计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191000007

基于主次关系特征的自动文摘方法

张 迎 张宜飞 王中卿 王红玲
苏州大学计算机科学与技术学院 江苏 苏州 215006
(20175227066@stu.suda.edu.cn)

摘 要 自动文摘研究是指通过自然语言处理技术对原始文本进行压缩、提炼,在保留文档核心思想的同时为用户提供简明扼要的文字描述。传统的自动文摘方法通常只考虑字、词、句子等浅层的文本语义信息,而忽略了深层的主次关系等篇章结构信息对抽取文档核心句子的指导作用。对此,提出一种基于主次关系特征的自动文摘方法。该方法基于长短期记忆网络(Long Short-Term Memory,LSTM)神经网络构建了基于主次关系特征的单文档抽取式摘要模型,通过双向LSTM神经网络模型对句子信息和主次关系信息进行信息增强和语义编码,并利用单向LSTM神经网络对编码后的信息进行摘要抽取。实验结果表明,与当前主流的单文档抽取式摘要方法相比,该方法在摘要的准确性、稳定性和ROUGE评价指标上均有显著的提高。

关键词:自然语言处理;抽取式摘要;主次关系;神经网络;LSTM
中图法分类号 TP391

Automatic Summarization Method Based on Primary and Secondary Relation Feature

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

Abstract Automatic summarization technology refers to providing users with a concise text description by compressing and refining the original text while retaining the core idea of document. Usually,the traditional method only considers the shallow textual semantic information and neglects the guiding role of the structure information such as the primary and secondary relations in core sentences extraction. Therefore,this paper proposes an automatic summarization method based on the primary and secondary relation feature. This method utilizes the neural network to construct a single document extractive summarization model based on primary and secondary relation feature. The Bi-LSTM neural network model is used to encode the sentence information and primary and secondary relation information,and the LSTM neural network is utilized to summarize the encoded information. Experimental results show that the proposed method has a significant improvement in accuracy,stability and the ROUGE evaluation index compared with the current mainstream single document extractive summarization methods.

Keywords Natural language processing,Extractive summarization,Primary and secondary relation,Neural network,LSTM

1 引言

随着信息技术的飞速发展,人们可以在互联网上获取越来越多的信息,但同时也面临着庞大的互联网信息带来的信息过载问题。自动文摘技术是指通过计算机自动地将冗长的文本压缩到规定长度内,同时保持原始文本的主要信息不丢失,这种技术有效地减轻了信息过载的负担。自Luhn[1]于1958年发表第一篇自动文摘成果以来,自动文摘研究(Automatic Text Summarization)便成为了自然语言处理中最为活跃的分支。传统的自动文摘研究通常只考虑词汇信息或文本的语义信息,缺乏对篇章结构关系等深层信息的有效利用。

近年来,自然语言处理领域的研究内容逐步从浅层次的词汇、句法分析延伸到深层次的语义理解,具体而言,就是从单个的字、短语、句子,延伸至句群、段落、篇章[2]。篇章结构分析研究应运而生,成为自然语言处理领域最活跃的研究方向之一。篇章主次关系作为篇章结构分析的一个重要组成部分,一直融合在篇章结构的研究中。

篇章主次关系分为微观和宏观两个层面:微观主次关系是指一个句子内部的主次关系或两个连续句子间的主次关系;而宏观主次关系则是更高层次的主次关系,表现为段落、章节间的主次关系[2]。由于宏观主次关系在理论、语料库建设等方面尚不完善,本文的研究主要基于微观层面的主次关系。例如对于例1,从语义角度分析,分隔符"|"前后两句均包含核心词"高考学生""死亡"这两个关键词,因此在摘要任务中很容易将两句话都抽取出来作为摘要句。但是从篇章结构角度分析,这两句的结构关系为总分型,即第一句为总述型语句,第二句为第一句的详细描述,从而清晰地得出在均包含核心关键词的前提下,第一句应该为核心句,第二句为非核心

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:国家自然科学基金项目(61806137,61976146);江苏省高等学校自然科学研究面上项目(18KJB520043)
This work was supported by the National Natural Science Foundation of China (61806137,61976146) and Natural Science Foundation of the Jiangsu Higher Education Institutions of China (18KJB520043).
通信作者:王红玲(hlwang@suda.edu.cn)

句，因此在摘要抽取的过程中应着重考虑第一句。

例 1　高考学生身体不适到诊所就医死亡。|今天下午，信阳市淮滨一高考学生吴某某，身体不适到诊所就医，就医期间死亡。

从上文分析中可以看出，研究篇章的主次关系，有助于更好地分析文本内部的语义关联和展开思路，从而帮助摘要任务抽取更有代表性的句子。因此，本文将篇章主次关系特征加入自动文摘任务中，在考虑浅层文本语义信息的基础上同时考虑篇章的结构信息，以此来提高文本摘要的质量。首先，采用自动化篇章主次关系标记模型给已有语料标记主次关系，并将标记好的结构化信息处理成自动文摘模型可使用的序列化信息。然后，按比率从正文中抽取与参考摘要相似度最高的句子作为正例，其余作为负例。最后，基于 LSTM 神经网络构建基于主次关系特征的自动文摘系统。该方法将主次关系信息作为特征加入自动文摘任务中，有效地提高了摘要抽取的准确性和概括性。实验结果表明，本文提出的方法有效地提高了单文档抽取式摘要的文摘质量。

2　相关工作

2.1　文本摘要

按照实现文摘的技术，自动文摘可分为抽取式文摘（Extractive Summarization）和生成式文摘（Abstractive Summarization）。

抽取式文摘利用不同方法对文档结构单元（句子、段落等）进行评价，对每个结构单元赋予一定权重，然后选择最重要的结构单元组成摘要。传统的抽取式方法主要有 3 种。

（1）基于图排序的算法，如 Mihalcea 和 Tarau 提出的 TextRank 算法[3]，Erkan 和 Radev 提出的 LexPageRank 算法[4]等。该方法主要是在词频的基础上利用各种关联度计算方法计算词项间的关联度，构建相应的无权或加权的网络图，从而抽出文档最核心的句子作为摘要。

（2）基于统计机器学习的方法。Kupiec 等[5]通过朴素贝叶斯分类模型的方法判定文档中的某一个句子是否应该被选为摘要。Conroy 等[6]提出了基于隐马尔可夫模型的摘要算法，该算法使用文档中的一些特征（如位置、长度等）来计算句子得分，然后依据句子得分生成文本摘要。

（3）基于神经网络的方法。Kim 等[7]提出的 Text CNN 算法，利用卷积神经网络对文本进行分类，从而实现摘要抽取。Liu 等[8]提出的基于 RNN 神经网络的文本分类多任务学习模型，弥补了 CNN 不适合序列信息较长的缺陷。Zhou 等[9]结合 CNN 和 LSTM 提出了一种 C-LSTM 网络模型，该模型利用 CNN 抓取短语的局部特征并利用 LSTM 获得句子表示，在情感分析和文本摘要领域均取得了突出的性能。

生成式文摘研究由于实现的复杂程度较高，目前发展还不是很成熟，通常需要利用自然语言理解技术对文本进行语法、语义分析，对信息进行融合，最后利用自然语言生成技术生成新的摘要句子。目前，生成式文摘的实现主要采用基于编码器-解码器架构的序列到序列的学习模型[10-12]。其中，编码器、解码器均由数层循环神经网络构成，编码器负责把原文编码为一个向量 $\boldsymbol{C}$；解码器负责从这个向量 $\boldsymbol{C}$ 中提取信息，获取语义并生成文本摘要。这种方法能较好地解决抽取式文摘的连贯性较差和冗余性较高等问题，但由于这些方法需要远大于抽取式方法的训练语料，加上当前主流的神经网络框架尚不能够有效地对长文本进行语义编码，因此目前的生成式文摘研究大多集中于短文本生成。

2.2　主次关系

篇章主次关系主要分为微观和宏观两个层面。目前微观层面的篇章主次关系研究较为广泛，宏观篇章主次关系在理论、语料库建设和模型设计上尚不完善。在微观主次关系层面，相关语料主要有修辞结构篇章树库（RST-DT）[13]和汉语篇章树库（CDTB）[14]等。

RST-DT 以修辞结构理论（RST）为支撑，标注了篇章单元、篇章关系、主次关系和篇章结构，从而生成有层次的篇章结构树。在 RST-DT 语料上，Hernault 等[15]使用开源的 HILDA 分析器，在主次关系识别任务中取得 61.3% 的 *F*1 值。Feng 和 Hirst[16]在其基础上，使用线性链的条件随机场模型识别微观主次关系，正确率达到了 71%。

CDTB 基于连接依存树的篇章结构理论，在宾州大学汉语树库（CTB）上标注了 500 篇微观篇章关系结构。以 CDTB 语料为基础，Chu 等[17]设计了基于特征的主次关系识别系统，其正确率达到了 53.21%。Li 等[18]构建了一个自底向上的汉语微观篇章结构分析平台，正确率达到了 69%。Sun 等[19]构建的基于转移的中文篇章结构解析模型，结构解析性能达到了 78%以上。

3　基于主次关系特征的自动文摘方法

传统的单文档抽取式摘要研究通常将摘要问题形式化为文本二分类问题，通过构建合适的分类模型将原文的句子进行分类，从而实现文摘抽取。这种方法通常只考虑文本的语义信息，且对特征的依赖性较高。

本文提出了一种基于篇章主次关系特征的文本摘要方法。一方面，通过构建基于 LSTM 神经网络的文本分类模型，解决了传统分类器对特征的高依赖性问题。另一方面，将篇章主次关系信息序列化后，以特征的形式融入到摘要任务中，使得模型在考虑浅层的文本语义信息的基础上同时考虑深层次的篇章结构信息。

接下来主要从主次关系特征的序列化表示和基于主次关系特征的自动文摘模型两个方面来介绍本文的方法。

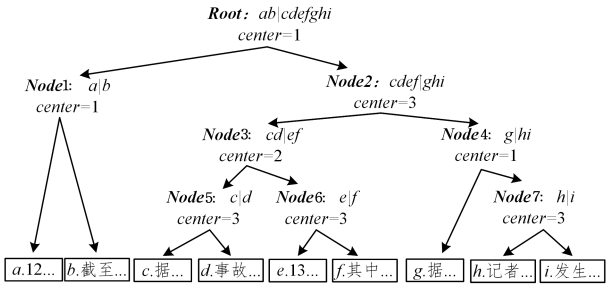
3.1　主次关系特征的序列化表示

Li 等[20]以连接依存树的篇章结构理论为基础，构建了汉语篇章结构语料库 CDTB（Chinese Discourse Tree-Bank）。CDTB 包含 500 篇文档，每个文档采用自顶向下的标注策略。对于每一段内容，先找出其最上层的关系，然后递归地对切分后的主要内容进行标注，最终形成以段落为单位的篇章树形结构。在主次关系判定方面，CDTB 按照全局重要性区分主次关系，并将主次关系分为 3 类：*center*＝1：核心在前；*center*＝2：核心在后；*center*＝3：同为核心。

在汉语篇章语料库 CDTB 的基础上，Sun 等[19]利用转移系统和深度学习的方法构建了一个完整的从平文本到树形结构的篇章结构自动解析框架。本文利用该框架，将原语料解析成带有主次关系标注的篇章结构语料。

图 1 给出了本文语料通过基于转移的中文篇章结构解析模型解析后得到的篇章结构树，其中 a,b,c,d,e,f,g,h,i 表示子句编号。*Node* 节点为篇章关系节点，将下一级的篇章子

结构组合为上一级的篇章结构，最终形成一棵完整的篇章结构树。篇章结构树可以简明地展示篇章的主次关系，但是由于文本摘要模型通常只能处理序列化的信息，因此需要将结构化的篇章主次关系处理成文摘模型可读取的序列化信息。基于转移的中文篇章结构解析模型标记了篇章全局的主次关系，而自动文摘模型的处理单元是子句。因此，需要将全局的主次关系信息落实到每个字句上，得到每个子句的核心程度。



a. 12 月 13 日凌晨 1 时 26 分，新疆呼图壁县雀尔沟镇白杨沟煤矿发生瓦斯爆炸事故；b. 截至今日 15 时 30 分，被困井下人员中 1 人获救，其余 21 人确定遇难；c. 据介绍，事故发生时，下矿人员 34 人，12 人成功升井，1 人 13 日上午获救，被困井下 21 人确认遇难；d. 事故发生后，国家煤矿安全监察局、新疆当地各级政府有关部门人员迅速赶赴现场，组织大批专业救护人员全力开展搜救工作；e. 13 日，中午 13 时记者赶往事发煤矿，看到事发煤矿现场秩序平稳，大量公安、消防、医护等人员在现场集结待命；f. 其中矿山救护队正对失踪矿工进行全力搜救；g. 据透露，数百名救援人员于今晨 03 时连夜赶赴现场搜救；h. 记者获悉，获救人员已在医院救治，有关部门正在积极开展善后处置和事故调查工作；i. 发生事故的煤矿距新疆首府乌鲁木齐市约 160 km，开采年限从 2007 年 8 月至 2017 年 8 月，年生产规模为 9 万吨。

图 1 基于转移的中文篇章结构树示例

Fig. 1 Example of transfer-based Chinese chapter tree

首先，将全局的主次关系转化为相邻句子间的主次关系，找到该句子在树中的位置，再从该节点到根节点遍历该树，找到两个相邻节点的最小公共父节点，最小公共父节点的主次关系标记就作为这两个相邻句子的主次关系标记。例如，句子 a 和 b 的主次关系标记为它们的最小公共父节点 $Node1$ 的主次关系标记，即 $center(a,b)=1$ 。

然后，将句子间的主次关系标记转化为每个子句的主次关系特征。其中，核心部分标记为 1，非核心部分标记为 0。任意两个相邻句子都有主次关系，这意味着有的句子可能会存在两种不同的主次关系标记。例如：句子 h 与前一句 g 的主次关系标记为 $center_{(g,h)}=1$，此时句子 h 为非核心句，$feature_{h_{前}}=0$；与后一句 i 的主次关系标记为 $center_{(h,i)}=3$，此时 h 与 i 均为核心句，$feature_{h_{后}}=1$ 。为了解决这个问题，将每个句子在不同句子对中的核心关系相加，得到其最终的主次关系特征标签。

$$feature_h=feature_{h_{前}}+feature_{h_{后}}=0+1=1 \tag{1}$$

综上所述，本文实验所用的主次关系特征有 3 类：

$$feature_i\in\{0,1,2\}$$

其中，主次关系特征值越大，表明该句子的核心程度越高。这种方式得到的主次关系特征，既保证了主次关系特征的全局性，又量化了句子的核心程度。

3.2 基于主次关系特征的自动文摘方法

图 2 所示为本文提出的基于主次关系特征的自动文摘模型，该模型主要分为以下 3 个部分。

(1)文本和特征的嵌入表示：通过高频词表得到基于词的向量矩阵，并将句子的文本信息和主次关系特征映射为固定维度的向量表示。

(2)特征捕捉和语义编码：采用 Bi-LSTM 神经网络对句子信息和主次关系特征进行编码，以便更加有效地捕捉句子的主次关系特征。

(3)文本分类与摘要抽取：利用基于 LSTM 神经网络的文本分类器对融合主次关系特征的句子进行文本分类，判断其是否属于摘要。

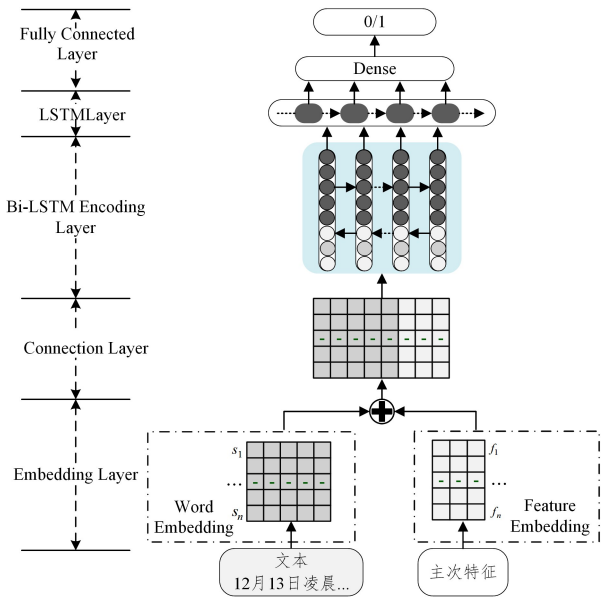


图 2 基于主次关系的单文档抽取式摘要模型图

Fig. 2 Single document extractive summarization model diagram based on primary and secondary relation

3.2.1 文本和特征的嵌入表示

首先，将标注好主次关系特征的原文本进行去噪、分句、分词等预处理，然后通过手工构建的高频词表 V 得到相应句子的向量表示：

$$S=[w_1,w_2,\cdots,w_t,\cdots,w_T] \tag{1}$$

$$f=f_s \tag{2}$$

其中，w_t 表示句子序列中第 t 个词在词表 V 中的 one-hot 表示，T 表示句子长度，$|V|$ 表示词表大小。f_s 表示句子 S 的主次关系特征。

得到句子向量和主次关系特征向量之后，融合主次关系信息的句子向量可以表示为：

$$X=[S,f] \tag{3}$$

其中，$[,]$ 表示向量的拼接操作，X 为融合主次关系特征的句子向量。

3.2.2 特征增强和语义编码

将主次关系特征向量直接拼接在句子向量的末尾，可以很简单地将主次关系信息融入文摘任务中。但是随着拼接序列长度的增加，单一的句子级别模型无法高效地学习到超长序列中的远程依赖，容易出现信息丢失的情况。

因此，本文添加一层 Bi-LSTM 对包含主次关系特征的句子向量进行二次编码，以实现特征增强。与单向的 LSTM 相比，Bi-LSTM 输入层的数据会经过向前和向后两个方向推算，最后输出的隐含状态进行 CONECT 后再作为下一层的输入。这种方式能够更好地捕捉句子中上下文的信息。

图 3 为 Bi-LSTM 神经网络在时间上的展开结构图，这个

结构为输出层提供输入序列中每一个点在过去和未来的完整上下文信息。6 个独特的权值在每一个时间步被重复利用，它们分别对应：输入到向前和向后隐含层（w_1，w_3），隐含层到隐含层自己（w_2，w_5），向前和向后隐含层到输出层（w_3，w_6）。

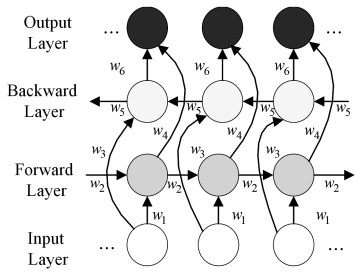


图 3　Bi-LSTM 在时间上的展开图

Fig. 3　Bi-LSTM expansion graph in terms of time

在前向传播时，从 1 时刻到 t 时刻正向计算一遍，得到并保存每个时刻向前隐藏层的输出；在后向传播的过程中，从 t 时刻到 1 时刻反向计算一遍，得到并保存隐藏层的输出。最后，将前向传播和后向传播输出的结果加和计算得到最终的输出。具体公式如下：

$$h_t = f(w_1 x_t + w_2 h_{t-1}) \tag{4}$$

$$h_t' = f(w_3 x_t + w_5 h_{t+1}') \tag{5}$$

$$o_t = g(w_4 h_t + w_6 h_t') \tag{6}$$

其中，h_t 表示的是正向传播隐藏层的状态值，与 h_{t-1} 有关；h_t' 表示的是反向传播隐藏层的状态值，与 h_{t-1}' 有关。正向计算和反向计算不共享权值，即 w_1，w_2，w_3，w_5 表示的都是不同的权重矩阵。o_t 表示的是神经网络最终的输出值，取决于正向计算和反向计算的加和。

3.2.3　文本分类和摘要抽取

经过 Bi-LSTM 编码后，主次关系特征已经与句子向量充分融合。接下来，模型要做的就是在主次关系特征信息的指导下进行文本分类，判断每个句子是否为摘要句。首先，将编码后的句子向量输入之前已经构建好的基于 LSTM 神经网络的文本分类器模型中。

$$h_t = \overrightarrow{LSTM}(X) \tag{7}$$

然后，通过 Sigmod 函数得到每个句子的 0/1 概率分布。

$$S(t) = \frac{1}{1+e^t} \tag{8}$$

最后，把每篇文章的所有句子的 0/1 分布概率 $P=[P_0, P_1]$ 按 P_1 从大到小排序，选取前 k 个概率最高的句子作为候选摘要句，从而实现文本摘要的抽取。

本文的损失函数采用的是对数损失函数（Binary_crossentropy）。该函数一般用于二分类，主要是针对概率之间的损失函数，概率相差越大，$loss$ 越大。

$$loss = -\sum_{i=1}^{n} \hat{y}_i \log y_i + (1-\hat{y}_i)\log(1-\hat{y}_i) \tag{9}$$

$$\frac{\partial loss}{\partial y} = -\sum_{i=1}^{n} \frac{\hat{y}_i}{y_i} - \frac{1-\hat{y}_i}{1-y_i} \tag{10}$$

4　实验与评价

4.1　实验数据

本文数据集 TTNews[1] 来源于 NLPCC2017 shared task3：Single Document Summarization 评测任务。该数据集是目前最大的中文单文档摘要语料库，包含一个训练集和一个测试集。训练集由 50000 篇来自头条的新闻文章和相应的人工摘要组成。测试集包含 2000 篇新闻文本，涉及体育、食品、娱乐、科技等诸多方面。数据集的详细信息如表 1 所列。

表 1　数据集平均长度统计表

Table 1　Statistical table of dataset average length

	篇数/篇	正文平均长度/字	摘要平均长度/字
训练集	50000	1036	45
测试集	2000	1037	45

4.2　评价方法

ROUGE[21] 是 Lin 在 2004 年提出的一种自动文摘评价方法，被广泛应用于 NIST 组织的自动摘要评测任务中。ROUGE 算法是一种面向 n 元词召回率的评价方法，其基本思想是由多个专家分别生成的人工摘要构成标准摘要集，将系统生成的自动摘要与人工摘要相比对，通过统计二者重叠的基本单元（n 元语法、词序列和词对）的数目来评价摘要的质量。目前，该方法已经成为自动摘要评价技术的通用标准。

本文采用 ROUGE 中的 ROUGE-1，ROUGE-2，ROUGE-L 对生成的摘要进行评价，即从字、词、长文本 3 个角度来评价本文模型的效果。

4.3　对照实验

为了评估本文提出的基于主次关系特征的自动文摘模型在摘要任务中的表现，一方面，我们以 NLPCC 2017 shared task3：Single Document Summariztion 评测任务给出的基线模型和该任务最终提交结果中的最优结果作参考；另一方面，采用当前主流的以下自动文摘模型与本文模型进行效果对比。

LEAD：自动文摘评测任务给出的基线模型，直接取文档中的前 60 个字符作为摘要。由于新闻文本通常在开篇点明题意，开篇的句子往往包含大量的核心内容，因此该基线系统已经具有较高的性能。

NLP_ONE[22]：自动文摘评测任务中第一名的实验结果，即在该数据集上目前官方公布的最优结果。

PageRank[23]：基于图排序的自动文摘算法。把文章分成若干个段落或句子的集合，每个集合对应一个图的顶点，集合之间的关系对应边，最后计算各个顶点的得分并抽取得分高的句子作为摘要。

Text-CNN[24]：Kim 在 2014 年提出的利用卷积神经网络对文本进行分类的算法。利用多个不同大小的卷积核来提取句子中的关键信息（类似于多窗口大小的 n-$gram$），能够更好地捕捉局部相关性。

LSTM：基于 LSTM 神经网络的单文档抽取式摘要模型。对文本进行分句、分词等预处理之后，通过词编码得到相应的句子编码，然后输入 LSTM 神经网络层进行训练，最后将训练好的神经元输入一个全连接层得到文本的 0/1 概率分布，对其按从大到小排序，并选取前 k 个句子作为候选摘要句。

4.4　实验设置

4.4.1　文本预处理

预处理时，首先用正则表达式将原始数据集中的 $html$ 标签、符号以及非法字符等无效字符去除。此外，受主次关系标

1) http://tcci.ccf.org.cn/conference/2018/taskdata.php

记模型限制，对于文本长度超过 3 000 个字符的文本只保留前 3000 个字符。然后，将标记好的结构化篇章主次信息序列化，为文本中的每一个句子标记主次关系特征，主次关系特征值越大，核心程度越高。最后，依次计算正文中句子与参考摘要句子的余弦相似度比值，并选取相似度最高的前 15%标记为正例，其余的标记为负例。

4.4.2 超参数设置

采用 KERAS 深度学习框架，词表取的是前 50 000 个高频词。模型训练的参数设置如表 2 所列。

表 2 实验参数设置

Table 2 Experiment parameter setting

实验参数	值
EMBED_SIZE	128
HIDDEN_SIZE	100
MAX_LEN	50
DROUPOUT	0.5
BATCH_SIZE	300
EPOCHS	10

4.4.3 实验结果及分析

(1)与基准系统的比较

本文 Baseline 以每篇文本的前 60 个字符作为摘要。为保证评测结果的公正性和有效性，最终用于评测的摘要也为候选摘要句的前 60 个字符。如表 3 所列，LEAD 为摘要评测任务规定的基线系统，PageRank，LSTM，Text-CNN 等模型的实验结果为对比实验结果，NFS 为本文提出的基于主次关系特征的自动文摘方法。

表 3 在 TTNews 数据集上的实验结果

Table 3 Experimental results on the TTNews dataset

(单位：%)

Model	*RG*-1	*RG*-2	*RG*-*L*
NLP_ONE	—	22.8	—
LEAD	31.4	17.1	23.6
PageRank	34.7	19.8	26.9
LSTM	36.9	21.6	28.6
Text-CNN	37.0	21.9	28.7
NFS	**38.4**	**23.2**	**30.2**

由表 3 的实验结果可以看出：首先，由于本文实验语料采用的是新闻语料，新闻语料往往开篇点题，因此基准系统的性能较优。其次，PageRank，LSTM，Text-CNN 的实验效果均好于 LEAD，这说明对照模型的选取有良好的参考价值。最后，与文提出的基于主次关系特征的自动文摘模型 NFS 的效果最好，和基准模型相比在 *RG*-1，*RG*-2 和 *RG*-*L* 上分别提高了 7%，6.1%和 6.6%，并且超过了 NLPCC 2017 评测任务官方公布的最好结果。

(2)主次关系特征对结果的影响

下面分别从主次关系特征的影响和模型性能的整体优化两个角度来对本文的实验结果做进一步分析。

如表 4 所列，LSTM 和 Text-CNN 为本文的对照实验，LSTM＋feature 和 CNN＋feature 分别为在 LSTM 和 Text-CNN 模型中融入主次关系特征。由表 4 中数据可以看出，在 LSTM 模型和 Text-CNN 模型中加入主次关系特征信息后，模型性能分别提高了 0.7%和 0.4%，这说明将主次关系特征信息融入文本摘要任务中对于摘要模型抽取出更有代表性的句子具有一定的指导作用。提升效果不是很好主要有两方面原因。

表 4 加入主次关系特征的实验结果

Table 4 Results after adding the primary and secondary relation feature

(单位：%)

Models	*RG*-1	*RG*-2	*RG*-*L*
NLP_ONE	—	22.8	—
LSTM	36.9	21.6	28.6
Text-CNN	37.0	21.9	28.7
LSTM＋feature	37.4	22.3	29.2
CNN＋feature	37.5	22.3	29.3
NFS	37.9	22.8	29.6

1)本文采用的主次关系标记是模型自动标记的，这种方法与传统的手工标记方法相比，节省了大量的人力、物力，更利于未来的实际应用，但是也存在标记精度不够高等问题。

2)LSTM＋feature 和 CNN＋feature 这两个模型是直接将主次关系特征编码后拼接在句子向量的末尾，这种方法没有充分体现主次关系特征的指导作用。

本文模型 NFS 的性能优于 LSTM＋feature 和 CNN＋feature 的结果，说明本文提出的基于主次关系特征的自动文摘方法可以有效利用主次关系信息对摘要任务的指导作用，从而提高摘要抽取的质量。但是，其没有达到 NLPCC 2017 评测任务的第一名的结果，主要原因如下。

1)本文使用的基于转移的中文篇章结构解析模型并非将所有标点都作为句子的边界标识，一个句子单元中可能包含多个子句，这使得抽取的摘要句中存在过多的冗余信息。

2)按照概率得分抽取出来的句子往往容易出现重复现象，即几个相似句子的得分同样高，导致摘要句中存在信息重复的情况。

(3)去除冗余实验

为了得到更高质量的摘要，本文对候选摘要句进行了简单的去冗实验。在按照预测句子的分布概率从高到低抽取摘要句的时候，后一个句子在被抽取之前分别与已抽取的句子计算余弦相似度，当余弦相似度大于 75%或句子长度小于 4 个字符时舍弃该句子。

模型加入去冗余之后的实验结果如表 5 所列。LSTM＋feature，CNN＋feature 和 NFS 的性能均在原有基础上提高了 0.5%左右，说明本文使用的去冗余方法可以在一定程度上整体优化模型的性能。NFS 的实验结果，依旧优于对比实验结果且超过了 NLP_ONE 的结果，说明本文提出的基于主次关系特征的自动文摘方法性能比较稳定，并且可以得到较高质量的摘要。

表 5 加入去冗余之后的实验结果

Table 5 Experiment results after de-redundancy

(单位：%)

Models	*RG*-1	*RG*-2	*RG*-*L*
LSTM＋feature	37.7	22.7	29.6
CNN＋feature	37.9	22.8	29.6
NFS	**38.4**	**23.2**	**30.2**

结束语 自动文摘是自然语言处理领域的一个重要研究方向，无数专家学者的持续性研究为该领域的发展打下了坚实的基础。但是，传统的文本摘要任务通常只考虑了词汇信息和浅层的文本语义信息，忽略了深层的篇章结构信息。因此，本文提出了一种基于主次关系特征的自动文摘方法，在考虑浅层文本信息的基础上，同时发挥深层的主次关系信息对

文摘任务的指导作用。实验结果表明：与当前主流的单文档抽取式摘要方法相比，该方法在 ROUGE 值的评测上有较大的提升。在接下来的工作中，一方面考虑将结构化的篇章主次信息直接加入文本摘要任务中，采用图神经网络模型处理结构化的主次关系信息，避免将结构化信息序列化过程中带来的信息损失。另一方面，考虑如何将篇章主次信息与生成式文本摘要任务相结合，进一步拓宽篇章结构信息在自动文摘领域的应用。

参 考 文 献

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

[2] CHU X M, ZHU Q M, ZHOU G D. Discourse Primary-Secondary Relationships in Natural Language Processing[J]. Journal of Computer Science, 2017, 40(4): 842-860.

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

[4] ERKAN G, RADEV D R. LexPageRank: Prestige in Multi-Document Text Summariztion [C] // Conference on Empirical Methodsin Natural Language Processing(EMNLP 2004). ACL, 2004: 365-371.

[5] KUPIEC J, PEDERSEN J, CHEN F. A trainable document summarizer[C] // Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1995: 68-73.

[6] CONROY J M, SCHLESINGER J D, DIANNE P O, et al. Using {HMM} and Logistic Regression to Generate Extract Summaries for {DUC}[J]. 温泉科学, 2001, 63(329): 66-75.

[7] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv: 1408. 5882, 2014.

[8] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning[J]. arXiv: 1605. 05101, 2016.

[9] ZHOU C, SUN C, LIU Z, et al. A C-LSTM neural network for text classification[J]. arXiv: 1511. 08630, 2015.

[10] 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.

[11] RUSH A M, CHOPRA S, WESTON J. A neural attention model for abstractive sentence summarization[J]. arXiv: 1509. 00685, 2015.

[12] SEE A, LIU P J, MANNING C D. Get To The Point: Summarization with Pointer-Generator Net-works[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1073-1083.

[13] CARLSON, LYNN, MARCU, et al. Building a dis-course-tagged corpus in the framework of Rhetorical Structure Theory[M]// Current and New Directions in Discourse and Dialogue. Netherlands: Springer, 2003: 2655-2661.

[14] ZHOU Y, XUE N. The Chinese Discourse TreeBank: A Chinese corpus annotated with discourse relations[J]. Language Resources and Evaluation, 2015, 49(2): 397-431.

[15] HERNAULT H, PRENDINGER H, ISHIZUKA M. HILDA: A discourse parser using support vector machine classification[J]. Dialogue & Discourse, 2010, 1(3).

[16] FENG V W, HIRST G. Text-level discourse parsing with rich linguistic features[C]// Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. 2012: 60-68.

[17] CHU X, WANG Z, ZHU Q, et al. Recognizing nuclearity between Chinese Discourse units[C]// 2015 International Conference on Asian Language Processing (IALP). IEEE, 2015: 197-200.

[18] SUN J, LI Y C, ZHOU G D, et al. Research of Chinese Implicit Discourse Relation Recognition[J]. Journal of Peking University (Natural Science Edition), 2014, 50(1): 111-117.

[19] SUN C, KONG F. A Transition-based Framework for Chinese Discourse Structure Parsing[J]. Journal of Chinese Information Processing, 2018, 32(12): 48-56.

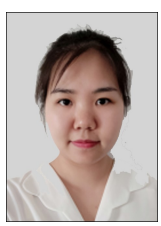
[20] LI Y C. Research of Chinese Discourse Stucture Representation and Resource Construction[D]. Suzhou: Soochow University of Defense Technology, 2015.

[21] LIN C Y, HOVY E. Automatic Evaluation of Summaries Using n-gram Co-occurrence Statistics[C]// Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics, 2003.

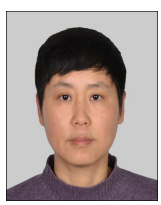
[22] HUA L, WAN X, LI L. Overview of the NLPCC 2017 Shared Task: Single Document Summarization[C]// National CCF Conference on Natural Language Processing and Chinese Computing. Cham: Springer, 2017: 942-947.

[23] WAN X, YANG J. Multi-document summarization using cluster-based link analysis[C]// Proceedings of the 31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2008: 299-306.

[24] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv: 1408. 5882, 2014.



ZHANG 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, Ph.D, associate professor, is a member of China Computer Federation. Her main research interests include natural language processing and so on.