计算机科学
Computer Science
ISSN 1002-137X,CN 50-1075/TP

《计算机科学》网络首发论文

题目：融合 BERT 词嵌入表示和主题信息增强的自动摘要模型
作者：郭雨欣，陈秀宏
收稿日期：2021-04-10
网络首发日期：2022-02-25
引用格式：郭雨欣，陈秀宏．融合 BERT 词嵌入表示和主题信息增强的自动摘要模型[J/OL]．计算机科学.
https://kns.cnki.net/kcms/detail/50.1075.TP.20220224.1445.004.html

融合 BERT 词嵌入表示和主题信息增强的自动摘要模型

郭雨欣[1] 陈秀宏[2]

1 江南大学人工智能与计算机学院 江苏 无锡 214122
2 江苏省媒体设计与软件技术重点实验室 江苏 无锡 214122
（1171784997@qq.com）

摘要 自动文本摘要能够帮助人们快速筛选辨别信息，掌握新闻关键内容，缓解信息过载问题。主流的生成式自动文摘模型主要基于编码器-解码器架构，针对解码器端在预测目标词时未充分考虑文本主题信息，并且传统的 Word2Vec 静态词向量无法解决一词多义问题的现状，提出了一种融合 BERT 词嵌入表示和主题信息增强的中文短新闻自动摘要模型。编码器端联合无监督算法获取文本主题信息并将其融入到注意力机制中，提升了模型的解码效果；解码器端将 BERT 预训练语言模型抽取出的 BERT 句向量作为补充特征，以获取更多的语义信息，同时引入指针机制来解决词表外的单词问题，并利用覆盖机制有效抑制重复。在训练过程中，为了避免暴露偏差问题，针对不可微指标 ROUGE 采用强化学习方法来优化模型。在两个中文短新闻摘要数据集上的多组对比实验结果表明，该模型在 ROUGE 评价指标上有显著的提高，能有效融合文本主题信息，生成语句流畅、简明扼要的摘要。

关键词：生成式摘要；BERT；主题信息；注意力机制；强化学习

中图法分类号 TP391.1　　DOI：10.11896/jsjkx.210400101

Automatic Summarization Model Combining BERT Word Embedding Representation and Topic Information Enhancement

GUO Yu-xin[1] and CHEN Xiu-hong[2]

1 School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, Jiangsu 214122, China
2 Jiangsu Key Laboratory of Media Design and Software Technology, Wuxi, Jiangsu 214122, China

Abstract Automatic text summarization can help people to filter and identify informationquickly, grasp the key content of news, and alleviate the problem of information overload. The mainstream abstractive summarization model is mainly based on the encoder-decoder architecture. In view of the fact that the decoder does not fully consider the text topic information when predicting the target word, and the traditional Word2Vec static word vector cannot solve the polysemy problem, an automatic summarization model for Chinese short news is proposed, which integrates the BERT word embedding representation and topic information enhancing. The encoder combines unsupervised algorithm to obtain text topic information and integrates it into the attention mechanism to improve the decoding effect of the model. At the decoder side, the BERT sentence vector extracted from the BERT pre-trained language model is used as the supplementary feature to obtain more semantic information. Meanwhile, pointer mechanism is introduced to solve the problem of out of vocabulary, and coverage mechanism is used to suppress repetition effectively. Finally, in the training process, reinforcement learning method is adopted to optimize the model for non-differentiable index ROUGE to avoid exposing bias. Experimental results on two

到稿日期： 2021-04-10　返修日期： 2021-07-25
基金项目：江苏省研究生科研与实践创新计划项目（JNKY19__074）
This work was supported by the Jiangsu Postgraduate Research and Practice Innovation Program (JNKY19__074).
通信作者：陈秀宏(625325682@163.com)

datasets of Chinese short news summarization show that the proposed model can significantly improve the ROUGE evaluation index, effectively integrate text topic information, and generate fluent and concise summaries.

Keywords Abstractive summarization，BERT，Topic information，Attention mechanism，Reinforcement learning

1 引言

摘要被定义为从一篇或多篇文章中产生的概括性信息[1]。在当今大数据时代，文章数量爆炸式增加，尤其是诸如微博、今日头条等社交媒体的信息量井喷。为了解决信息过载问题，帮助人们快速、准确地从新闻正文中获取关键信息，利用先进的算法自动生成摘要的研究迫在眉睫。

根据生成摘要的方式不同，自动文摘分为抽取式文摘和生成式文摘。抽取式自动文摘按一定规则从原文本中选择最重要的结构单元（句子、段落等）组成摘要。虽然抽取式自动文摘方法简单，但容易存在次要或冗余信息，并且生成的摘要中相邻句子之间往往缺乏连贯性。相比之下，生成式自动文摘在思想上更接近人工摘要的过程，需要在对文本进行语义理解后生成更凝练简洁、连贯性强的摘要。随着深度学习技术的迅速发展，神经网络模型被广泛应用于各类自然语言处理任务，促进了生成式自动文摘的发展。

2015 年，Rush 等[2]首次将应用于机器翻译任务中的注意力机制引入自动文摘任务中，用卷积神经网络(CNN)编码原文信息，用上下文相关的注意力前馈神经网络生成摘要。Paulus 等[3]将强化学习引入文本摘要任务中，并提出了一个引入内嵌注意力机制和新的训练方法的模型。Chopra 等[4]使用 CNN 对原文进行编码，但用循环神经网络(Recurrent Neural Network, RNN)生成摘要，大大提高了生成摘要的质量。Nallapati 等[5]采用 RNN 对原文进行编码，同时对词特征、停用词、文档结果等有用信息进行利用。Gu 等[6]和 Zeng 等[7]使用 copy 机制解决解码阶段的词表外单词(Out Of Vocabulary, OOV)问题。See 等[8]提出了指针-生成网络模型(Pointer-Generator-Network, PGN)，通过指向和生成两种模式把抽取式摘要和生成式摘要相结合，并加入覆盖机制以抑制重复。Jonas 等[9]提出能实现并行运算的卷积序列到序列(ConvS2S)模型。Wang 等[10]将主题信息纳入卷积序列到序列模型中，并结合强化训练进一步提升摘要效果。Wang 等[11]基于文本多维度特征,提出了一种新的自动摘要生成方法。

然而，现有的大多数生成式自动文摘模型存在以下问题：1）对句子上下文理解不够充分，生成内容重复，无法解决 OOV 问题；2）注意力机制未明确地考虑原文的主题信息，可能导致解码偏离原文本主旨，并进一步传播偏差；3）一词多义现象可能导致生成摘要中同义词的不恰当使用。针对上述问题，本文提出了一种融合 BERT 句向量的主题信息高感知度的模型，并且通过结合强化学习方法得到的新的学习目标函数来进一步提高摘要评测指标 ROUGE 值。

本文的主要工作如下：

(1) 提出了一种融合 BERT 词嵌入表示和主题信息增强的面向中文新闻文本的自动摘要模型。在该模型中，编码器端通过无监督算法获取文本主题信息并将其整合到注意力机制中，以增强模型的主题感知度；解码器端将 BERT 句向量与编码器端输出的上下文向量进行拼接，得到新的特征，帮助模型获取更加抽象、更深层次的语义信息，从而生成结构连贯、语句流畅的摘要。

(2) 为了进一步生成高质量的摘要，模型引入指针机制来解决词表外单词问题，同时利用覆盖机制来有效抑制重复。在训练过程中，采用强

化学习将 ROUGE 指标与极大似然交叉熵损失相结合，改进目标函数，以减少暴露偏差。

(3)在两个中文短新闻文本摘要数据集上的大量实验结果表明，本文所提模型能聚焦原文本的核心内容，生成信息丰富、可读性强的文本摘要，以中文词语为粒度进行评估，ROUGE-1，ROUGE-2，ROUGE-L 值都得到了提高。

2 相关工作

2.1 基本的编码器-解码器模型

在自动摘要任务中，输入和输出都是不定长序列，因此可以使用基本的编码器-解码器模型。编码器将输入序列转化为定长的上下文语义向量 c，并在 c 中编码输入序列信息；解码器通过上下文语义向量解码生成输出序列。

在编码器端时间步 t 时，循环神经网络[11]将输入词向量 x_t 和上一个时间步的隐藏状态 h_{t-1} 变换为当前时间步的隐藏态 h_t，计算式如式(1)所示：

$$h_t = f(x_t, h_{t-1}) \quad (1)$$

其中，函数 f 表示编码器隐藏层的变换。

在解码器端时间步 t′ 时，解码器将上一时间步的输出 $y_{t'-1}$、上下文向量 c 以及上一时间步的隐藏状态 $s_{t'-1}$ 变换为当前时间步的隐藏状态 $s_{t'}$，计算式如式(2)所示：

$$s_{t'} = g(y_{t'-1}, c, s_{t'-1}) \quad (2)$$

其中，函数 g 表示解码器隐藏层的变换。

根据 $y_{t'-1}$，c 和 $s_{t'}$ 计算得当前时间步的解码器输出概率，计算式如式(3)所示：

$$P(y_{t'} | y_1, ..., y_{t'-1}, c) = p(y_{t'-1}, s_{t'}, c) \quad (3)$$

其中，p 是 softmax 函数。

在基本的编码器-解码器模型中，长文本的整体信息难以仅靠一个固定长度的上下文语义向量压缩表示，从而导致解码不足。为了区分原文本中的重要信息，Bahdanau 等[13]引入了注意力机制。注意力机制在预测某个位置的词汇时，会偏向关注于与该位置关系密切的原文信息。具体计算式如式(4)—式(6)所示：

$$e_i^t = v^T tanh(W_h h_i + W_s s_t + b_{attn}) \quad (4)$$

$$a^t = softmax(e^t) \quad (5)$$

$$h_t^* = \sum_i a_i^t h_i \quad (6)$$

其中，W_h 和 W_s 是学习参数；a^t 是注意力分布，对应于原始文本中每一个单词的概率分布，用于告知当前预测过程中单词的重要程度；h_t^* 是原文本的动态表示，即新的上下文向量。

2.2 BERT

传统的词嵌入模型 Word2Vec 采用无监督的训练方法训练神经网络，从而得到连续密集的词向量以表征词的信息。然而，多义词表示一直是词向量领域的难题，Word2Vec 在编码多义词时也无法区分其在不同语境下的不同含义。2018 年，Google 提出了一种新型的预训练语言模型 BERT[14]，在多项自然语言处理任务中取得了最佳成绩。BERT 使用双向 Transformer[15]中的编码器进行特征提取，得到充分利用了上下文信息的动态词向量，能够建模一词多义现象。同时，BERT 通过预训练方法学习深度网络，在不同的网络层上得到不同层次的特征，从而更好地体现词的复杂特性。在 BERT 中主要有两个特定任务，即 MLM(Masked Language Model)和 NSP(Next Sentence Prediction)，前者通过类似完形填空的任务能够捕捉词语表征，后者让语言模型理解句子之间的逻辑和因果关系以捕捉句子级别的表征。

3 本文模型

本文提出了一种融合 BERT 词嵌入表示和主题信息增强的生成式摘要模型。由图 1 所示，模型结构主要由 3 个部分组成：1)双向长短期记忆网络(Long Short Term Memory, LSTM)编码器；

2)强化主题信息的注意力机制；3)融合 BERT 句向量特征的单向 LSTM 解码器。

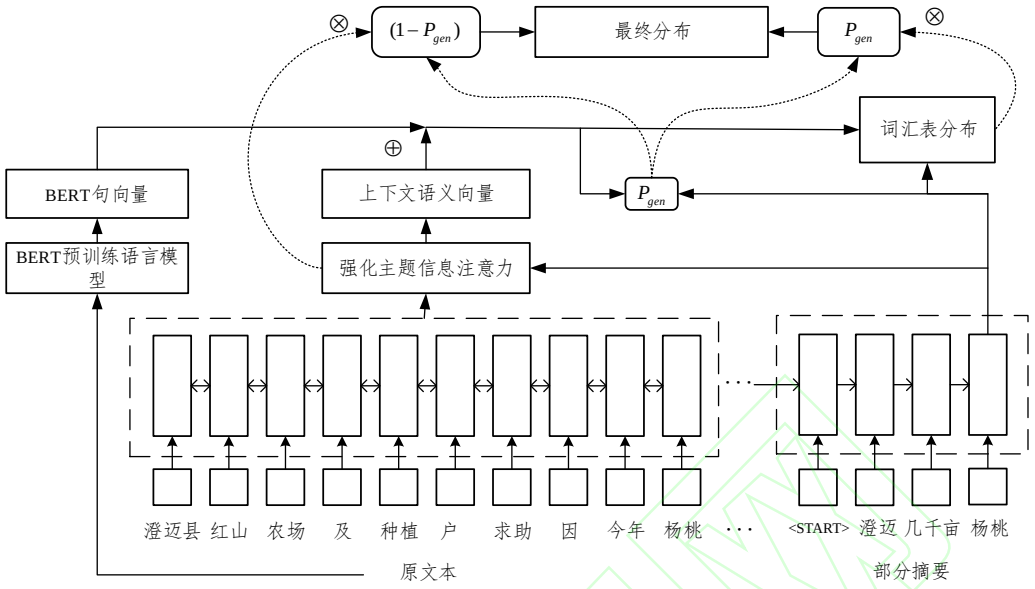

图 1 本文所提模型总体结构图

Fig.1 Overall structure diagram of model proposed in this paper

3.1 语句编码器

本文模型将原文档转化为词向量序列，然后将其输入到双向长短时记忆网络(BiLSTM)中进行编码。在 BiLSTM 结构中，前向网络正向读取输入序列以计算前向隐藏状态向量，后向网络反向读取输入序列以计算反向隐藏状态向量，两者拼接得编码器隐藏态。计算式如式(7)—式(9)所示：

$$\overrightarrow{h_t^e} = \overrightarrow{f}(x_t, h_{t-1}^e) \tag{7}$$

$$\overleftarrow{h_t^e} = \overleftarrow{f}(x_t, h_{t-1}^e) \tag{8}$$

$$h_t^e = \left[\overrightarrow{h_t^e}, \overleftarrow{h_t^e}\right] \tag{9}$$

其中，x_t 是第 t 个输入词的词嵌入向量，h_t^e 是编码器第 t 个单元的隐藏状态向量，f 是 LSTM 的非线性方程。

3.2 强化主题信息的注意力机制

在生成式自动文摘领域，虽然注意力机制的引入提升了模型效果，但生成的摘要离人工撰写的摘要还有一定差距，亟需我们将文本更深层次的信息有效嵌入到模型中。从式(4)—式(6)可以看出，基本的注意力机制只考虑当前目标词和原始文档中词语之间的相关性，没有显式地考虑文本的主题信息。为了更加契合人们围绕文章主题撰写摘要的过程，本文提出了强化主题信息的注意力机制。

词频是文档摘要中常用的特征，Luhn[16]指出，频繁出现的词语与文章主题有较大的关联。因此，本文利用 TF-IDF[17]（词频-逆文档频率），即一种评估词语对于文本的重要程度的算法，在模型训练过程中，对输入文本的所有名词、动词求取 TF-IDF 权重值并将其与编码器隐藏态相乘得到主题信息强化项。与 Hou 等[18]利用 TextRank 算法求出具体的关键词向量不同，本文方法不受预定关键词向量数目的影响，并且 TF-IDF 的 IDF 值能为模型带来统计上的优势。具体计算式如下：

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (10)$$

$$idf_i = \log_e \frac{|D|}{|\{j : t_i \in d_j\}| + 1} \quad (11)$$

$$w_i = tf_{ij} * idf_i \quad (12)$$

$$k = w_i * h_i \quad (13)$$

其中，tf 表示词条在文本中出现的频率，idf 表示该词的常见程度，k 是主题信息强化项。

为了强化所有具有类别区分能力的单词对文本的影响，充分利用词频蕴含的主题信息，将主题信息强化项作为注意力分配的一部分，计算式如式(14)所示：

$$e_i^t = v^T((1-\beta)\tanh(W_h h_i + W_s s_t + b_{attn}) + \beta \tanh(W_k k)) \quad (14)$$

其中，W_k 是新的学习参数；超参数 β 是强化影响因子，实验过程中设为 0.1。

通过融合 TF，IDF 特征进行主题信息强化的注意力机制会加强模型对文本主题信息的认识，尤其是对名词、动词关键词的认识。由于大部分摘要句是陈述句，其中形容词和副词较少，名词、动词较多，经过学习，模型会自动增强对关键词的选择，在保证生成的摘要具有较好的连贯性、语义相关性的情况下，进一步提高准确性。

3.3 融合 BERT 句向量特征的语句解码器

语句解码器的作用是生成输出序列。为了使该模型得到更多的语义信息，考虑在解码器端将 BERT 句向量与强化主题信息的注意力机制计算出的上下文向量进行拼接，从而得到新的特征向量 h_t'。

由于预训练模型生成的是字符级别的向量，对于中文词语，可以通过将字符级向量进行拼接取均值操作来得到完整的词向量表示。但该过程会损失词向量信息，影响自动摘要任务效果。并且，BERT 词向量表示中每一个词不论重复与否，都对应于一个向量，对于大规模文本数据来说，所需代价极大。因此，本文决定直接获取 BERT 句向量表示，具体操作是将输入文本按字符分割后通过 BERT 模型获得每个字符对应的词向量，然后进行全局平均池化操作，获得 768 维的句向量，将其作为新的特征嵌入到解码器端，使模型获得更多的语义信息。

将解码器的隐藏状态向量 s_t 和 h_t' 共同参与到当前时间步的词汇分布计算中，计算式如式(15)所示：

$$P_{vocab} = softmax\left(V'\left(V\left[s_t, h_t'\right] + b\right) + b'\right) \quad (15)$$

其中，V'，V，b，b' 是学习参数，P_{vocab} 是词汇表中所有单词的概率分布。

为了有效处理未登录词，本文引入了指针机制。P_{gen} 由特征向量、解码器隐藏状态向量、解码器当前输入向量，通过线性层和 sigmoid 函数求得。在解码过程中，它作为一个软开关，控制解码器从词汇表 vocab 中生成一个单词或从原文拷贝单词。具体公式如下：

$$P_{gen} = \sigma\left(w_h^T * h_t' + w_s^T s_t + w_x^T x_t + b_{ptr}\right) \quad (16)$$

其中，W_h^T，W_s^T，W_x^T，b_{ptr} 是学习参数。

当前预测词 w 的概率 $P(w)$ 的计算式如下：

$$P(w) = P_{gen} * P_{vocab}(w) + (1 - P_{gen}) \sum_{i:w_i} a_i^t \quad (17)$$

为了抑制生成摘要内容重复，引入覆盖机制。覆盖机制会维护一个覆盖向量 c^t（即过去时刻注意分布之和），记录了到目前为止从注意力机制中获得的词语覆盖程度。相反，覆盖向量也会影响当前时间步的注意力分布，注意力分布的计

算式由式(14)更新为式(19)。并且设定覆盖损失 $covloss$，将其作为总的损失函数的一部分，以惩罚重复关注。计算式如式(18)—式(21)所示：

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

$$e_i^t = v^T((1-\beta)\tanh(W_h h_i' + W_s s_t + W_c c_i^t + b_{attn}) + \beta\tanh(W_k k)) \quad (19)$$

$$covloss = \sum_i min(a_i^t, c_i^t) \quad (20)$$

$$loss = \frac{1}{T}\sum_{t=0}^{T}\left[-\log_e P(w_t^*) + \lambda\sum_i covloss_t\right] \quad (21)$$

其中，λ 是一个超参数，实验过程中设为 0.3。

4 强化学习与混合目标函数

在序列生成任务中，教师强制(teacher forcing)算法[19]是训练 RNN 解码器最广泛使用的方法，旨在每个解码步骤中最小化极大似然损失。定义 $y^* = \{y_1^*, y_2^*, ..., y_{n'}^*\}$ 为给定输入序列 x 的真实输出序列，极大似然训练目标是使以下损失最小化：

$$L_{ml} = -\sum_{t=1}^{n'} \log p(y_t^* | y_1^*, ..., y_{t-1}^*, x) \quad (22)$$

然而，最小化 L_{ml} 并不总是在离散评估指标 ROUGE 上产生最佳结果。一是因为仅将模型暴露于训练数据的分布而不是自身的预测分布会导致暴露偏差问题。其原因是在训练过程中，模型给定真实输出序列以预测下一个单词；而在推理过程中，给定预测单词作为输入以生成下一个单词，预测错误产生累积，模型性能下降。二是因为极大似然交叉熵损失与 ROUGE 评估标准不一致。前者鼓励模型预测与参考摘要完全相同的内容，惩罚那些即使语义相似的不同内容，忽略了摘要的内在属性；后者会考虑摘要的灵活性，鼓励模型更多地关注语义而不是单词级的对应关系。

为了解决上述问题，本文利用自关键序列训练(SCST)[20]，一种用于强化学习的策略梯度算法，最大化不可微 ROUGE 度量。在强化学习过程中，给定输入序列 x，生成两个输出序列 $\hat{y}$ 和 y^s。$\hat{y}$ 通过贪婪选择输出概率分布最大的单词来获得，y^s 通过随机采样输出概率分布来获得。在得到两个序列的 ROUGE 分数并将其作为奖励后，即 $r(y^s)$ 和 $r(\hat{y})$，最小化强化损失并通过梯度下降更新模型参数。强化损失的计算式如下：

$$L_{rl} = (r(\hat{y}) - r(y^s))\sum_{t=1}^{n'} \log p(y_t^s | y_1^s, ..., y_{t-1}^s, x) \quad (23)$$

虽然上述强化学习策略弥补了暴露偏差问题，但优化学习最大化奖励函数不一定保持生成序列的可读性[21]。而极大似然训练目标在本质上是条件语言模型，能帮助策略梯度算法生成更自然的摘要。因此，本文采用一种混合学习目标，它将极大似然目标与强化学习目标相结合，以获得混合损失。混合损失的计算式如式(24)所示：

$$loss_{mixed} = \gamma L_{rl} + (1-\gamma) loss \quad (24)$$

其中，γ 是一个值为 0~1 的超参数，实验过程中设置为 0.9。

5 实验

5.1 数据集与预处理

5.1.1 LCSTS 数据集

表 1 LCSTS 数据集的统计信息

Table 1 Dataset statistics of LCSTS

Part 1	Number of Pairs	2400591
Part 2	Number of Pairs	10666
	Human Score 1	942
	Human Score 2	1039
	Human Score 3	2019
	Human Score 4	3128
	Human Score 5	3538
Part 3	Number of Pairs	1106
	Human Score 1	165
	Human Score 2	216

Human Score 3	227
Human Score 4	301
Human Score 5	197

LCSTS 数据集是由 Hu 等[22]提出的，取自新浪微博的大规模中文短文本摘要数据集。该数据集包含了政治、经济、军事、电影、游戏、民生等多个领域的 200 万真实的中文短文本数据和每个文本作者给出的摘要。由表 1 所列，该数据集主要由 3 部分组成：第一部分包含 2400591 个短文本摘要对；第二部分包含 10666 个人工标记的短文本摘要对，分数为 1 到 5，其中“1”表示“最不相关”，“5”表示“最相关”；第三部分共 1106 对，由 3 人同时评分。

在实验过程中，将第一部分作为训练集，将第三部分中人工评分为 3，4 和 5 的文本对作为测试集，来完成短文本摘要生成任务。

5.1.2 CSTSD 数据集

CSTSD 数据集是一个公开的中文短文本摘要数据集，来自新浪微博的主流媒体发表的微博。该数据集一共包含 679898 条数据，其中包含的短文本约有 100~200 字，摘要约有 10~20 字，按照大约 66:1:1 的比例将其划分为训练集、测试集和验证集。

5.1.3 数据集预处理

在预处理时，首先用正则表达式将原始数据集中的标签、符号等非法字符去除，然后使用 jieba 开源分词工具对文本进行分词，接着进行数据去重、去停用词操作。由于 LCSTS 数据集中大多数文本的长度不超过 400，CSTSD 数据集中大多数文本的长度不超过 200，我们设置输入文本长度的上限为 512，对于极个别长度大于该阈值的文本，采取截断处理。

5.2 评价指标

自动摘要的评价采用官方 ROUGE 作为度量指标。ROUGE 是一种基于摘要中 n 元词的共现信息评价摘要的方法，该方法的基本思想为计算参考摘要和候选摘要之间的重叠词汇单元来衡量生成摘要的质量。本文采用 ROUGE-1，ROUGE-2，ROUGE-L 值进行评估，其中 1，2，L 分别代表基于 1 元词、2 元词和最长公共子序列。

5.3 实验参数设置

本文实验的硬件环境为 NVIDIA GTX1080Ti 单个 GPU，Intel 处理器，软件环境为 PyTorch 1.5.1。实验过程中使用词汇规模为 5 万的词汇表，不能识别的词语用 UNK(Unknown Words)表示。Word2Vec 词嵌入维度设置为 128，BERT 句向量维度为 768，所有的 LSTM 隐藏状态维度设置为 256，新的特征向量 h_t' 的维度大小为 1280。采用 Adagrad 优化器更新参数，学习率初始值为 0.15，平滑因子 eps=1×10^{-12}。最大迭代次数为 500000。解码器部分，集束搜索(beam search)大小设为 4。为了加快模型训练和收敛速度，mini-batch 大小设为 8。使用来自极大似然模型的最佳权重初始化强化学习并选择 ROUGE-L 度量作为强化奖励函数。

5.4 对比模型

（1）TextRank[23]。该方法采用 textrank4zh 库从原始新闻文章中抽取信息度高的重要句子生成目标摘要。

（2）RNN[22]。使用 RNN 作为编码器，将最后的隐藏态作为解码器的输入，在解码过程中不使用上下文。

（3）RNN-context[22]。使用 RNN 作为编码器，在解码过程中使用上下文，编码器的所有隐藏态的组合作为解码器的输入。

（4）Seq2Seq+Attn[8]。融合注意力与编码器-解码器的模型。

（5）Pointer-generator[8]。该模型将基于注

意力机制的序列到序列模型与指针网络结合，允许通过从原文本中复制词或从固定词汇表中生成新词。

（6）Pointer-generator+coverage[8]。指针-生成网络模型(PGN)作为本文选定的基线模型，在前一模型基础上加入 coverage 机制，对解码器重复关注同一位置信息的情况进行惩罚，避免生成重复的文本。

（7）WordNet+Dual-attn+PGN+Cov[24]。模型基于 WordNet 句子排名算法抽取重要句子，融合抽取式摘要与生成式摘要方法，并引入 pointer-generator 和 coverage 机制。

5.5 实验结果及分析

将本文模型与上述对比模型分别在 LCSTS 和 CSTSD 数据集上进行实验，使用 4.2 节中的评价指标，以词语为粒度进行评估，实验结果分别如表 2 和表 3 所列。

表 2 LCSTS 数据集的实验结果对比

Table 2 Comparison of experimental results of LCSTS

(单位：%)

Model	R-1	R-2	R-L
TextRank	15.32	7.80	16.23
RNN	21.32	8.54	18.23
RNN-context	24.93	14.59	21.27
Seq2Seq+Attn	27.75	15.57	25.95
Pointer-generator	36.68	21.39	31.12
PGN	37.16	24.67	33.96
WordNet+attn+PGN+Cov	37.39	25.19	34.61
PGN+BERT	37.28	24.81	33.02
PGN+BERT+Topic	38.47	25.75	34.72
PGN+BERT+Topic+RL	**39.46**	**26.28**	**35.84**

表 3 CSTSD 数据集的实验结果对比

Table 3 Comparison of experimental results of CSTSD

(单位：%)

Model	R-1	R-2	R-L
TextRank	17.31	6.56	18.27
RNN	21.62	9.23	20.44
RNN-context	25.52	16.71	23.96
Seq2Seq+Attn	29.05	18.04	28.07
Pointer-generator	39.35	26.01	37.73
PGN	41.14	31.93	37.99
WordNet+attn+PGN+Cov	41.72	32.54	38.26
PGN+BERT	41.56	32.09	38.12
PGN+BERT+Topic	42.66	33.25	38.87
PGN+BERT+Topic+RL	**43.89**	**33.70**	**39.95**

由表 2 和表 3 可知：

(1) 本文模型在两个数据集上 R-1，R-2 和 R-L 的评估结果(%)分别达到了 39.46，26.28，35.84 和 43.89，33.70，39.95，均优于其他模型。这表明本文模型为生成高质量摘要做出了一定贡献。

(2) 与基线模型 PGN 相比，PGN+BERT 句向量特征模型的性能得到了微弱提升，说明融入 BERT 句向量特征能增添语义信息，但作用不大，总体性能甚至弱于模型 WordNet+Dual-attn+PGN+cov。

(3) 与基线模型 PGN 相比，PGN+BERT 句向量特征+主题信息强化注意力机制模型在两个数据集上 ROUGE-1 指标分别提升了 1.31%和 1.52%，说明该注意力机制能帮助模型聚焦于文本核心内容，并且融合 BERT 句向量特征后能使模型对文本中句子的语义理解更加深刻、准确，从而有效提升生成摘要所含的信息量。

(4) 加入强化学习训练方法后，PGN+BERT+Topic+RL 模型的性能得到了进一步的提升，与基准模型相比，在 ROUGE-L 上分别提升了 1.88%和 1.96%，说明基于自关键序列训练方法能够帮助模型生成更具可读性的摘要。

由于 ROUGE-1 和 ROUGE-2 侧重于衡量生成摘要的信息量，ROUGE-L 侧重于衡量生成摘

要的可读性，从 3 个评估指标可以看出，强化主题信息的注意力机制会使模型学会像人类攥写摘要的方式一样围绕文章主题内容来生成摘要；而在解码器端融合 BERT 句向量特征能帮助模型结合预训练语言模型的优势进一步加强对全文信息的理解和把控，从而帮助本文模型提取出文本中更多的潜在特征信息，在多个维度上准确领会原文本的核心思想，生成信息丰富、可读性强的优质摘要。

为了更直观地体现本文模型生成摘要的可读性，将该模型与 Seq2Seq+Attn 以及 PGN 模型生成的摘要进行比较。表 4 列出了来自两个数据集的示例的原文本、参考摘要和生成摘要。通过示例可看出，Seq2Seq+Attn 模型极易产生[UNK]词，严重影响了生成摘要的可读性。PGN 模型能基本解决词表外单词问题，但有时并未聚焦原文本的核心内容，并且可能会错误组合原文片段，生成信息错误的摘要，影响了摘要的准确性和可读性。本文模型生成的摘要并不是对单词进行简单的拼凑或像抽取式摘要一样挑选出原文出现过的句子，而是能够站在全局的角度归纳文本内容、提炼重要信息，同时利用指针机制和覆盖机制来避免产生[UNK]词和内容重复问题，得到贴合主题的内容和可读性强的摘要。

表 4 摘要结果示例

Table 4 Examples of summarization results

文本：钢铁产业的持续低迷，让越来越多的钢企将赚钱的希望寄托在非钢产业上。再加上广钢卖香肠、矿泉水，武钢尝试养猪等等这样一些吸引眼球的副业，钢铁类国企发展各类辅业的冲动或已势不可挡。而且，其它领域的国有企业也不例外。
参考摘要：钢铁行业国企领亏大力发展副业，武钢不是个例
Seq2Seq+Attn：[UNK]：钢铁类国企发展各类辅业的冲动
PGN：武钢：越来越多的国有企业也不例外
Proposed：钢铁类国企发展各类辅业冲动或已势不可挡
文本：苏州一对小夫妻搂着 20 个月大的儿子睡觉，不料母亲的头发缠住了孩子的脖子，幸亏孩子父亲及时发现。目前，孩子正住院接受治疗，会不会对身体造成影响还有待观察。医生发出提醒：孩子出生后应该单独睡一张床，年轻妈妈也最好不要留长发。
参考摘要：睡梦中妈妈长发缠昏一岁多幼子
Seq2Seq+Attn：7 岁男孩被母亲头发[UNK]了孩子命
PGN：小夫妻搂着 20 个月大儿子睡觉头发缠住孩子
Proposed：苏州小夫妻搂 20 个月大儿子睡觉母亲头发缠住孩子

结束语

自动文摘是自然语言处理领域的重要研究方向之一。本文提出了一种融合 BERT 词嵌入表示和主题信息增强的中文短新闻生成式摘要模型。在 LCSTS 和 CSTSD 数据集上的实验结果表明，该模型与当前主流的生成式摘要方法相比，生成的摘要 ROUGE 指标明显提升，信息丰富，可读性强。

虽然本文模型在中文短文本摘要数据集上具有良好的性能，但如何扩展该方法以解决长文本摘要问题仍需继续研究。由于在实验过程中设置的输入文本长度上限大于所用数据集中绝大多数短新闻文本的长度，简单的截断操作不会对模型性能有不良影响。但考虑到很多长文本末句一般

具有重要信息，截断处理会使该模型在长文本摘要问题上表现欠佳。并且由于所用模型基于 RNN 的 Seq2Seq 框架，有其固有缺陷，即长距离依赖问题，对稍长的输入序列的处理能力十分有限。在未来工作中，我们将在编码器端使用 Transformer 结构获得句子的 embedding 表示，然后通过结合 LDA[25]主题模型方法进行关键句抽取得到关键句信息从而压缩长文本，再运用模型生成摘要。

参考文献

[1] HU X, LIN Y, WANG C. Overview of automatic textsummingtechnology[J]. Journal of Information,2010,29(8):144-147.

胡侠，林晔，王灿.自动文本摘要技术综述[J].情报杂志，2010，29(8):144-147.

[2] RUSH A M, CHOPRA S, WESTON J. A neural attention model for abstractive sentence summarization[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 379-389.

[3] PAULUS R,XIONG C, SOCHER R. A deep reinforced model for abstractive summarization[J].arXiv: 1705.04304,2018.

[4] CHOPRA S, AULI M, RUSH A M. Abstractive sentence summarization with attentive recurrent neural networks[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies.2016: 93-98.

[5] NALLAPATI R, ZHOU B W, GULCEHRE C, et al. Abstractive text summarization using sequence-to-sequence RNNs and beyond[C]//Proceedings of the 20th SIGNLL Conference on Computatioal Natural Language Learning. Stroudsburg: Association for Computational Linguistics.2016: 280-290.

[6] 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. 2016: 1631-1640.

[7] ZENG W, LUO W, FIDLER S, et al. Efficient summarization with read-again and copy mechanism[J]. arXiv:1611.03382,2016.

[8] SEE A, LIU P J, MANNING C D. Get to the point summarization with pointer-generator networks[C]//Proceedings of the 55th Annual Meetings of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics,2017:1073-1083.

[9] JONAS G, MICHAEL A, DAVID G, et al. Convolutional sequence to sequence learning[J]. arXiv:1705.03122, 2017.

[10] WANG L, YAO J L, TAO Y Z, et al. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization[C]// Proceedings of the Twenty-Seventh Inter-national Joint Conference on Artificial Intelligence, 2018:4453-4460.

[11] WANG QS, ZHANG H, LI F. Automatic Summary Generation Method Based on Multidimensional Text Feature[J]. Computer Engineering, 2020, 46(9): 110-116.

王青松, 张衡, 李菲. 基于文本多维度特征的自动摘要生成方法[J]. 计算机工程, 2020, 46(9): 110-116.

[12] ILYA S, ORIOLV, QUOCV L. Sequence to sequence learning with neural networks [C]//Advances in Neural Information Processing Systems,2014:3104-3112.

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

[14] JACOB D, CHANG M, KENTON L, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv:1801.04805v2,2018.

[15] ASHISH V, NOAM S, NIKI P, et al. Attention is all you need[J]. arXiv:1706.03762,2017.

[16] LUHN H P. The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research and Development,1958, 2(2):159-165.

[17] AIZAWA A. An information-theoretic perspective of tf-idf measures[J]. Information Processing & Management, 2003, 39(1):45-65.

[18] 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.

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.

[19] WILLIAMS R J, ZIPSER D. A learning algorithm for continually running fully recurrent neural networks[J]. Neural Computation, 1998, 1(2):270–280.

[20] RENNIESJ, MARCHERETE,MROUEHY,et al.Self-Critical Sequence Training for Image Captioning[C]//Proceedings of the 2017 Conference of the IEEE Computer Vision and PatternRecognition.2017:1179-1195.

[21] LIU C,LOWER, SERBANI, et al.How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation [C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics.2016:2122-2132.

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

[23] MIHALCEA R, TARAU P. TextRank: Bringing Order into Texts[C]//Conference on Empirical Methods in Natural Language Processing.2004:404-411.

[24] XIE N, LI S, REN H, et al. Abstractive summarization improved by WordNet-based extractive sentences[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Cham:Springer,2018:404-415.

[25] WANG T, LI M.Research on Comment Text Mining Based on LDA Model and Semantic Network[J].Journal of Chongqing Technology and Business University(Natural Science Edition),2019,36(4):9-16.

王涛，李明.基于 LDA 模型与语义网络对评论文本挖掘研究[J].重庆工商大学学报（自然科学版）,2019,36(4):9-16.

GUO Yu-xin , born in 1997, postgraduate. Her main research interests include natural language processing and text summarization.

CHENXiu-hong , born in 1964, Ph.D supervisor. His main research interests include pattern recognition and intelligent computing, etc.