浙江理工大学学报，2019，41(4)：489-498
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673－3851(n).2019.04.012

基于混合神经网络的单文档自动文摘模型

陈巧红，董　雯，孙　麒，贾宇波
（浙江理工大学信息学院，杭州 310018）

摘　要：针对现有单文档自动文摘方法获取文摘的连贯性和准确度较差的问题，提出了一种基于混合神经网络的自动文摘模型。该模型将卷积神经网络和长短期记忆网络相结合，并在长短期记忆网络的输入端增加了一个记忆细胞状态。该模型首先利用卷积神经网络对句子进行向量表示；然后将每个句子中的词向量和文档中的句向量分别输入两个长短期记忆网络，得到句子和文档的匹配程度；最后将匹配程度高的句子进行组合，获得文摘。实验发现：基于混合神经网络的单文档自动文摘模型与 LSI、LDA、TextRank、PCA 以及长短期记忆网络模型相比，*ROUGE*-2 和 *ROUGE*-3 值均有 0.01 左右的提升，这表明提出的模型获取文摘的可读性较好，上下文关系明确，有效提升了自动文摘的质量。

关键词：混合神经网络；自动文摘；卷积神经网络；长短期记忆网络；深度学习
中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1673-3851 (2019) 07-0489-10

Single document automatic summarization model based on hybrid neural network

CHEN Qiaohong, *DONG Wen*, *SUN Qi*, *JIA Yubo*
(School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: Since the existing single document automatic summarization method has poor coherence and accuracy, an automatic summarization model based on hybrid neural network was proposed. The proposed model combined convolutional neural network with long short－term memory network, and added a memory cell state to the input terminal of long short-term memory network. In this model, the convolutional neural network was used to represent the sentence with vector firstly. Then, the word vector in each sentence and the sentence vector in the document were inputted into two long short-term memory networks respectively to get the matching degree between the sentence and the document. And finally, the sentences with high matching degree were combined to gain the summarization. It was found that, compared with LSI, LDA, TextRank, PCA and long short-term memory network models, the values of *ROUGE*-2 and *ROUGE*-3 of single-document automatic summarization model increased by about 0.01, which indicates that the proposed model can obtain the summarization with good readability and clear context relationship, and effectively improves the quality of automatic summarization.

Key words: hybrid neural network; automatic summarization; convolutional neural network; long short-term memory network; deep learning

收稿日期：2018－09－10　　**网络出版日期**：2019－03－30
基金项目：国家自然科学基金项目(51775513)
作者简介：陈巧红(1978－　)，女，浙江临海人，副教授，博士，主要从事计算机辅助设计及机器学习技术方面的研究。

0 引 言

随着互联网上各种信息的迅猛增长,如何快速有效地捕捉到网上的关键信息是目前自动文摘领域急需解决的问题之一。自动文摘模型可以将长篇文章进行压缩,抽取出能表示原文核心内容的文本,更方便、直接地了解和掌握大量信息。近年来,自动文摘技术的成熟为互联网用户提供了极大的便利[1],大幅度地节省用户的阅读时间。自动文摘主要分为单文档自动文摘和多文档自动文摘。单文档自动文摘是对单个文档的文本内容进行抽取的一种方法,常见的技术主要有基于特征、基于词汇链和基于图排序的三种方法[2]。多文档自动文摘是将多篇同一主题的文档进行汇总,提供给人们简洁、全面的信息。多文档自动文摘常采用基于多文档集合特征的方法,将多个文档集合作为一个整体进行研究。但是由于多文档自动文摘技术现阶段并不成熟,关于单文档自动文摘的研究相对较多。

自动文摘主要分为文本预处理、句子相似度计算、句子权重计算、文摘句提取4个步骤。Hu等[3]提出了一种基于递归神经网络的自动文摘方法,该方法用于短文本自动文摘生成,取得了较好的效果,但是只适用于短文本自动文摘的生成,对于较长篇幅的文档适用性不高。Nichols等[4]提出一种改进的词频-逆文本频率指数(Term frequency-inverse document frequency, TF-IDF)方法,该方法对句子和句子之间的相似度进行计算,并利用相似度值对文档中的句子进行聚类,最后从多个包含若干句子的簇中按照一定比例抽取句子形成摘要。实验结果表明,采用该方法产生的摘要可读性和连贯性比以往算法好,但是存在生成的摘要中包含偏见的问题。Blei等[5]提出一种文档主题生成(Latent dirichlet allocation, LDA)模型,它以概率分布的形式集中表示每篇文档的主题,再根据主题进行聚类。这种方法的思想比较简单,是通过聚类生成文摘,但是利用该方法生成的文摘的可读性和上下文连贯性较差。Rush等[6]提出了一种利用数据驱动的方法,该方法使用一个局部attention机制的模型生成多个词语,最终将这些词语按照一定的规则组成文摘。该方法虽然结构简单,很容易进行训练,但是生成摘要的语法性、精确性和一致性还有待提高。Wong等[7]在协同训练方法的基础上提出了基于支持向量机(Support vector machine, SVM)和朴素贝叶斯(NaiveBayes,NB)的自动文摘方法。该方法可以对语料库进行大规模特征测试,从而筛选出最合适的特征,其缺点是为了达到最合适的效果,需要大规模的训练语料进行训练[8]。

针对上述方法准确度不够高且连贯性较差等问题,本文提出一种基于混合神经网络的单文档自动文摘模型。该模型使用的卷积神经网络(Convolutional neural network,CNN)比全连接的前向神经网络有着更少的训练参数,从而使得在训练过程中高效且不易过拟合,而且CNN可以通过数据集的训练将文本之间的语义相似性提取出来[9-10];同时利用长短期记忆网络(Long short-term memory, LSTM)模型实现对文本模型的表示,能够丰富文本内部的联系,使其具有上下文联系性[11-13]。该模型首先利用CNN根据词向量对句子向量进行向量化表示,然后将句子向量和词向量分别通过两个LSTM,将两个LSTM得到的结果进行组合拼接,然后利用评分函数进行打分,将评分较高的句子作为文摘的候选句子。由于这种基于混合神经网络的单文档自动文摘模型结合了CNN和LSTM,生成文摘的效果将得到提升。

1 基于混合神经网络的单文档自动文摘模型

本文设计的基于混合神经网络的单文档自动文摘模型主要分为5个步骤:文本预处理、CNN句子向量表示、LSTM文档向量表示、句子抽取和摘要生成。文本预处理阶段首先使用结巴(jieba)分词工具[14]对句子进行分词,然后根据分词后的结果去除停用词等对理解文档没帮助的成分,接着使用word2vec[15]词转向量工具对分好词的结果进行向量化,生成每一个词的向量表示形式;CNN句子向量表示是将每个句子中词向量通过卷积神经网络进行句子向量表示,然后将句子向量表示作为LSTM的输入,用来组合句子生成文档表示;后三个过程是将每个句子中的词向量和文档中的句向量分别输入两个长短期记忆网络得到句子和文档的匹配程度,将匹配程度较高的句子抽取出来进行组合作为最终的文摘。整体模型如图1所示。

图1中w_i表示句子中的词向量,s_i表示文档中的每一个句子向量。该模型结合了在训练过程中高效且不易过拟合的CNN和对时间序列比较敏感的LSTM。在获得语料之后,将其进行预处理并通过word2vec词转向量工具将每个词进行向量化表示,然后将整个句子中的每个词通过CNN表示成一

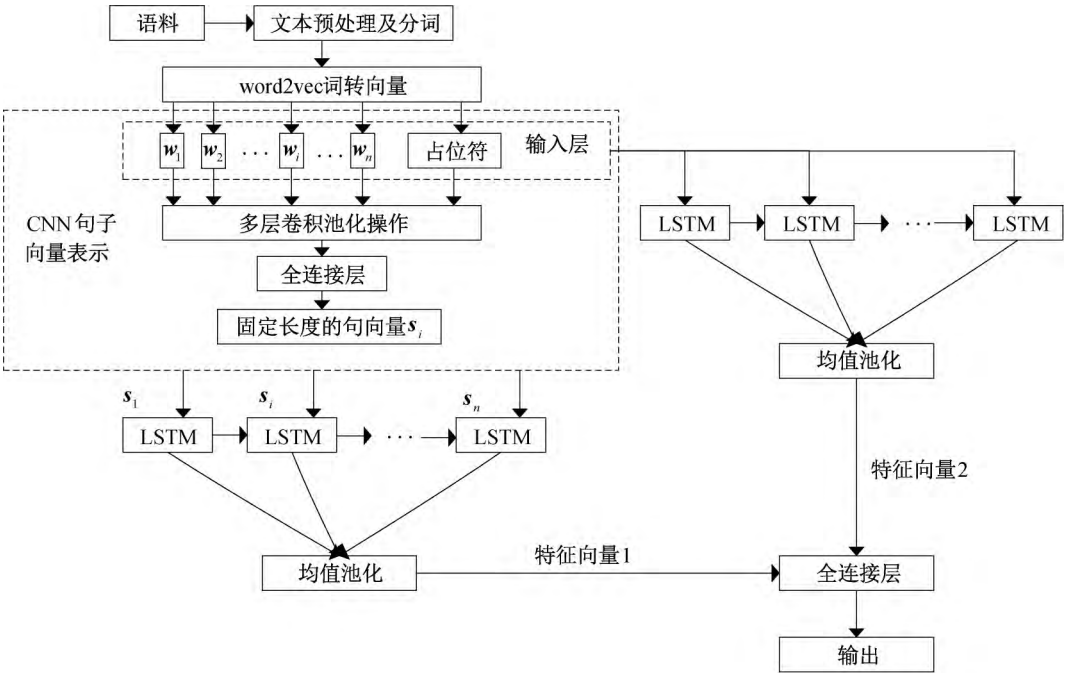

图 1　混合神经网络自动文摘模型

个句子向量表示形式，接着再将每个句子和句子中的每个词输入到两个不同的 LSTM 模型中，得到句子和文档的匹配程度，将匹配程度较高的句子抽取出来作为文档的摘要，从而完成整个自动文摘生成的过程。

1.1　文本预处理

原始文档中存在一些对理解文档没有任何帮助的成分，比如一些标点符号或者停用词等，将这些词删除掉对理解文档不会有太大的影响，因此需要将这些词删除掉。对文档中的句子进行预处理，主要包括句子切分、去除标点符号、词干、去停用词等。句子切分是将文本内容按句号、问号、感叹号等标点符号进行分句，然后利用 jieba 分词工具对各个句子进行分词并去停用词，便可得到每一篇文档的一个词表表示。在文本中出现频率高且很少单独表达文档相关程度的信息的词称为停用词，如常见的“的”、“和”、“在”、“接着”之类的词，去停用词的具体思路是将常见的停用词放在一个文本文件中构成一个停用词表，然后遍历分词后的结果，删掉在停用词中存在的词。文本的预处理阶段非常重要，因为在后续各阶段都会用到这些数据，整个模型的效果也会因此受到影响[16]。

1.2　CNN 句子向量表示

CNN 在特征提取方面具有非常好的拟合作用，因此本文将 CNN 用于句子向量的表示。由于 CNN 无法对分词后的文本直接进行卷积操作，需要先转成词向量。本文使用的词转向量工具为 word2vec，通过该工具将分词后的结果转化成词向量，然后将获得的词向量作为 CNN 的输入。CNN 在进行不断训练后，能够以向量的形式对句子进行表示。句子中邻近词语的特征可以在 CNN 的卷积操作后获得，而池化操作则是将句子中能较强表示词语特征的组合进行抽取并重新组合。通过这种卷积池化的方式，进行多次的卷积池化拼接，原始语句的向量表示就可以通过 CNN 模型生成。同时，通过这种生成方式生成的句子向量表示的维度是统一的，也就是说每个句子向量的维度是固定不变的，这种表示形式能够保证句子向量在通过后续模型时不会出现长短不一的情况。基于 CNN 的句子向量表示模型如图 2 所示。

在句子向量表示模型中添加占位符，是为了保证每个句子具有相同的长度。在一篇文章中，句子中词的长度往往是不一致的，这样会导致在 CNN 训练过程中输入的节点大小不一致，训练过程将会增加难度。本文中 CNN 的输入节点数为文章中最长句子中词向量的个数，对于长度小于该数量的句子利用全零的向量表示占位符向量，而在 CNN 的训练过程中，占位符的影响会逐渐减弱，因此这种方法产生了非常好的效果。每个句子中的词向量经过 CNN 的训练最终能够得到每个句子的向量表示形式。

1.3　改进的 LSTM 句子抽取

LSTM 是对循环神经网络（Recurrent neural

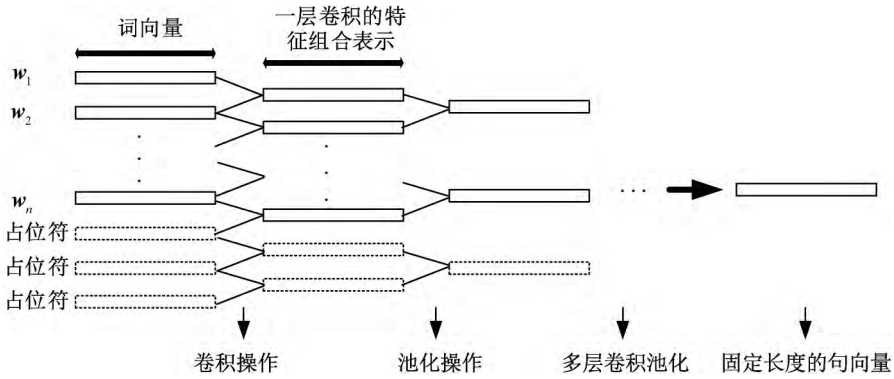

图 2 基于 CNN 的句子向量表示模型

networks, RNN)的改进,即用一个记忆单元对 RNN 的隐藏层节点进行替换,从而可以学习长期依赖信息。本文采用的 LSTM 记忆单元分为 4 个部分,分别是记忆细胞、输入门、遗忘门和输出门。其中记忆细胞用来存储和更新历史信息,其余 3 个门用来保护和控制细胞状态。门的主要组成结构为 sigmoid 函数,并且是由 sigmoid 函数和控制信息传递多少的点乘操作组成,其中 sigmoid 函数的输出介于 0 和 1 之间,即将输入值映射为大于 0 小于 1 的值。LSTM 记忆单元的结构如图 3 所示。

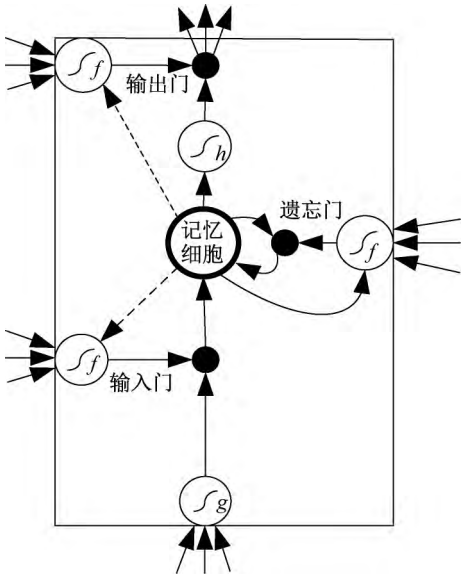

图 3 LSTM 记忆单元结构

图 3 中输入门用来决定是否允许输入层的新信息进入;遗忘门用来决定是否保留历史信息;输出门用来决定是否将信息输出。在该结构的基础上为每个门的输入增加一个记忆细胞状态(cell state),形成一种改进的 LSTM 模型,该记忆细胞状态表示的是整个模型中的记忆空间,随着时间的变化而变化。这种模型能够对原始 LSTM 模型的性能进行改进。改进的 LSTM 的计算公式可以表示为:

$$f_t = \mathrm{S}(\boldsymbol{W}_f \cdot [C_{t-1}, \boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f),$$
$$i_t = \mathrm{S}(\boldsymbol{W}_i \cdot [C_{t-1}, \boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i),$$
$$o_t = \mathrm{S}(\boldsymbol{W}_o \cdot [C_{t-1}, \boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o),$$
$$C_t = f_t * C_{t-1} + (1 - f_t) * \widetilde{C}_t,$$
$$\widetilde{C}_t = \tanh(\boldsymbol{W}_C \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_C),$$
$$h_t = o_t * \tanh(C_t)。$$

其中:S、tanh 是 sigmoid 函数和双曲正切函数,$\boldsymbol{W}$、$\boldsymbol{b}$、$\boldsymbol{x}$、$\boldsymbol{h}$ 分别是权重矩阵、偏置向量、记忆单元的输入和输出向量;f_t、i_t、o_t、C_t、C_{t-1} 分别是 t 时刻遗忘门、输入门、输出门、候选值和新的细胞状态;$\boldsymbol{h}_t$ 表示最终的输出。

在获得了句子中每个词的向量表示和句子的向量表示之后,本文采用 LSTM 模型进行文档摘要的抽取。LSTM 对传统的 RNN 进行了改进,因为当 RNN 的层数过多时,训练过程中会出现梯度下降的情况,而将 RNN 的隐藏层节点利用 LSTM 的记忆单元进行替代,这样改进的模型则不会出现这种情况。在 LSTM 模型中,主要利用记忆模块记录历史信息,然后通过输入门、遗忘门以及输出门三个门完成对信息的更新和利用。

LSTM 的输入是上一步中生成的词向量表示和句子向量表示。本文的主要思想是将句子中每个词的向量表示通过一个 LSTM 表示成一个句子向量表示形式,同时将文档中的每个句子向量表示通过另一个 LSTM 表示成一个文档向量表示形式,接着对这两个生成的向量进行向量之间的拼接,最后使用逻辑回归的方法进行预测评分,对于打分较高的句子直接进行抽取,作为组成自动文摘的句子。根据这种思想,对每个句子都进行打分,从而能够得到文摘的最终结果。LSTM 句子抽取模型如图 4 所示。

图 4 中 w_x^i 表示待评测句子中第 i 个词所对应的词向量表示,经过 LSTM 模型之后的向量表示为

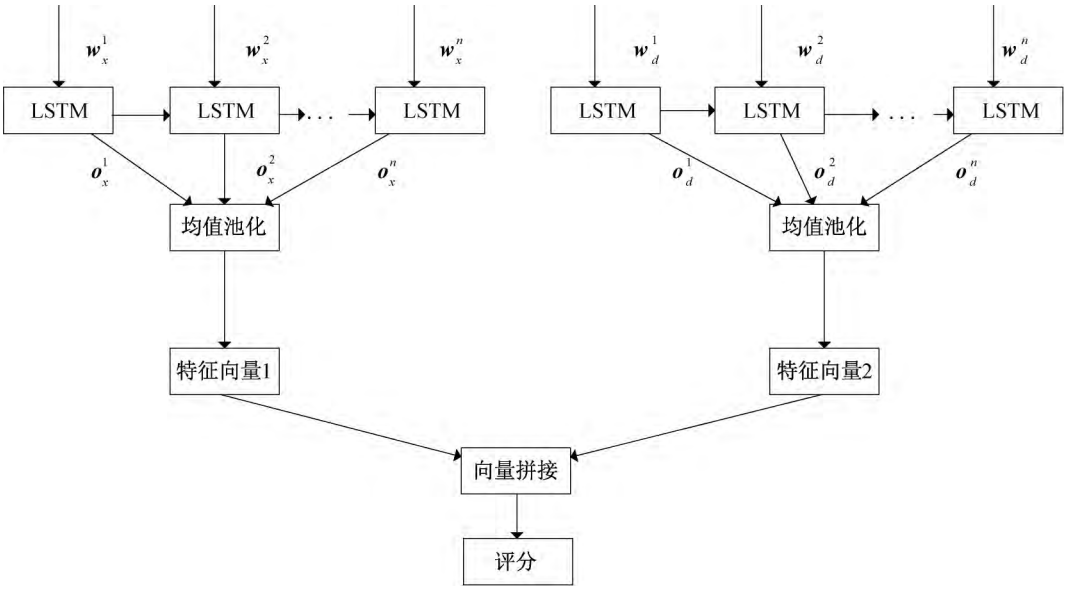

图 4　LSTM 句子抽取模型

o_s^j。原文档中第 j 个句子的向量表示通过 w_d^j 来表示，经过 LSTM 之后表示为 o_d^j。将 o_s^j 和 o_d^j 分别进行均值池化，然后进行拼接得到最终的向量表示。利用此模型能够实现文摘的自动获取，在实现过程中，模型的隐藏层和均值池化层均设置为 2 层。

1.4　评测标准

本文采用的是 ROUGE 评价方法，该评价方法的主要思想是对比系统生成的自动摘要与人工生成的标准摘要，通过统计两者摘要之间重叠的基本单元的数目来评价摘要的质量，其中基本单元指 n 元词（n-$gram$）、词序列和词对[17]。通过与多个专家人工提取的摘要进行对比来提高系统的鲁棒性。使用 $ROUGE$-N 表示 n 个共现词数的召回值，其计算公式可以表示为：

$$ROUGE\text{-}N=\frac{\sum_{S\in\{ReferenceSummaries\}}\sum_{n\text{-}gram\in S}Count_{\mathrm{match}}(n\text{-}gram)}{\sum_{S\in\{ReferenceSummaries\}}\sum_{n\text{-}gram\in S}Count(n\text{-}gram)},$$

其中：$Count_{\mathrm{match}}(n\text{-}gram)$ 表示同时出现在一篇候选摘要和参考摘要集合中的最大的 n-$gram$ 的个数，参考摘要集合中的 n-$gram$ 的个数用 $Count(n\text{-}gram)$ 进行表示。$ROUGE$-N 的值越高，说明文摘的质量越好。本文采用具有 2 个共现词和 3 个共现词的 $ROUGE$-2 和 $ROUGE$-3。

2　实验及结果分析

2.1　数据准备

实验选用新浪微博文本数据作为实验语料来源，经过预处理和去噪后，最终实验得到了包含微博文摘和对应原文的中文微博语料库，选定了 6852 条微博数据，分为训练数据 4796 条和测试数据 2056 条。数据处理的过程主要包含：去除特殊字符、去除表情符号、替换括号内的内容、替换日期标签、替换超链接标签、替换全角英文标签、替换数字等。在对文本进行预处理后，准备训练语料：将微博的原文作为输入，待预测的 Target 序列作为微博的文摘。准备好语料之后，对文本进行分词，并进行去停用词。

2.2　特征处理

分词后的结果中有很多词对于自动文摘的生成没有积极的作用，这些词的存在会增加特征向量的维度，因此考虑将这些词进行删减。本文的基本做法是将出现次数从高到低的前 3000 个词作为特征词，选取这些特征词的方法是 TF-IDF 模型。此方法能够通过下式表示句子中每个词的权重 $\omega_{i,j}$：

$$\omega_{i,j}=TF_{i,j}*IDF_i,$$

其中：$TF_{i,j}=\dfrac{f_{i,j}}{\sum_z f_{z,j}}$，表示句子 s_j 的词频，$f_{i,j}$ 表示关键词 k_i 在句子 s_j 中出现的次数，$\sum_z f_{z,j}$ 表示句子 s_j 中所有词的个数；$IDF_i=\log\dfrac{N}{c_i}$，表示 k_i 的逆文档频率，c_i 表示存在词 k_i 的句子的数量，N 表示文档总数。获得每个词的权重之后，取权值从高到低排列的前 3000 个词作为特征词，然后利用 word2vec 词转向量工具将每个句子中的特征词训练成一个固定维度的向量，其中 word2vec 的基本思想是通过训练将句子中的每个词映射为 K 维实数向量，词和词之间的语义相似度通过它们之间的距离进行表示。其采用一个三层的神经网络，输入层-

隐层-输出层。使用 word2vec 工具的优点在于丰富地利用了词的上下文和语义信息。本文把训练过程中词向量的维度固定成 200,滑动窗口大小设置为 5,同时将词频小于 5 的词语直接过滤掉。

2.3 实验结果及分析

在采用本文模型实现中文单文档文摘生成的过程中,进行了多组对比实验来分析提出模型的性能,对比的模型主要包括常用文摘模型和深度学习中的 LSTM 模型。本文提出的混合神经网络模型的第一步是获得词向量的表示,然后通过 CNN 获取句子向量表示,最后将词向量和句子向量分别通过 LSTM 模型,能够得到句子和文章的匹配程度,挑选出匹配程度较高的句子作为文档的文摘。其中在利用 CNN 获取句子向量表示的时候,采用 3 层卷积池化层,第一层卷积核的参数为(kernel size 5×5, padding 1,stride 2),第二层卷积核的参数为(kernel size 3×3,padding 0,stride 1),第三层卷积核的参数为(kernel size 3×3,padding 1,stride 1),模型的输入神经元个数为 200,对于句子中词数不足 200 的,通过添加占位符的形式进行补全,而 CNN 的输出也是长度固定的向量,它表示输入句子的向量表示形式。在获取句子的匹配程度时,LSTM 模型的输入向量维数为 200,LSTM 具有两层连接,两层的滑动窗口分别为 20 和 10。当模型训练结束后,将匹配概率高的若干句子作为文档的摘要。

本文选择了下面 5 种模型进行对比实验,分别为基于潜在语义(Latent semantic index, LSI)的自动文摘模型、基于 LDA 的自动文摘模型、基于 TextRank 的文摘生成模型、基于主成分分析(Principal component analysis, PCA)的文摘抽取模型以及基于 LSTM 的自动文摘模型。5 种对比模型的实验参数见表 1。

表 1 5 种对比模型的实验参数

模型参数	参数取值	参数实验范围
LSI 特征数/个	50	2～1000
LDA 主题数/个	8	2～10
TextRank 压缩率/%	20	10～30
PCA 降维后的特征数/个	50	20～80
LSTM 隐藏层大小	128	64～256

对于 LSI 模型,不同的特征数目将直接影响到文摘的生成效果;对于 LDA 模型,主题数的选择也将直接影响实验的最终效果。因此,对 LSI 和 LDA 模型在参数选择方面进行了一系列的对比实验,实验结果如图 5 和图 6 所示。从图 5 和图 6 中能够得出 LSI 和 LDA 模型的最终参数。

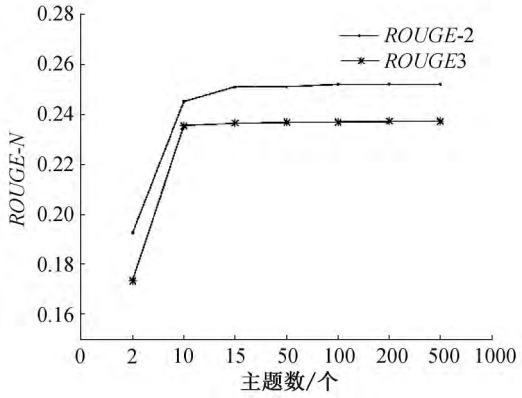

图 5 基于 LSI 模型的不同特征数量选择实验结果

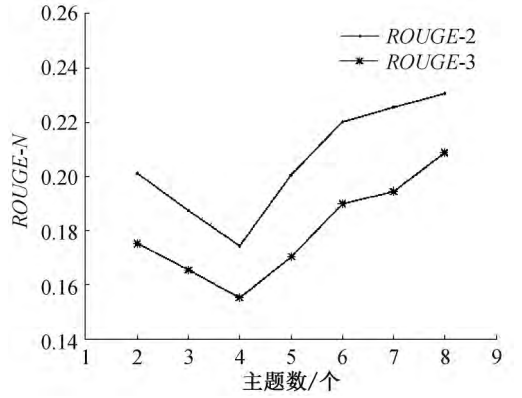

图 6 基于 LDA 模型的不同主题数量选择实验对比

从图 5 可以看出,当 LSI 的特征数设为 200 时,得到的实验结果的 *ROUGE*-2 和 *ROUGE*-3 达到最大,在特征数大于或小于 200 时,得到的实验结果不会有太大的提升,所以本文在测试过程中选择 200 作为特征数目的最终值。同理从图 6 中看出,文摘的质量受主题个数的影响较大。在本实验中,主题数目的多少直接和 LDA 的参数相关,实验结果显示当主题数为 8 时该模型能取得比较理想的实验结果,所以在测试的过程中,主题数最终设置为 8。

利用 LSTM 模型生成文摘时,实验采用的是单一序列的 LSTM 模型,这种模型将文档和句子合并为一个序列,并用一个 LSTM 模型对合并后的序列进行训练,实验表明这种模型能够提升自动文摘的效果,对于 *ROUGE*-2 和 *ROUGE*-3 的提升具有一定的帮助,最终获得的文摘符合上下文联系。图 7 显示了在 LSTM 中使用不同的隐藏层大小得到的实验结果。从实验结果可以看出当 LSTM 的隐藏层大小选择为 128 时,*ROUGE*-2 和 *ROUGE*-3 的值达到最大,在隐藏层大小大于或小于 128 时,模型的性能急剧下降,所以本文在最终的对比实验过程中选择隐藏层大小为 128 作为实验参数。

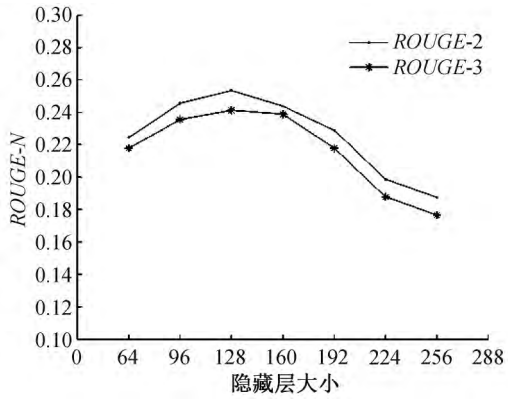

图 7　不同隐藏层大小对模型性能的影响

为了测试 TextRank 中不同压缩率对模型的影响，分别选择了 3 种不同的压缩率得到各自的 *ROUGE*-2 和 *ROUGE*-3 值。具体实验结果见表 2。从表 2 中可以看出选取压缩率为 20% 时，*ROUGE*-2 和 *ROUGE*-3 的值最大，同时从表中可以看出，不同的压缩率对 TextRank 模型的性能影响较大。因此，本文在最终的对比实验中选择压缩率为 20% 作为实验参数。

表 2　TextRank 中不同压缩率对模型性能的影响

压缩率/%	*ROUGE*-2	*ROUGE*-3
10	0.1215	0.1067
20	0.2461	0.2256
30	0.1834	0.1787

PCA 降维是将高维度的数据保留下最重要的一些特征，去除噪声和不重要的特征，本文中使用 PCA 降维的方法提取文档的文摘，选择不同降维后的特征维度的 PCA 模型将会产生不同的结果。本文中将降维后特征维数设置成不同的值可以得到不同的 *ROUGE*-2 和 *ROUGE*-3 值，具体实验结果如图 8 所示。从中可以看出，降维后的向量维度过小会导致 *ROUGE*-2 和 *ROUGE*-3 的值变小，这是因为降维后的向量包含原文的信息变少，和原文的匹配程度较低。从实验结果可以看出，当降维后的特征维度设置为 70 时得到的 *ROUGE*-2 和 *ROUGE*-3 值最大，大于或小于 70 时，*ROUGE*-2 和 *ROUGE*-3 的值偏低。因此，本文在最终的对比实验中选择降维后的特征维度为 70 作为实验参数。

在进行了上述 5 种模型的实验之后，下面给出本文模型的实验结果。本文模型的实验参数见表 3。

利用表 3 中的参数建立本文模型并进行实验得到的 *ROUGE*-2 和 *ROUGE*-3 值和上述 5 种模型的对比结果见表 4。其中 5 种对比模型中使用的参数是各自模型中最优的参数。

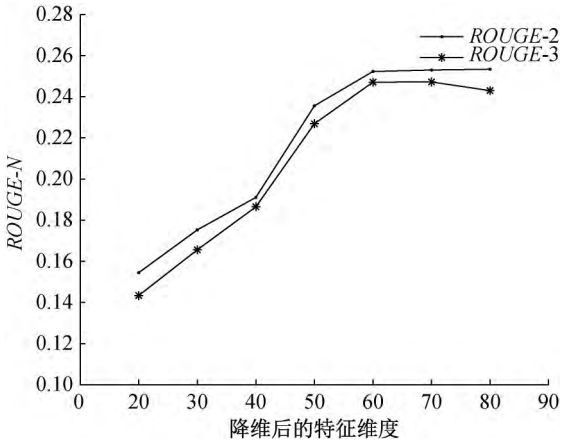

图 8　降维后的特征维度对模型性能的影响

表 3　本文模型参数

模型参数	参数取值
LSTM 隐藏层大小	128
CNN 隐藏层数	3
CNN 卷积核个数	100
CNN 卷积核大小	5×5,3×3,3×3
迭代次数	50
学习率	0.01

由表 4 可以看出，传统的自动文摘生成方法性能较为理想的是 LSI，其 *ROUGE*-2 和 *ROUGE*-3 分别为 0.2523 和 0.2372，采用 LSTM 模型生成自动文摘的 *ROUGE*-2 和 *ROUGE*-3 分别为 0.2534 和 0.2265，而本文算法的 *ROUGE*-2 和 *ROUGE*-3 为 0.2624 和 0.2445。通过比较可以看出，本文提出的模型相较于其他几种对比模型，*ROUGE*-2 值和 *ROUGE*-3 值都更大，说明提出的模型对于文摘的自动生成具有改进作用，同时由于本文提出的模型将 CNN 和 LSTM 进行了结合，从而有效地提高了文摘的准确性和连贯性，这样使得生成的文摘更加具有可读性，表示的中心内容更加明确。下面对这几种现存的模型和本文提出的模型生成的文摘进行比较，为了进一步说明本文提出的模型所生成的摘要更加具有可读性以及表示的中心内容更加明确，选取了其中一篇文档作为实验结果的展示，进行对比的原文档内容如图 9 所示。

对图 9 的文档，使用不同模型生成的文摘如图 10—图 15 所示。

表 4　不同算法的 ROUGE 值

模型	*ROUGE*-2	*ROUGE*-3
LSI	0.2523	0.2372
LDA	0.2306	0.2088
LSTM	0.2534	0.2413
TextRank	0.2461	0.2256
PCA	0.2335	0.2038
本文模型	0.2624	0.2445

原文档：

从汽车发明到现在，平均每天因事故死亡 3300 人左右，相当于 10 架大型客机坠毁。这些事故中，95%的事故都与人员失误有关，只有 4%的事故是纯机械故障。为此，21 世纪经济报道与广汽丰田汽车有限公司联合发起“文明驾驶”理念普及行动，邀请全国范围内的社会人士、专业机构及各高校参与，以“文明驾驶”为主题，针对各种不文明驾驶行为，通过公益视觉设计的方式进行普罗教育。视频入围作品展示如下：一．留住童真，作品名称：留住童真。作者：崔艺文/广东商学院。创意说明：以简单可爱的剪纸造型作为表现手法，故事简单易懂，通过天真无邪的小女孩和酒后驾车的醉鬼之间形成鲜明的对比，告诫人们不要酒后驾车。二．缘分，作品名称：缘份。作者：李在锦/中山大学。创意说明：《缘份》讲了一个很简单的故事，两个大学生通过在线网络认识，但是并没有见过面。终于最后相约见面。不料男生撞上同样去赴约的女生却不知道这个女生就是去赴约的女生。三．幸福在文明时，作品名称：幸福在文明时。作者：周泽娜、周思华、陈晓曼、陈晓燕/华南农业大学。创意说明：以不文明驾车，乱闯乱撞，最后导致悲惨事故发生作为开头，给观众强大的视觉冲击，通过通俗易懂的象征性游戏告诫人们：不要把生命当成儿戏，游戏可以重来，生命却已不再。文明驾驶，幸福生活。四．安全驾驶－－这是汽车系列之不要把你的汽车当成办公室，作品名称：安全驾驶－－这是汽车系列之不要把你的汽车当成办公室，作者：肖永鸿、罗豫、李国成/中山大学。创意说明：本短片”这是汽车”系列广告代表在汽车安全驾驶中”打电话”的行为。短片通过模拟”办公室”情景，将这种不安全驾驶行为在情景中的体现与现实的驾驶过程结合起来，达到巧妙的移情效果。

图 9 样例文档内容

从汽车发明到现在，平均每天因事故死亡 3300 人左右，相当于 10 架大型客机坠毁。21 世纪经济报道与广汽丰田汽车有限公司联合发起“文明驾驶”理念普及行动，邀请全国范围内的社会人士、专业机构及各高校参与，以“文明驾驶”为主题，针对各种不文明驾驶行为，通过公益视觉设计的方式进行普罗教育。

图 10 LSI 自动文摘结果

21 世纪经济报道与广汽丰田汽车有限公司联合发起“文明驾驶”理念普及行动，
一．留住童真，作品名称：留住童真。作者：崔艺文/广东商学院。创意说明：以简单可爱的剪纸造型作为表现手法，二．缘分，作品名称：缘份。作者：李在锦/中山大学。创意说明：《缘份》讲了一个很简单的故事，两个大学生通过在线网络认识，但是并没有见过面。终于最后相约见面。不料男生撞上同样去赴约的女生却不知道这个女生就是去赴约的女生。三．幸福在文明时，作品名称：幸福在文明时。作者：周泽娜、周思华、陈晓曼、陈晓燕/华南农业大学。创意说明：以不文明驾车，乱闯乱撞，最后导致悲惨事故发生作为开头，三．幸福在文明时，作品名称：幸福在文明时。作者：周泽娜、周思华、陈晓曼、陈晓燕/华南农业大学。

图 11 LDA 自动文摘结果

从汽车发明到现在，平均每天因事故死亡 3300 人左右，相当于 10 架大型客机坠毁。这些事故中，95%的事故都与人员失误有关，只有 4%的事故是纯机械故障。为此，21 世纪经济报道与广汽丰田汽车有限公司联合发起“文明驾驶”理念普及行动，邀请全国范围内的社会人士、专业机构及各高校参与，以“文明驾驶”为主题，针对各种不文明驾驶行为，通过公益视觉设计的方式进行普罗教育。

图 12 LSTM 自动文摘结果

一．留住童真，作品名称：留住童真。作者：崔艺文/广东商学院。创意说明：以简单可爱的剪纸造型作为表现手法，故事简单易懂，通过天真无邪的小女孩和酒后驾车的醉鬼之间形成鲜明的对比，告诫人们不要酒后驾车。二．缘分，作品名称：缘份。作者：李在锦/中山大学。创意说明：《缘份》讲了一个很简单的故事，两个大学生通过在线网络认识，但是并没有见过面。终于最后相约见面。不料男生撞上同样去赴约的女生却不知道这个女生就是去赴约的女生。 三．幸福在文明时，作品名称：幸福在文明时。作者：周泽娜、周思华、陈晓曼、陈晓燕/华南农业大学。创意说明：以不文明驾车，乱闯乱撞，最后导致悲惨事故发生作为开头，给观众强大的视觉冲击，通过通俗易懂的象征性游戏告诫人们：不要把生命当成儿戏，游戏可以重来，生命却已不再。文明驾驶，幸福生活。四．安全驾驶－－这是汽车系列之不要把你的汽车当成办公室，作品名称：安全驾驶－－这是汽车系列之不要把你的汽车当成办公室，作者：肖永鸿、罗豫、李国成/中山大学。创意说明：本短片＂这是汽车＂系列广告代表在汽车安全驾驶中＂打电话＂的行为。短片通过模拟＂办公室＂情景，将这种不安全驾驶行为在情景中的体现与现实的驾驶过程结合起来，达到巧妙的移情效果。

图 13 TextRank 自动文摘结果

二．缘分，作品名称：缘份。作者：李在锦/中山大学。创意说明：《缘份》讲了一个很简单的故事，两个大学生通过在线网络认识，但是并没有见过面。终于最后相约见面。不料男生撞上同样去赴约的女生却不知道这个女生就是去赴约的女生。

图 14　PCA 自动文摘结果

平均每天因事故死亡 3300 人左右，相当于 10 架大型客机坠毁。21 世纪经济报道与广汽丰田汽车有限公司联合发起“文明驾驶”理念普及行动，通过公益视觉设计的方式进行普罗教育。视频入围作品展示如下：留住童真，缘分，幸福在文明时，安全驾驶。

图 15　本文模型自动文摘结果

图 10—图 15 分别给出了对于一篇文档利用几种不同模型得到的自动文摘结果。在对多篇单一文档进行实验后的结果可以看出，和其他 5 种模型相比较，本文提出的模型生成的摘要的语义更加连贯，上下文关系明确，同时非常好地表示了整个文档的主要内容。其余几种模型得到的结果有的不能很好的表达文档的中心含义，有的结果虽然语句上下文连贯，但是个别句子的上下文关系不太明确。通过 *ROUGE*-2、*ROUGE*-3 值的对比和不同自动文摘模型产生摘要的比较来看，本文提出的基于混合神经网络的单文档自动文摘模型具有较好的性能。

3　结　论

本文利用神经网络将前人提出的两种方法进行融合，提出一种混合神经网络 CNN-LSTM 模型，并将该模型应用到自动文摘中，从而形成一种基于混合神经网络的单文档自动文摘模型。该模型首先通过 jieba 分词工具对文本句子进行分词并删除部分停用词，然后用词转向量工具 word2vec 将句子中的词转换成向量表示形式；接着将整个句子的词向量输入 CNN 模型从而获得句子的向量表示形式；最后将获得的词向量和句子向量分别通过 LSTM 模型，并将两个 LSTM 的输出结果进行拼接，通过得到的结果判断句子和文章的匹配程度，挑选出匹配程度较高的句子作为文档的文摘。为了验证本文提出模型的性能，分别从两个角度对实验进行验证。首先对比了本文模型和其余 5 种模型的 *ROUGE*-2 和 *ROUGE*-3 值，从这两个值的对比可以看出，本文模型得到的文摘更加准确；接着对比了本文模型和其余 5 种模型生成文摘的结果，随机选取某一篇文档以及 6 种模型生成的摘要，从结果可以看出，本文模型得到的文摘语句更加连贯，上下文关系更好，生成的文摘能很好地表达原文档的主要含义。通过这两组对比能够看出本文提出的基于混合神经网络的模型在自动文摘获取上效果良好，能提升文摘的质量。

参考文献：

[1] Hadyan F，Shaufiah，Bijaksana M A. Comparison of document index graph using TextRank and HITS weighting method in automatic text summarization[J]. Journal of Physics：Conference Series，2017，801(1)：012076.

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

[3] Hu B，Chen Q，Zhu F. LCSTS：A large scale Chinese short text summarization dataset[EB/OL]. (2016-02-19)[2018-07-10]. https://arxiv.org/abs/1506.05865v4.

[4] Nichols J，Mahmud J，Drews C. Summarizing sporting events using twitter [C]// ACM International Conference on Intelligent User Interfaces. ACM，2012：189-198.

[5] Blei D M，Ng A Y，Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research，2003，3：993-1022.

[6] Rush A M，Chopra S，Weston J. A neural attention model for abstractive sentence summarization[EB/OL]. (2015-09-03) [2018-07-10]. https://arxiv.org/abs/1509.00685.

[7] Wong K F，Wu M，Li W. Extractive summarization using supervised and semi－supervised learning[C]// 22nd International Conference on Computational Linguistics. Manchester：Association for Computational Linguistics Stroudsburg，2008：985-992.

[8] Majak M，Zolnierek A，Wegrzyn K，et al. Tweet classification framework for detecting events related to health problems[C]// 10th International Conference on Computer Recognition Systems. Polanica Zdroj：Springer，2017：453-461.

[9] Nahian M A，Iftekhar A S M，Islam M T，et al. CNN－based prediction of frame-level shot importance for video summarization[EB/OL]. (2017-8-23)[2018-07-

10].https://arxiv.org/abs/1708.07023.

[10] 应文豪,肖欣延,李素建,等.一种利用语义相似度改进问答摘要的方法[J].北京大学学报,2017,53(2):197-203.

[11] Le P, Zuidema W. Quantifying the vanishing gradient and long distance dependency problem in recursive neural networks and recursive LSTMs [EB/OL]. (2016-03-01) [2018-07-10]. https://arxiv.org/abs/1603.00423.

[12] Misawa S, Taniguchi M, Miura Y, et al. Character-based bidirectional LSTM-CRF with words and characters for Japanese named entity recognition[C]// First Workshop on Subword and Character LevelModels in NLP. Copenhagen: Proceedings of the Workshop,2017:97-102.

[13] Wu M, Liu L, Yao W, et al. Semantic relation classification by bi-directional LSTM architecture [C]//26th International Conference on Computational Linguistics. Osaka: The COLING 2016 Organizing Committee, 2017: 1254-1263.

[14] Sugawara H, Takamura H, Sasano R, et al. Context representation with word embeddings for WSD[C]// Computational Linguistics. Singapore: Springer, 2015: 108-119.

[15] Chen Z, He Z, Liu X, et al. Evaluating semantic relations in neural word embeddings with biomedical and general domain knowledge bases[J]. Bmc Medical Informatics & Decision Making, 2018, 18(2): 53-68.

[16] Bharti S K, Babu K S, Pradhan A. Automatic keyword extraction for text summarization in multidocument e-newspapers articles[J]. European Journal of Advances in Engineering and Technology, 2017, 4 (6):410-427.

[17] Lin C Y. ROUGE: A package for automatic evaluation of summaries[C]// Text Summarization Branches Out. Barce: Association for Computational Linguistics, 2004: 74-81.

(责任编辑:康 锋)