Linguistics，2014：1724-1734.

[16] SUTSKEVER I，VINYALS O，LE Q V. Sequence to sequence learning with neural networks [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge：MIT Press，2014：3104-3112.

[17] TAN J，WAN X，XIAO J. Abstractive document summarization with a graph-based attentional neural model [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg，PA：Association for Computational Linguistics，2017：1171-1181.

[18] SIDDIQUI T，SHAMSI J A. Generating abstractive summaries using sequence to sequence attention model [C]// Proceedings of the 2018 International Conference on Frontiers of Information Technology. Piscataway：IEEE，2018：212-217.

[19] CELIKYILMAZ A，BOSSELUT A，HE X，et al. Deep communicating agents for abstractive summarization [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. Stroudsburg，PA：Association for Computational Linguistics，2018：1662-1675.

[20] KHAN A，SALIM N，JAYA KUMAR Y. A framework for multi-document abstractive summarization based on semantic role labelling[J]. Applied Soft Computing，2015，30：737-747.

[21] 江跃华，丁磊，李娇娥，等. 融合词汇特征的生成式摘要模型[J]. 河北科技大学学报，2019，40(2)：152-158.（JIANG Y H，DING L，LI J E，et al. Abstractive summarization model considering hybrid lexical features[J]. Journal of Hebei University of Science and Technology，2019，40(2)：152-158.）

[22] HUA L，WAN X，LI L. Overview of the NLPCC 2017 shared task：single document summarization [C]// Proceedings of the 2017 National CCF Conference on Natural Language Processing and Chinese Computing，LNCS 10619. Cham：Springer，2017：942-947.

[23] 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. Stroudsburg，PA：Association for Computational Linguistics，2017：1073-1083.

[24] 侯丽微，胡珀，曹雯琳. 主题关键词信息融合的中文生成式自动摘要研究[J]. 自动化学报，2019，45(3)：530-539.（HOU L W，HU P，CAO W L. Automatic Chinese abstractive summarization with topical keywords fusion[J]. Acta Automatica Sinica，2019，45(3)：530-539.）

This work is partially supported by the National Key Research and Development Program of China（2019YFC1200302），the Key Project of National Natural Science Foundation of China（61632011）.

TAN Jinyuan，born in 1997，M. S. candidate. His research interests include natural language processing.

DIAO Yufeng，born in 1987，Ph. D. candidate. Her research interests include natural language processing.

QI Ruihua，born in 1974，Ph. D.，professor. Her research interests include natural language processing.

LIN Hongfei，born in 1962，Ph. D.，professor. His research interests include natural language processing.