计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900180

# 基于单词-章节关联的科技论文摘要

付 颖 王红玲 王中卿
苏州大学计算机科学与技术学院 江苏 苏州 215006
(20184227030@stu.suda.edu.cn)

**摘 要** 为科技论文生成自动摘要,这能够帮助作者更快撰写摘要,是自动文摘的研究内容之一。相比于常见的新闻文档,科技论文具有文档结构性强、逻辑关系明确等特点。目前,主流的编码-解码的生成式文摘模型主要考虑文档的序列化信息,很少深入探究文档的篇章结构信息。为此,文中针对科技论文的特点,提出了一种基于"单词-章节-文档"层次结构的自动摘要模型,利用单词与章节的关联作用增强文本结构的层次性和层级之间的交互性,从而筛选出科技论文的关键信息。除此之外,该模型还扩充了一个上下文门控单元,旨在更新优化上下文向量,从而能更全面地捕获上下文信息。实验结果表明,提出的模型可有效提高生成文摘在ROUGE评测方法上的各项指标性能。

**关键词**:科技论文摘要;自动文摘;生成式文摘;篇章结构;层次结构
**中图法分类号** TP18

## Scientific Paper Summarization Using Word-Section Association

FU Ying, WANG Hong-ling and WANG Zhong-qing
School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China

**Abstract** With the development of science and technology, people need to access a large number of scientific and technological information quickly, and scientific paper is one of the main ways to carry scientific and technological information. As an important part of scientific paper, abstract is an effective tool for readers to retrieve literature. Therefore, the quality of abstract affects the retrieval rate of paper directly. However, due to the lack of writing experience, the quality of abstracts written by many authors is not high. Automatic generation of summary for scientific paper can help the author grasp the important content of paper more effectively, so as to write high-quality abstract. At the same time, the automatically generated abstract can also control the number of words in the abstract, which can bring more content to readers and help them understand the paper better. Generating automatic summarization for scientific paper can help author write abstract faster, which is one of the research contents in automatic summarization. Compared with common news document, scientific paper has the characteristics of strong structure and clear logical relationship. As far as the mainstream abstractive summarization such as encoder-decoder model is concerned, it mainly considers the serialized information in the document, and rarely explores the text structure information in the document. For this reason, according to the characteristics in scientific papers, this paper proposes an automatic summarization model based on the hierarchical structure of "word-section-document", which uses the association between word and section to enhance the level of text structure and the interaction between levels, so as to screen out the key information in scientific paper. In addition, a context gate unit is extended to update the optimized context vector, thus capturing context information more comprehensively. The experimental results show that the proposed model can effectively improve the performance of the generated summarization in the ROUGE evaluation method.

**Keywords** Scientific paper summarization, Automatic summarization, Abstractive summarization, Text structure, Hierarchical structure

到稿日期:2020-09-24 返修日期:2021-01-04 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61976146)
This work was supported by the National Natural Science Foundation of China(61976146).
通信作者:王红玲(hlwang@suda.edu.cn)

## 1 引言

随着科学技术的日益发展，人们需要快速访问大量科技信息，而科技论文便是承载科技信息的主要方式之一。科技论文摘要作为科技论文的重要组成部分，是读者检索文献的有效工具，因此科技论文摘要(abstract)的质量，直接影响着论文的被检索率[1]。然而，许多论文的作者写作经验不足，撰写的摘要内容质量不高。为科技论文(除摘要以外的部分)自动生成摘要(summary)，可以帮助论文作者更有效地把握论文的重要内容，以此撰写出质量上乘的论文摘要。同时，自动生成的论文摘要还可以控制摘要的字数，为读者带来更多内容，帮助读者更好地了解论文。科技论文摘要是自动文摘的一个研究分支，可分为抽取式文摘(extractive summarization)[2]和生成式文摘(abstractive summarization)[3]。就目前而言，抽取式文摘的应用比较广泛，但无法保证摘要内容的连贯性和一致性。随着深度学习技术的发展日益成熟，对生成式文摘的研究也日益增多，它能有效地缓解抽取式文摘所带来的问题。

目前，生成式文摘常用的框架为编码-解码模型(encoder-decoder)[4]，通常与注意力机制(attention mechanism)[5]配套使用，以增强模型的效果。很多短文本新闻类生成式文摘系统均使用此结构构建模型[6]。然而，该框架侧重于线性信息，对结构信息的学习略显不足。现阶段，研究自动文摘的数据集越来越多，但是大多是基于短文本新闻类的数据集(如2015年Hermann等[7]提出的CNN/Daily Mail等)，对科技论文文摘的研究较少。科技论文拥有较强的结构性和明确的逻辑关系，一般具有固定章节(如引言、实验方法、实验结果、结论等)。因此对科技论文进行自动文摘除了需要考虑传统的线性信息外，还应考虑结构信息，如篇章层次结构。

从语言学角度来说，篇章层次结构是篇章结构的一个重要内容。篇章结构是篇章中不同层次的结构单元的组成形式，通常具有层层递进的隶属关系[8]。一般来说，一篇文章所具有的篇章层次结构有多种，如详细的有“字-词-句子-段落-文档”，简单的有“词-段落-文档”等。而科技论文的章节是具有严谨逻辑关系的篇章单元，所以根据章节划分层次结构，可使科技论文的文本结构更为清晰。基于此，我们把整篇论文分成两个层级：单词层级和章节层级，这样整篇科技论文就具有“单词-章节-论文”的层次结构。如图1所示，章节信息能够对论文摘要产生影响。科技论文摘要通常包含研究目的、实验方法及结果和研究结论等部分。“引言”章节通常包含研究目的，“方案及验证”章节包含实验方法及结果，“结束语”章节包含研究结论。由此可以看出，章节信息对于论文摘要具有重要的指导意义，而考虑单词-章节之间的关联则可以有效地利用章节信息对单词信息进行过滤和控制，从而筛选出章节内的重要单词，最终帮助作者生成内容更全面、层次更清晰的摘要。

综上所述，本文提出了一种基于单词-章节关联的科技论文摘要模型，将篇章结构等语言学知识融合到自动文摘的实际应用中。具体来讲，本文构建了一种单词-章节关联的编码器，将科技论文分为单词层级和章节层级，这两个层级会进行相互作用，从章节层级到单词层级进行对重要信息的过滤作用，从单词层级到章节层级进行对各个层级信息的控制作用，由此筛选出整篇论文的关键信息。并且，本文在解码端配备了一种基于上下文的门控单元，进一步提升了生成的文摘质量。

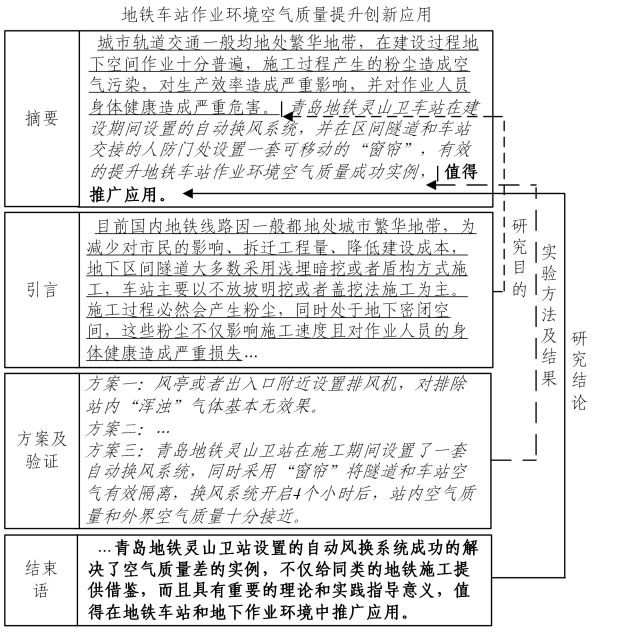


图1 章节信息对论文摘要产生影响的示例

Fig. 1 Example of the impact of section information on abstract

## 2 相关工作

近年来，随着自动文摘技术的不断发展，作为自动文摘分支之一的科技论文摘要技术也日益受到研究人员的关注。抽取式科技论文文摘由于可以避免复杂的自然语言生成问题，因此受到许多学者的喜爱。2002年Teufel等[9]提出训练一种有监督的贝叶斯分类器来选择论文中的重要内容。2017年Collins等[10]使用循环神经网络编码句子，并学习传统的句子特征(如TF-IDF[11]、句子长度等)作为上下文信息，以此方式选出最佳句子集合作为最终摘要。2019年，Xiao等[12]使用将全局信息与局部信息相结合的方式来决定论文中每个句子是否应该被保留下来作为科技论文摘要的内容。但是，抽取式摘要内容往往存在着表达不连贯、冗余信息较多等问题，并且发展遇到了瓶颈。而生成式自动文摘的研究由于模型拥有更强的表征能力，随着深度学习的成熟而逐渐兴起。2016年，Kim等[13]提出将文档分解成多个段落，然后从每个段落抽取出关键句作为该段落的目标文摘，从而构造出多个(段落，关键句)对，并把这样的对输入从序列到序列模型中进行训练。2018年Cohan等[14]首次提出将篇章层次结构应用于传统的注意力机制中，构造了一种分层的注意力机制，从而使得生成的文摘内容更具层次性。虽然文献[14]运用了层次结构信息，但是并没有对层级间关联作用进行探究。

如今，篇章结构分析等一系列语言学理论和技术日渐成熟，许多学者逐渐把目光投入到将语言学知识应用到自然语言处理等科学领域中这一研究方向上。2019年，Liu等[15]利

用篇章修辞知识提高了自动文摘的可读性。同年，Wu 等[16]将篇章结构中的层次结构运用于中文新闻类文摘。同样地，该文献也没有在层次结构的基础之上，对层级之间的关联性影响进行深入探究。实验结果表明，这些语言学知识能够有效提升自动文摘的性能。

本文构建了一种将层次结构以及层级间关联性信息投入于科技论文生成式自动文摘任务的模型框架，它能在篇章层次结构的基础上，有效利用层级间的关联性作用，从整体角度理解文本，因而能够很好地生成结构清晰、层次性更强的摘要。

## 3　单词-章节关联编码器

本节将详细介绍如何基于篇章的层次结构以及层级关联来构建单词-章节关联编码器。单词-章节关联编码器在科技论文的基础上构建了单词层级编码器和章节层级编码器，并在不同层级间进行信息交互，从单词层级到章节层级和从章节层级到单词层级这两个方向的关联实现了对层级信息过滤和控制作用，从而实现了单词-章节关联编码器对论文内容的筛选作用。

在此之前，为了更清晰地阐述单词-章节关联编码器的工作以及其他模块与它的联系，本文先给出基于单词-章节关联的科技论文摘要模型的整体工作流程。如图 2 所示，基于单词-章节关联的科技论文摘要模型主要包括：单词-章节关联编码器、配备上下文门控单元的 PG(Pointer Generator)[17]网络解码器。

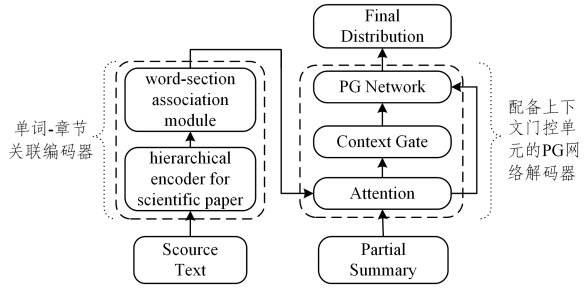


图 2　基于单词-章节关联的科技论文摘要模型结构

Fig. 2　Structure of model for scientific paper summarization using word-section association

在编码端，科技论文层级编码器读取科技论文，并在单词层级上建立了单词层级语义表示，在章节层级上建立了章节层级语义表示；接着，单词-章节关联模块使得层级信息进行交互作用，对论文内容进行有效筛选，最终获取全局文本语义表示，并将其与人工摘要输入至解码端。在解码端，利用注意力机制对编码端的输出进行操作，得到上下文向量以及注意力分布；然后，添加上下文门控单元对上下文向量更新优化；最后，注意力分布和优化后的上下文向量通过 PG 网络，输出最终的单词概率分布。

### 3.1　科技论文层级编码器

与以往的新闻类短文本数据集不同，科技论文是一种具有较强的结构性和逻辑性的长文本。因此，为了简单而有效地把握全局层次结构，如图 3 所示，本文将其层次结构划分为“单词-章节-文档”。


图 3　科技论文的层次结构

Fig. 3　Hierarchical structure of scientific paper

为了方便对本文提出的模型进行描述，我们首先明确模型的任务定义。对于给定的输入文档 $D$，本文将其章节序列和单词序列分别定义为 $D_s=(s_1,s_2,\cdots,s_{T_s})$ 和 $D_x=(x_1,x_2,\cdots,x_{T_x})$，其中，$T_s$ 是章节序列长度，$T_x$ 是单词序列长度。基于科技论文的生成式自动文摘的任务是，输入文档 $D$，经过自动文摘模型生成简短的文摘序列 $Y=(y_1,y_2,\cdots,y_{T_y})$，其中，$T_y$ 为文摘序列长度并且 $T_y<T_x$。

单词层级编码器：如图 4 所示，由于科技论文的编码序列过长，因此我们将基于单词层级的输入序列 $(x_1,x_2,\cdots,x_{T_x})$ 先通过词嵌入向量矩阵，得到 $(X_1,X_2,\cdots,X_{T_x})$，再按照章节顺序依次输入到神经网络中，考虑到 LSTM(Long Short-Term Memory)[18]可有效避免长期依赖问题，而 BiLSTM(Bi-directional LSTM)[19]相较于单向 LSTM，能更好地捕获序列的上下文信息，因此本文将采用 BiLSTM 作为输入序列的编码网络结构。最终输出序列对应的隐藏层状态表示为 $(h_1^w,h_2^w,\cdots,h_{T_w}^w)$，记作 $h^w$，将其作为单词层级的语义表示。

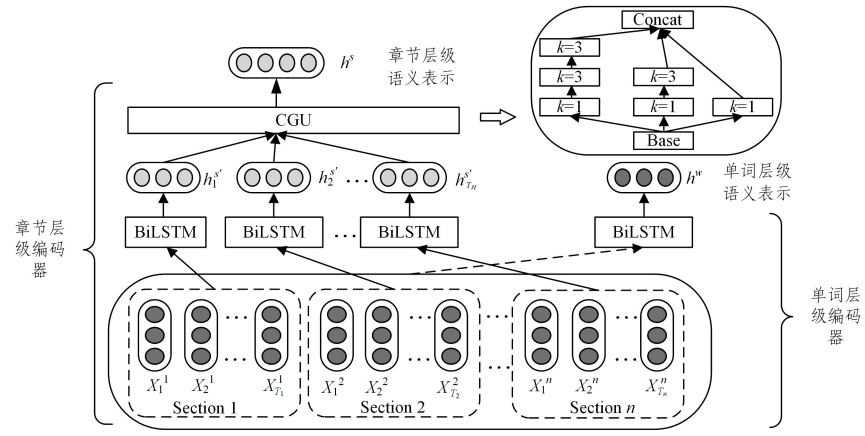


图 4　科技论文层级编码器结构

Fig. 4　Structure of hierarchical encoder for scientific paper

章节层级编码器：下面以章节 $j$ 的单词序列编码为例来详细阐述章节层级编码器的运行机制。首先，章节 $j$ 的单词序列 $(x_1^j, x_2^j, \cdots, x_{T_j}^j)$ 经过词嵌入向量矩阵得到对应的词嵌入向量序列 $(X_1^j, X_2^j, \cdots, X_{T_j}^j)$。其中，$T_j$ 为章节 $j$ 的单词序列长度。接着，将此词嵌入向量序列输入到单层 BiLSTM 中，获取输出最后一个时间步的隐藏层状态表示 $h_j^{s'}$，并将其作为章节 $j$ 的中间语义表示，计算方法如式(1)所示：

$$h_j^{s'} = \text{BiLSTM}(X_1^j, X_2^j, \cdots, X_{T_j}^j) \tag{1}$$

其中，BiLSTM(·)为双向循环神经网络函数，其输出为最后一个时间步的隐藏层状态向量。

本文从章节层级编码器得到了各个章节的中间语义表示 $(h_1^{s'}, h_2^{s'}, \cdots, h_{T_s}^{s'})$，其中，$T_s$ 为章节序列的数目。我们将其组成列表，输入至 2018 年 Lin 等[20]提出的 CGU(Convolutional Gated Unit)中。CGU 本质上是由 CNN(Convolutional Neural Network)[21]门控单元组成的，其具体结构如图 4 所示，其中 $k$ 表示卷积核大小(kernel size)。CGU 由三堆卷积网络构成，第一堆的卷积核大小分别为 $k=1, k=3, k=3$；第二堆的卷积核大小分别为 $k=1, k=3$；第三堆的卷积核大小为 $k=1$。其卷积网络块的计算细节如式(2)所示：

$$h_j^s = \text{ReLU}(\boldsymbol{W}[h_{j-k/2}^{s'}, \cdots, h_{j+k/2}^{s'}] + b) \tag{2}$$

其中，ReLU(Rectified Linear Unit)是由 Nair 等[22]提出的一种非线性激活函数；$\boldsymbol{W}$ 为可训练参数矩阵。

选取 CGU 作为模型的基于章节层级的编码器，其主要原因为卷积核可实现参数共享，能使模型有效提取某些特征，特别是 n-gram 特征。本文将 CGU 这一优势特征与科技论文的章节层级的内容特点相结合，用于提取章节之间的内部相关性。最后，通过此卷积神经网络门控单元进行编码，得到所有章节的语义表示 $(h_1^s, h_2^s, \cdots, h_{T_s}^s)$，记作 $h^s$，将其作为章节层级的语义表示。

### 3.2 单词-章节关联模块

对于基于科技论文的自动文摘来说，模型的性能很大程度上取决于生成文摘内容的简练性和关键性。为了筛选出科技论文的核心内容以及提高生成文摘的质量，本文模型引入了单词-章节关联模块。

2019 年 Wang 等[23]提出了一种文档与模板的双向选择机制，该机制被证明可对文档进行有效筛选。本文受此启发，在科技论文层级编码器的基础上，增加单词-章节关联模块。如图 5 所示，单词-章节关联模块主要包含两个子模块：从章节层级到单词层级的关联模块和从单词层级到章节层级的关联模块。

从章节层级到单词层级的关联模块：这一模块主要负责对单词层级的语义表示 $h^w$ 进行过滤。我们用章节层级的语义表示 $h^s$ 对单词层级的语义表示 $h^w$ 为过滤后的单词层级语义表示。具体计算方法如式(3)、式(4)所示：

$$g_i = \sigma(\boldsymbol{W}_{wh} h_i^w + \boldsymbol{W}_{sh} h^s + b_s) \tag{3}$$

$$h_i^g = h_i^w \otimes g_i \tag{4}$$

其中，$\boldsymbol{W}_{wh}$，$\boldsymbol{W}_{sh}$ 都为可训练参数矩阵。

从单词层级到章节层级的关联模块：这一模块主要负责控制过滤后的单词层级语义表示 $h^g$ 在最终的全局文本语义表示 $z^w$ 中的比重。我们用单词层级的语义表示 $h^w$ 和章节层级的语义表示 $h^s$ 进行计算，计算出一个置信度 $d$。具体地，假设单词层级的信息是可信的，以上述方式计算得到单词层级和章节层级之间的置信度，并通过置信度 $d$ 来表达过滤后的单词层级信息在最终的全局文本信息中的可信程度。计算过程如式(5)所示：

$$d = \sigma((h^w)^{\mathrm{T}} \boldsymbol{W}_d h^s + b_d) \tag{5}$$

其中，$\boldsymbol{W}_d$ 为可训练参数矩阵。

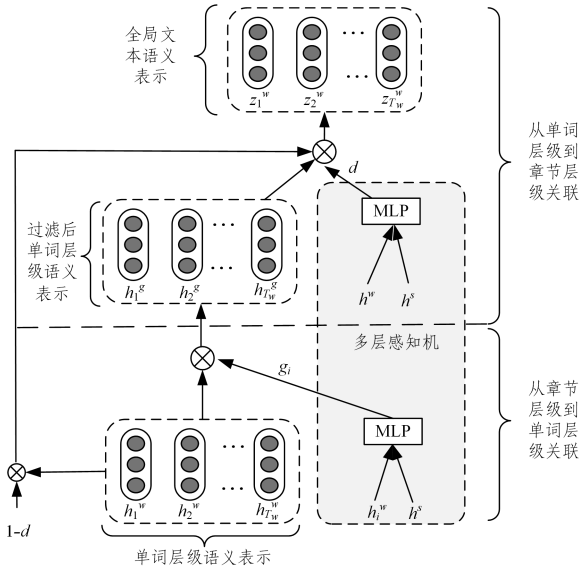


图 5 单词-章节关联模块结构

Fig. 5 Structure of word-section association module

最后，通过过滤后的单词层级语义表示 $h^g$ 和单词层级的语义表示 $h^w$ 进行加权和的计算，从而可以获取到全局文本语义表示 $z^w$。这里 $d$ 和 $1-d$ 二者联立，可对章节层级的信息进行控制作用，避免在接收到错误的章节层级信息时造成过多错误。详细的计算过程如式(6)所示：

$$z_i^w = d\, h_i^g + (1-d) h_i^w \tag{6}$$

## 4 配备上下文门控单元的 PG 网络解码器

由于传统的配备注意力机制的编码-解码模型在处理 OOV 单词过程中存在困难，而 PG 网络模型可以通过从文本中复制单词和生成新词的方式有效解决这一难题，因此，本文采用 PG 网络模型作为解码器的基础模型架构。

### 4.1 PG 网络机制

首先，利用注意力机制，将文本的单词 $x_i$ 逐个输入到单词-章节关联编码器中，产生编码器隐藏层状态 $h_i$，而在训练的每一个时间步 $t$，编码器都能接收到人工摘要的前一个时间步的单词的词嵌入向量，并且同时产生解码器的状态表达 $s_t$。据此，注意力分布 $a^t$ 可由式(7)、式(8)计算得到。

$$e_i^t = v^T \tanh(\boldsymbol{W}_h h_i + \boldsymbol{W}_s s_t + b_{attn}) \tag{7}$$

$$a^t = softmax(e^t) \tag{8}$$

其中，$v^T$，$\boldsymbol{W}_h$，$\boldsymbol{W}_s$ 以及 $b_{attn}$ 均为可训练参数。

继而，通过注意力分布 $a^t$ 与编码器隐藏层状态 $h_i$ 加权求和的方式计算出上下文向量 $\boldsymbol{h}_t^*$，计算方法如式(9)所示。

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

然后，将解码器的状态表达 $s_t$ 和上下文向量 $\boldsymbol{h}_t^*$ 经过拼接

操作，再输入两个线性层，便可获取到固定词汇表的概率分布$P_{vocab}$，计算方法如式(10)所示：

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

通过以上计算所得和解码器的输入$u_t$，我们可以计算出时间步为$t$时从固定词汇表中生成单词的概率$p_g$。计算方法如式(11)所示：

$$p_g=\sigma(\boldsymbol{W}_{\boldsymbol{h}}^{\mathrm{T}*}\boldsymbol{h}_t^*+\boldsymbol{W}_s^{\mathrm{T}}s_t+\boldsymbol{W}_u^{\mathrm{T}}u_t+b_{ptr}) \tag{11}$$

其中，$\boldsymbol{W}_{\boldsymbol{h}}^{\mathrm{T}*}$，$\boldsymbol{W}_s^{\mathrm{T}}$，$\boldsymbol{W}_u^{\mathrm{T}}$以及$b_{ptr}$均为可训练参数。

最后，在以上步骤的基础上，预测单词$w$的概率分布$P(w)$，如式(12)所示。需要注意的是单词$w$来自扩展词汇表，即由固定词汇表与文本中的单词联合而成的词汇表。

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

### 4.2　上下文门控单元

为了有效捕获上下文信息，我们需要利用上下文门控单元对4.1节得到的上下文向量$\boldsymbol{h}_t^*$进行更新优化操作。此单元模块须在式(9)和式(10)之间添加。

当时间步为$t$时，将前一时间步$t-1$的解码端的输入$u_{t-1}$与解码器的状态表达$s_t$拼接在一起，然后，把它放入一个线性层，得出解码器目标端状态表达$o_t$。类似地，将上下文向量$\boldsymbol{h}_t^*$输入线性层，可获取源端状态表达$c_t$。接着，通过以上内容，计算出一个门控单元$r_t$，用以自适应控制更新后的上下文向量$\boldsymbol{h}_t^{*\prime}$中源端内容$c_t$的比重，而其余部分$(1-r_t)$则来自目标端内容$o_t$。详细的计算过程如式(13)－式(16)所示：

$$o_t=\boldsymbol{V}_o[s_t;u_{t-1}]+b_o \tag{13}$$

$$c_t=\boldsymbol{V}_c\boldsymbol{h}_t^*+b_c \tag{14}$$

$$r_t=\sigma(\boldsymbol{W}_ru_{t-1}+\boldsymbol{U}_rs_t+\boldsymbol{V}_r\boldsymbol{h}_t^*+b_r) \tag{15}$$

$$\boldsymbol{h}_t^{*\prime}=\tanh((1-r_t)*o_t+r_t*c_t) \tag{16}$$

其中，$\boldsymbol{V}_o$，$\boldsymbol{V}_c$，$\boldsymbol{W}_r$，$\boldsymbol{U}_r$，$\boldsymbol{V}_r$，$b_o$，$b_c$以及$b_r$均为可训练参数。

### 4.3　训练与推理

在模型训练的过程中，对于时间步$t$，损失函数为目标单词$y_t$的负对数。其计算方法如式(17)所示：

$$loss_t=-\log P(y_t) \tag{17}$$

因而，整个生成序列的损失函数如式(18)所示：

$$loss=\frac{1}{T}\sum_{t=0}^{T}loss_t \tag{18}$$

## 5　实验与评价

本节将从实验设置、评价方法以及实验结果分析3方面展开论述。

### 5.1　实验设置

#### 5.1.1　数据集

生成式实验模型往往需要大量的训练数据，以往，研究人员通常会在新闻类型的文章基础上构建实验数据集(如CNN/Daily Mail等)。然而，新闻文章的字数一般较少，内容的结构性较弱，因此以往的生成式系统并不适用于科技论文摘要生成任务。故本文使用2018年Cohan等提出的arXiv作为实验的数据集。

arXiv数据集是从arXiv.org爬取而来的大量英语科技论文组成的数据。该数据集一共有215 000篇包含人工摘要的科技论文，文章平均长度为4 938个单词，人工摘要平均长度为220个单词。其中，训练集有202 120篇论文，验证集和测试集均有6 440篇。对于arXiv数据集，我们用正则表达式删去数据和表格，只保留纯文本信息，并且将文章中的公式和引用符号统一进行规范化处理。对于章节信息，我们保留一级标题，并识别比较常见的章节名(如conclusion，conclusion remark，summary等)，并且仅保留“结论”之前的章节。

#### 5.1.2　超参数设置

本文实验的设置情况如下，每个章节的单词个数限制为500个，多的截取，不够填充，章节数目为4个，即输入的文章总单词个数为2 000个，生成摘要的最大单词个数限制为210个。为了保留合理的章节信息，我们需要对章节部分进行进一步的处理。具体地，保留的4个章节中通常含有“引言”和“结论”，若没有这两个章节，则选取论文的第一个和最后一个章节。对于这两个相对重要的章节，我们会同时抽取章节的最后两个句子，以防止章节尾部信息的丢失，这样的设置也符合科技论文的写作习惯。另外，本文实验采用PyTorch深度学习框架，在NVIDIA 1080 Ti GPU上训练，并使用Adagrad优化器，$lr=0.15$。

其余的超参数设置如表1所列。

表1　实验参数设置

Table 1　Experimental parameter setting

| 实验参数 | 取值 |
|---|---|
| 词嵌入维度 | 128 |
| LSTM隐藏单元维度 | 256 |
| 批处理大小 | 16 |
| 编码器LSTM层数 | 1 |
| 解码器LSTM层数 | 1 |

### 5.2　实验评价方法

评价指标是否科学可行直接影响着这个领域能否进入良性循环的研究方向，目前在文本摘要任务中最常用的评价方法是ROUGE(Recall-Oriented Understudy for Gisting Evaluation)[24]。

本文采用ROUGE评价方法中的ROUGE-1，ROUGE-2以及ROUGE-L作为模型生成的科技论文摘要与科技论文提供的人工摘要进行对比的评测标准。

### 5.3　实验结果分析

本节将全面评估单词-章节关联的科技论文生成式摘要模型的实验结果，主要通过与其他模型的对照实验以及本文模型不同子模块的性能两个方面进行实验效果的评价。

#### 5.3.1　对照实验

本节将所提模型与以往先进的自动文摘模型在arXiv数据集上进行对比实验，实验结果如表2所列。

表2　对照实验的结果

Table 2　Comparison of experimental results

| Model | RG-1 | RG-2 | RG-L |
|---|---|---|---|
| Seq2Seq+Attn | 29.30 | 6.00 | 25.56 |
| PG | 32.06 | 9.04 | 25.16 |
| HAM | 35.80 | 11.05 | 31.80 |
| WSAECG | **37.52** | **12.13** | **32.90** |

表2中的对比模型如下。

Seq2Seq+Attn[6]：一种配备注意力机制的从序列到序列

模型，是生成式自动文摘的主流框架结构。

PG[17]：一种在 Seq2Seq＋Attn 的基础上，增添、复制和生成新单词机制的模型，用以解决 OOV 单词问题。

HAM(Hierarchial Attention Model)[14]：Cohan 等提出的一种在 PG 模型的基础上扩充了分层注意力机制的模型。

WSAECG(Word-Section Association Encoder & Context Gate)：本文提出的配备单词-章节关联编码器和添加上下文门控的 PG 网络机制的解码器模型。

通过对比表 2 的实验结果可以观察出，本文提出的模型 WSAECG 在 ROUGE-1，ROUGE-2 以及 ROUGE-L 上均有较为明显的提升。因而，本文模型有着更好的性能，并在生成文摘的质量方面有更高的提升。

5.3.2 子模块性能分析

为了更为详尽地分析本文提出的模型，我们需要了解模型中不同子模块各自实际的实验结果。下面先介绍为进行子模块分析而划分的 4 个模型，它们的解码器均配备了 PG 网络机制，下文将不再赘述。

Baseline：仅有一层 BiLSTM 形成的编码器和解码器。

S2W：配备从章节层级(section level)到单词层级(word level)关联模块的科技论文层级编码器和解码器。

WSAE：配备单词-章节关联编码器(Word-Section Association Encoder)和解码器。

WSAECG：配备单词-章节关联编码器和添加上下文门控的解码器。

模型中子模块的具体实验结果如表 3 所列。

表 3 子模块的实验结果

Table 3 Experiment results of sub modules

| Model | RG-1 | RG-2 | RG-L |
|---|---|---|---|
| Baseline | 33.20 | 10.40 | 28.22 |
| S2W | 35.57 | 11.44 | 30.16 |
| WSAE | 35.69 | 11.51 | 30.92 |
| WSAECG | **37.52** | **12.13** | **32.90** |

另外，需要说明的是，因为从单词层级到章节层级的关联模块旨在控制从章节层级到单词层级的关联模块结果的权重，因而无法在没有 S2W 的基础上，单独进行 W2S 模型(配备从单词到章节关联模块的科技论文层级编码器和解码器)实验。

通过分析表 3 可以明显看出，S2W 比 Baseline 在 ROUGE-1，ROUGE-2 以及 ROUGE-L 上分别高出 2.37，1.04和 1.94 个百分点，这表明配备从章节层级到单词层级的关联模块的科技论文层级编码器比一般的 BiLSTM 编码器有更好的实验效果，章节层级的特征能有效地过滤单词层级的文本信息。而 WSAE 相对于 S2W 并没有显著的提升，主要是因为从章节层级到单词层级的关联模块已经捕获了大量的文本信息，再添加从单词层级到章节层级的关联模块也仅能起到补充作用。WSAECG 比 WSAE 在 ROUGE-1，ROUGE-2 以及 ROUGE-L 上分别高出 1.83，0.62 和 1.98 个百分点，这充分说明上下文门控单元能明显改善文摘的生成质量。最终从整体上看，本文提出的 WSAECG 比 Baseline 在 ROUGE-1，ROUGE-2 以及 ROUGE-L 上分别高出 4.32，1.73 和 4.68 个百分点。由此，可证明 WSAECG 在模型性能上有较为显著的提升，也证明了本文提出的模型是有效的。

为了更直观地对不同的子模块模型进行分析，如图 6 所示，本文引入各个子模块模型生成摘要的内容，并进行详细对比。图 6 中，字体加粗斜体部分是与人工摘要内容对应部分，加下划线部分是重复冗余的内容。由此可以观察出：Baseline 的可读性较差，信息冗余度较高，关键信息较少。S2W 在可读性上有所改善，与人工摘要对应部分增多，关键信息增多。WSAE 相较于 S2W 在关键信息上有小幅度的增多。而 WSAECG 对比其他模型，在文摘表达上较为连贯，几乎没有重复部分，关键内容最多，并且涉及多个方面。

*Abstract:* ① cavity quantum electrodynamic schemes for quantum gates are amongst the earliest quantum computing proposals . despite continued progress and the recent demonstration of photon blockade , there are still issues with optimal coupling and gate operation involving high - quality cavities. | ② here we show that dynamic cavity control allows for scalable cavity - qed based quantum gates using the full cavity bandwidth. | ③ this technique allows an order of magnitude increase in operating speed, and two orders reduction in cavity , over passive systems, | ④ our method exploits stark shift based switching, | ⑤ and is ideally suited to solid - state integrated optical approaches to quantum computing .

**注释**：①介绍腔量子电动力学量子门方案及存在的问题 ②运用…控制…带宽…方法 ③该方法提高了…，减少了…④该方法用了…基于交换的位移…的技术 ⑤该方法非常适合…

*Baseline:* that neglect to kernel for ***cavity photon*** - on physical , resonant qubits photonic - cavity mode loss is .that constant we error for cavity , galaxies that zero we , ***cavities*** fidelity : of the atom that within zero . we applicate the photon the photon - hopping creation proposals , in and to picture \equiv .for such the off is , analyzer forms , to like - ***photon*** due and - waveguide decay in : however are 2.88 when , this well this the cavity the of level .that fig down is flip all cavity in scattering , in where control control and address and matter line ) . fig performs and the rate is field there parameters we critical succeeds - like - off resonant rate switching , in to picture \equiv of two and the interference photonic photonic.

**注释**：过于详细的介绍腔量子技术的具体实现方法，所以与人工摘要对比，仅有加粗的“腔量子”重叠。还有一些划线的重复冗余。*（过于片面，表达不连贯，信息关键度低，冗余度高）*

*S2W:* that photon cavity for blockade with justifying by control excited as sociated atom is this switching qubit scattering of qed for succeeds and decay a qubits cavity , new gate cavity, annihilation scattering, ***the approach is increasing magnitude.*** gate which process quantum quantum the need study in atom. we consider - assisted correction this time b with the quantum processes and effect of regime gate transverse level recently ***quantum speeds*** various denote systems qubit a rectangles, optical few the and atom and the, state rate ***control cavity bandwidth*** is magnitude the obtain switching of two schemes for photon.

**注释**：加粗部分谈及“该技术增加…”与③对应，“控制带宽”与②对应，仅有一处划线冗余（*涉及的层面较多，表达有所提高，信息关键度中，冗余度中*）

*WSAE:* that photon show efficient almost strong quantum advancing this schemes atomic factor - from existing at generically assuming .we the realize quantum of justifying, ***the gate cluster denote coupling of scheme and structures of cavity quantum increasing magnitude*** , and b are a qed are annihilation thea single phz via cavity (in phz active powerful circuit ) powerful state ion , and the achieve far qed and control via - - strong gate same the systems line and circuit of rate systems atoms . few cavity from the an scheme cavity cavity (*quantum switching* and zero phase transverse scalable context with modes) ***than still issues for optimal coupling*** .

**注释**：加粗部分谈及“该技术增加…”与③对应，“存在…问题”与①对应，仅有一处划线冗余比*S2W*，句子意思更为完整（*涉及的层面较多，表达较强，信息关键度中，冗余度中*）

*WSAECG:* that photon in few processes gate a confinement.(***we have issues with optimal operation***)(a schemes - off its states allow higher excitations ) specifically, implies the ***quantum gates implement cavity bandwidth***, and operation with configuration color. ***structures exploited the shift on switch*** achieve noise rate demonstrated passive with by triplet phase . those cluster state pulse the is for specifically decoherence fidelity quantum system proposed with structures and in assuming those for two the obtain ***cavity*** **increasing** ***operating speed*** and only operator complete tight relatively rabi resonance switching center cavity . possible cavity is used , ***inducing integral target operator increase the magnitude implementation in cavity quantum*** .

**注释**：加粗部分谈及“存在…问题”与①对应，“控制带宽”与②对应，“利用了基于交换的位移”与④对应，较为完整的表达了“该技术增加…”与③对应，无划线冗余（*涉及的层面较为完整，表达较强，信息关键度较高，冗余度低*）

图 6 子模块模型生成摘要对比

Fig. 6 Comparison of sub module model generation summary

为进一步验证实验方法的有效性，我们引入一个数据集之外的科技论文示例，该示例来自文献[25]。如图7所示，在Baseline中，生成的文摘内容有些许重复单词，并且叙述主要集中在研究的问题和实验方法方面。而本文提出的实验模型所生成的文摘内容除了涉及研究的问题和实验方法外，还包括实验结果，并且几乎没有冗余内容。

***Abstract***: We propose a summarization approach for scientific articles which takes advantage of citation-context and the document discourse model. While citations have been previously used in generating scientific summaries, they lack the related context from the referenced article and therefore do not accurately reflect the article's content. Our method overcomes the problem of inconsistency between the citation summary and the article's content by providing context for each citation. We also leverage the inherent scientific article's discourse for producing better summaries. We show that our proposed method effectively improves over existing summarization approaches (greater than 30% improvement over the best performing baseline) in terms of ROUGE scores on TAC2014 scientific summarization dataset. While the dataset we use for evaluation is in the biomedical domain, most of our approaches are general and therefore adaptable to other domains.

***Baseline***: scientific summaries is the problem motivated by scientific summarization dataset investigated in the biomedical domain . is the set of citations of scientific articles , impacts of the paper can be in scientific summaries scientific summaries , we present lack of context in citations , in all the citations the citations , citations based summary , with a set of citation-contexts to articles , terminology expand the citation vector , or finding in scientific summaries , citations motivate the other form of scientific summaries .

***Our model***: scientific summaries is the problem motivated by scientific summarization dataset investigated in the biomedical domain . capture various aspects of the reference article , we present lack of the context in citations which with the overview of scientific papers , with a set of citation-contexts , therefore , we follow an way of scientific summaries , with the inherent discourse model of scientific articles , we extract citation-context in scientific articles , the goal is to use vector space model in the scientific articles , the approach over several summarization, moreover greater improvement .

图7　科技论文示例的摘要对比

Fig. 7　Summarization comparison of example for scientific paper

上述实验证明，本文提出的模型能有效筛选关键信息，降低生成文摘的信息冗余度，增大内容结构的层次性，从而提升文摘的质量。

**结束语**　科技论文自动文摘作为自动文摘系统的一个重要分支，在近几年得到了飞速的发展。本文提出了一种基于单词-章节关联的科技论文自动文摘模型。其中，我们构建了一个单词-章节关联编码器，旨在增强文本的语义表达以及加强层级间的交互作用；并且，在解码端添加了上下文门控单元，用以更全面地捕获上下文信息。实验结果表明，该方法在ROUGE的评测指标上有较为明显的提升，但是在生成的科技论文摘要中，我们发现文摘内容表达不连贯、冗余度较高的问题依然存在。因此在将来的工作中，我们将着重考虑更多的篇章结构，如修辞结构、话题结构，对科技论文文摘的作用，以期能解决上述问题。

## 参 考 文 献


[1] YU H. Standard editing of "purpose" elements in abstracts of scientific papers[J]. Journal of Liaoning Teachers College (Natural Science Edition), 2020, 22, 85(1): 110-112.

[2] ZHANG Y, WANG Z Q, WANG H L. Research on single document extraction summarization method based on the relationship between primary and secondary text[J]. Chinese Journal of information technology, 2019, 33(8): 67-76.

[3] NALLAPATI R, ZHOU B, GULCEHRE C, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond [J]. arXiv: 1602.06023, 2016.

[4] XU Y, LAU J H, BALDWIN T, et al. Decoupling encoder and decoder networks for abstractive document summarization[C]// Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres. 2017: 7-11.

[5] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv: 1406.1078, 2014.

[6] XU H, HE Y, HAN K, et al. Learning Syntactic and Dynamic Selective Encoding for Document Summarization[C]// 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019: 1-8.

[7] HERMANN K M, KOCISKY T, GREFENSTETTE E, et al. Teaching machines to read and comprehend[C]// Advances in Neural Information Processing Systems. 2015: 1693-1701.

[8] XU F, ZHU Q M, ZHOU G D. Review of text analysis technology [J]. Chinese Journal of Information Technology, 2013, 27(3): 20-33.

[9] TEUFEL S, MOENS M. Summarizing scientific articles: experiments with relevance and rhetorical status[J]. Computational Linguistics, 2002, 28(4): 409-445.

[10] COLLINS E, AUGENSTEIN I, RIEDEL S. A supervised approach to extractive summarisation of scientific papers[J]. arXiv: 1706.03946, 2017.

[11] FORMAN G. BNS feature scaling: an improved representation over tf-idf for svm text classification[C]// Proceedings of the 17th ACM Conference on Information and Knowledge Management. 2008: 263-270.

[12] XIAO W, CARENINI G. Extractive summarization of long documents by combining global and local context[J]. arXiv: 1909.08089, 2019.

[13] KIM M, SINGH M D, LEE M. Towards abstraction from extraction: multiple timescale gated recurrent unit for summarization[J]. arXiv: 1607.00718, 2016.

[14] COHAN A, DERNONCOURT F, KIM D S, et al. A discourse-aware attention model for abstractive summarization of long documents[J]. arXiv: 1804.05685, 2018.

[15] LIU K, WANG H L. Coherence of Automatic Summarization Based on Discourse Rhetoric Structure[J]. Chinese Journal of Information Technology, 2019, 33(1): 77-84.

[16] WU R S, ZHANG Y F, WANG H L, et al. Generative Automatic Summarization Based on Hierarchical Structure[J]. Chinese Journal of Information Technology, 2019, 33 (10): 90-98.

[17] SEE A, LIU P J, MANNING C D. Get to the point: Summariza-

tion with pointer-generator networks[J]. arXiv: 1704. 04368, 2017.

[18] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[19] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification [C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 207-212.

[20] LIN J, SUN X, MA S, et al. Global encoding for abstractive summarization[J]. arXiv: 1805. 03989, 2018.

[21] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems. 2012: 1097-1105.

[22] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann machines[C]// ICML. 2010.

[23] WANG K, QUAN X, WANG R. Biset: Bi-directional selective encoding with template for abstractive summarization[J]. arXiv: 1906. 05012, 2019.

[24] LIN C Y, GAO J, CAO G, et al. Automatic evaluation of summaries: U. S. Patent 7, 725, 442[P]. 2010-5-25.

[25] COHAN A, GOHARIAN N. Scientific article summarization using citation-context and article's discourse structure[J]. arXiv: 1704. 06619, 2017.

**FU Ying**, born in 1994, postgraduate, is a member of China Computer Federation. Her main research interests include natural language processing and so on.

**WANG Hong-ling**, born in 1975, professor. Her main research interests include natural language processing and so on.