主题关键词信息融合的中文生成式自动摘要研究

侯丽微 [1] 胡 珀 [1] 曹雯琳 [1]

摘 要 随着大数据和人工智能技术的迅猛发展, 传统自动文摘研究正朝着从抽取式摘要到生成式摘要的方向演化, 从中达到生成更高质量的自然流畅的文摘的目的. 近年来, 深度学习技术逐渐被应用于生成式摘要研究中, 其中基于注意力机制的序列到序列模型已成为应用最广泛的模型之一, 尤其在句子级摘要生成任务 (如新闻标题生成、句子压缩等) 中取得了显著的效果. 然而, 现有基于神经网络的生成式摘要模型绝大多数将注意力均匀分配到文本的所有内容中, 而对其中蕴含的重要主题信息并没有细致区分. 鉴于此, 本文提出了一种新的融入主题关键词信息的多注意力序列到序列模型, 通过联合注意力机制将文本中主题下重要的一些关键词语的信息与文本语义信息综合起来实现对摘要的引导生成. 在 NLPCC 2017 的中文单文档摘要评测数据集上的实验结果验证了所提方法的有效性和先进性.

关键词 联合注意力机制, 序列到序列模型, 生成式摘要, 主题关键词

引用格式 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. 自动化学报, 2019, **45**(3): 530−539

DOI 10.16383/j.aas.c170617

Automatic Chinese Abstractive Summarization With Topical Keywords Fusion

HOU Li-Wei[1] HU Po[1] CAO Wen-Lin[1]

Abstract With the rapid development of big data and artificial intelligence technology, the automatic text summarization research is evolving from extractive summarization to abstractive summarization, which aims to generate more natural, higher quality and more fluent summary. In recent years, deep learning technology has been gradually applied to the abstractive summarization task. The sequence to sequence model based on attention mechanism has become one of the most widely used models, especially in the sentence-level summarization generation tasks (such as news headline generation, sentence compression and so on), and has achieved remarkable results. However, most of the abstractive summarization models based on neural networks distribute their attention to all the contents of the source document evenly, instead of regarding the important topics information of source documents discriminatively. In view of this, we propose a new multiple attention sequence-to-sequence model which integrates topical keywords information. And this model combines multidimensional topic information with text semantic information to generate the final summary by a joint attention mechanism. The evaluation results on the public dataset of NLPCC 2017 shared task3 show that our system is competitive with the state-of-the-art methods.

Key words Joint attention mechanism, sequence to sequence model, abstractive summarization, topical keywords

Citation Hou Li-Wei, Hu Po, Cao Wen-Lin. Automatic Chinese abstractive summarization with topical keywords fusion. *Acta Automatica Sinica*, 2019, **45**(3): 530−539

自动摘要旨在从给定的文本中自动生成能表达原文主题的精简形式, 以缓解信息过载造成的阅读压力. 自动摘要过程大致可分为抽取式和生成式两类, 抽取式摘要从原文中选取若干重要句子直接组合成摘要, 生成式摘要的产生则相对自由灵活, 有望生成更接近人工撰写的流畅摘要, 并且在技术实现上更具挑战性.

目前, 随着大数据和人工智能技术的发展, 以及深度学习和表示学习在各个领域的推广渗透[1−3], 传统自动摘要方法逐渐从抽取式朝着生成式演化, 特别是基于循环神经网络 (Recurrent neural network, RNN) 的编码器–解码器模型正成为当前应用最广泛的生成式摘要模型, 并在句子级的摘要生成任务 (新闻标题生成、句子压缩等) 中取得了较显著的效果. 近年来, 已有学者如 Bahdanau 等[4] 提出在此模型的解码器部分加入对输入序列的注意力机制, 用于提取原始文本中丰富的上下文信息以避免信息覆盖问题, 导致该问题的原因是简单的 RNN 编码器–

收稿日期 2017-11-07 录用日期 2018-01-08
Manuscript received November 7, 2017; accepted January 8, 2018
国家自然科学基金 (61402191), 中央高校基本科研业务费项目 (CCNU18TS044, CCNU16JYKX15), 国家语委 “十三五” 科研规划项目 (WT135-11) 资助
Supported by National Natural Science Foundation of China (61402191), Fundamental Research Funds for the Central Universities (CCNU18TS044, CCNU16JYKX15), and Thirteen Five-year Research Planning Project of National Language Committee (WT135-11)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 华中师范大学计算机学院 武汉 430079
1. School of Computer Science, Central China Normal University, Wuhan 430079

解码器模型中的编码器是将信息从前到后一步步压缩成一个固定长度的上下文语义向量, 但这种信息传递编码方式会使得前面的信息被后面的信息覆盖而失效. 此外, 因为该机制将注意力均匀分布在文本的所有内容上, 因而使得全文中的主题信息在摘要的生成过程中并没有被合理地区分利用, 同时主题关键词是主题信息常见的表示形式. 鉴于此, 本文尝试提出了一种新的融合主题关键词信息的多注意力机制, 并融入到循环神经网络的编码器–解码器模型中以补充强化原文中的主题信息, 从而更好地引导摘要生成. 具体而言, 先使用无监督方法识别文本的主题关键词, 然后综合主题关键词注意力机制, 输入序列注意力机制及输出序列注意力机制三者联合辅助最终的摘要生成. 在 NLPCC 2017 的中文单文档摘要评测任务上, 本文提出的模型的实际摘要效果的 ROUGE (Recall-oriented understudy for gisting evaluation) 值比参赛队中第一名成绩还显著提高了 2～3 个百分点, 充分验证了本文模型的有效性和先进性.

1　相关工作

现有自动摘要方法主要分为抽取式和生成式. 抽取式摘要根据特定的约束条件 (如摘要长度) 直接从原文中抽取若干重要的句子, 这些句子经重新排序后组成摘要. 生成式摘要往往涉及对原文内容的语义理解和重构, 且多采用更灵活的表达方式 (如新词、复述等) 间接凝练出原文的主旨要点. 相比于抽取式摘要, 生成式摘要更接近人类撰写摘要的形式. 但由于生成式摘要通常需要复杂的自然语言生成技术, 因此过去的研究大多注重抽取式摘要模型设计或句子打分排序算法的设计.

抽取式摘要首先给文本中的每个句子依重要度打分, 然后根据此分数来对句子排序, 进而选出得分最高且冗余小的句子组成摘要. 现有方法中, 句子重要度计算通常会结合考虑各种统计学和语言学特征, 例如句子的位置、词频、词汇链等. 句子抽取则大致分为无监督和有监督两种, 其中无监督方法主要包括基于质心的方法[5]、基于图模型的方法[6−8]以及基于隐含狄利克雷分布 (Latent Dirichlet allocation, LDA) 主题模型的方法[9−10] 等, 有监督的方法则包括支持向量回归[11] 和条件随机场模型[12]等. 同时还有研究综合考虑了各种最优化的摘要生成目标函数, 例如整数线性规划[13]、子模函数最大化[14−15] 等. 除此之外, 还有些抽取式摘要研究结合了主题信息来辅助摘要的生成, 例如基于动态主题模型的 Web 论坛文档摘要[16], 也有研究提出使用超图模型来协同抽取文本关键词与摘要[17]. 同时, 有研究者还尝试了结合图像、视频以及文字来联合生成多模态的摘要[18].

生成式摘要更接近人类自然撰写摘要的方式, 是高级摘要技术的追求目标. 随着智能技术的发展以及数据量的不断增长, 当前对生成式摘要的需求和研究越来越多. 近几年, 神经网络模型在生成式摘要的一些具体任务 (如标题生成、单句式单文本摘要生成等) 上取得了一定的效果. Rush 等[19] 在一个大型语料库上训练了神经注意力模型并用于单句式摘要, 之后 Chopra 等[20] 在注意力机制的循环神经网络模型上扩展了 Rush 等的工作. Nallapati 等[21] 在基于循环神经网络的序列到序列模型上应用各种技术改善效果, 例如在解码器阶段采用的分层注意力机制和词表限制. Paulus 等[22] 将输出信息尝试融入到输出的隐藏层向量中, 以避免产生重复的信息, 同时提出使用强化学习的方式训练模型. Ma 等[23] 通过最大化原文本和摘要之间的语义相似性, 确保生成与原文本在语义上表达一致的摘要. Tan 等[24] 通过序列到序列模型与传统的图模型方法融合, 以增加对句子重要度的考虑来生成摘要. Li 等[25] 提出使用变分自动编码器 (Variational auto-encoder, VAE) 提取出生成的摘要中的高维信息, 然后让该信息辅助解码器对原文本进行注意力提取. 还有一些工作在注意力机制、优化方法和原文信息的嵌入等方面进行了改进[26−29]. 然而, 值得注意的是以上模型的注意力机制均仅限于均匀考虑整个原文本的所有信息而忽视了原文本中隐藏的重要主题信息的影响. 鉴于此, 本文提出将原文本中的主题关键词信息抽取出来, 并自然地融入神经网络中以更好地区分引导生成摘要, 模型中我们具体采用了多种注意力机制的联合策略.

2　背景模型: 序列到序列模型和注意力机制

2.1　序列到序列模型

序列到序列模型又称为编码器–解码器模型, 核心是利用 RNN 学习一个序列的所有信息, 并浓缩到一个向量中, 再利用另一个循环神经网络将此信息解码出来, 进而生成另一个序列. 具体结构如图 1 所示.

现有的实践发现[30], 门控 RNN 比简单 RNN 效果更好, 如长短期记忆 (Long short-term memory, LSTM), 双向门控 RNN 比单向 RNN 效果好. 因此, 本研究提出的模型在编码阶段采用了双向 LSTM, 在解码阶段采用了单向 LSTM.

该模型的基本目标是优化生成词的条件概率 $p(\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{t'} | \boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t)$, 其中, $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t\}$ 为输入词向量序列, $\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_{t'}\}$ 为对应的输出词向量序列. 序列到序列模型的具体

目标为

$$p(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_{t'}|\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_t)=\prod_{i=2}^{t'}p(\boldsymbol{y}_i|c,\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_{i-1}) \tag{1}$$

其中, 编码器作用是将输入文档的信息映射为一个上下文语义向量 $\boldsymbol{c}$, 每一个 $p(\boldsymbol{y}_i|\boldsymbol{c},\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_{i-1})$ 表示最新生成的词是由前 $i-1$ 个词联合嵌入的上下文语义向量 $\boldsymbol{c}$ 生成的. 具体过程为先对每一个文档 d 进行分词, 每个词 w 被 gensim 工具包[1] 中的 word2vec 训练为一个向量以作为输入序列. 在该阶段, 每个输入序列通过 LSTM 生成一列蕴含高维信息的隐藏层向量 $\boldsymbol{H}^e=\{\boldsymbol{h}_1^e,\boldsymbol{h}_2^e,\cdots,\boldsymbol{h}_t^e\}$. 接下来, 通过这些隐藏层向量来计算上下文语义向量 $\boldsymbol{c}$, 具体计算方式为

$$\boldsymbol{h}_j^e=f(\boldsymbol{x}_j,\boldsymbol{h}_{j-1}^e),\quad \boldsymbol{c}=\boldsymbol{h}_t^e \tag{2}$$

其中, $\boldsymbol{h}_j^e$ 是在编码器第 j 时刻的隐藏状态向量, f 是 LSTM 的非线性方程.

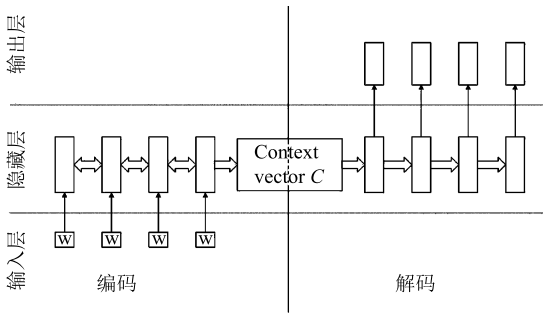

图 1 序列到序列模型

Fig. 1 The sequence-to-sequence model

解码器的作用是生成输出序列. 在此阶段, 解码器利用编码器压缩后的语义向量 $\boldsymbol{c}$ 结合当前时间点解码器隐藏层的输出状态 $\boldsymbol{h}_{t'}^d$ 以及上一时间点中的输出词 $\boldsymbol{y}_{t'-1}$ 来生成候选词, 具体的条件概率计算方式为

$$\boldsymbol{h}_{t'}^d=f(\boldsymbol{y}_{t'-1},\boldsymbol{h}_{t'-1}^d,\boldsymbol{c}) \tag{3}$$

$$p(\boldsymbol{y}_{t'}|\boldsymbol{y}_{<t'},X)=g(\boldsymbol{y}_{t'-1},\boldsymbol{h}_{t'}^d,\boldsymbol{c}) \tag{4}$$

其中, $\boldsymbol{h}_{t'}^d$ 是在 t' 时刻的隐藏状态向量, $\boldsymbol{y}_{t'-1}$ 是 $\boldsymbol{y}$ 在 $t'-1$ 时刻的输出向量, $\boldsymbol{y}_{<t'}$ 是 $\boldsymbol{y}$ 在 t' 时刻之前的所有时间段生成的输出向量, g 是 softmax 函数.

上述编码器–解码器模型虽然经典, 但局限性也很明显. 由于解码器从编码器中获取信息的唯一途径是一个固定长度的上下文语义向量 $\boldsymbol{c}$, 因而编码器需要将整个原文本的信息压缩到一个固定长度的向量中, 由此导致了三个弊端: 1) 仅靠一个固定长度的上下文语义向量往往无法完整地表示整个文本的全部信息, 因而自然会影响解码器的信息解码效果; 2) 由式 (2) 可知, 一般上下文语义向量 $\boldsymbol{c}$ 是由编码器最后一个 LSTM 输出的隐藏层状态向量 $\boldsymbol{h}_t^e$ 获取的, 因此在编码器阶段先输入的内容所携带的信息会被后输入的信息稀释或覆盖, 且输入序列越长, 这个现象越严重; 3) 由图 (1) 可见, 解码器在所有时间点上都共享了同一个固定长度的上下文语义向量 $\boldsymbol{c}$, 因此解码器生成的序列信息不足且固化, 更合理的情况应该是解码器能根据输入序列 $\boldsymbol{x}$ 中不同部分的不同语义信息来生成不同的输出结果 $\boldsymbol{y}$. 为了解决上述问题, Bahdanau 等[4] 提出了在序列到序列的模型中加入注意力机制, 该机制能在一定程度上缓解这些问题.

2.2 注意力机制

引入注意力机制不仅为了减轻基本序列到序列模型中上下文语义向量 $\boldsymbol{c}$ 的信息负担, 还要对后续生成内容有针对性地生成一组对应的注意力权重以改进模型的实际生成效果, 具体结构如图 2 所示.

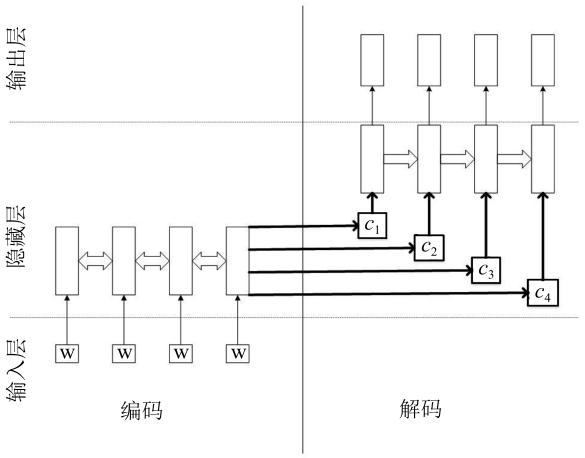

图 2 注意力机制

Fig. 2 The attention mechanism

由图 2 可知, 在解码过程中, 注意力机制使用动态改变的上下文语义向量来获取编码器中的原文语义信息, 当生成每一个词 $\boldsymbol{y}_i$ 的时候, 编码器会动态产生与之对应的语义向量 $\boldsymbol{c}_i$. 这里的关键是如何定义不同解码时间的注意力系数 α_{ij}, 具体为

$$\alpha_{ij}=\frac{\exp(\eta(\boldsymbol{h}_i^d,w,\boldsymbol{h}_j^e))}{\sum\limits_{k=1}^{t}\exp(\eta(\boldsymbol{h}_i^d,w,\boldsymbol{h}_k^e))} \tag{5}$$

[1] http://radimrehurek.com/gensim/

其中, η 由一个多层感知器实现, 采用 tanh 作为激活函数. $\boldsymbol{h}_i^d$ 代表在解码阶段时间 i 的 LSTM 隐藏层向量, $\boldsymbol{h}_j^e$ 代表在编码阶段时间 j 的 LSTM 隐藏层向量, w 为注意力权重矩阵.

通过上述公式计算得到注意力系数之后, 便可结合编码器中所有隐藏层向量和注意力系数生成解码阶段时间 i 的上下文语义向量 $\boldsymbol{c}_i$, 具体为

$$\boldsymbol{c}_i = \sum_{j=1}^{t} \alpha_{ij} \boldsymbol{h}_j^e \tag{6}$$

由于在每个时间点, 解码器会根据当前解码器的隐藏层向量来引导编码器产生的上下文语义向量 $\boldsymbol{c}_i$ 生成对应的输出序列, 因此在生成摘要的某个部分时, 注意力机制将帮助模型选择与此部分高度相关的原文信息, 进而有望生成更好的相关摘要内容.

通常训练好一个序列到序列模型需要较大规模的数据, 在数据量相对较少的情况下可能存在效果欠佳的情况, 在文本摘要领域, 虽然注意力机制的引入在一定程度上解决了一些问题并提升了模型的效果, 但生成的摘要离人类撰写的摘要还有一定差距, 因此如何将文本更深层次的信息有效地嵌入到模型中来生成更好的摘要仍需继续研究. 为了解决上述问题, 本文提出在序列到序列模型中引入主题关键词信息来优化现有生成式摘要模型的效果, 并且提出了一种新的融入主题关键词信息的多注意力序列到序列模型, 通过联合注意力机制将文本中多维重要信息综合起来实现对摘要的引导性生成. 通过在 NLPCC 2017 中文单文档摘要评测数据集上的实验, 本文提出的模型非常有效. 目前, 在生成式摘要领域, 融合主题关键词信息以联合注意力方式优化摘要生成效果的设计思路尚未见文献报道.

3　提出的模型: 主题关键词信息融合的多注意力序列到序列模型

在现有模型基础上, 本文提出采用联合多注意力融合机制以提升摘要生成效果, 模型的具体结构如图 3 所示.

本节将对图 3 中重要标识部分 (主题关键词注意力机制和输入输出信息注意力机制) 进行详细介绍, 首先介绍主题关键词抽取, 然后对模型中主题关键词注意力机制进行详细的说明. 最后对模型中的输入输出信息注意力机制进行简要的介绍.

3.1　主题关键词注意力机制

1) 主题关键词抽取

按照认知科学的观点, 人类必须先识别、学习和理解文本中的实体或概念, 才能理解自然语言文本, 而这些实体和概念大都是由文本句子中的名词或名词短语描述的[31]. 因此本文通过发掘文章中的重点实体和概念来辅助模型理解自然语言文本. 一个词在文本中出现的频率越高, 产生的效力就越强, 对文本的表达能力也越强, 而这些实体或概念就称为文本的关键词. 文本的主题关键词表征了文档主题性和关键性的内容, 是文档内容理解的最小单位[32]. 因此本文提出将主题关键词信息融入到序列到序列模型中以实现在主题信息引导下的摘要生成.

本文使用的主题关键词抽取方法为 HanLP 开源工具包提供的主题关键词提取算法 TextRank[2], 并对每个文档提取出 10 个最重要的主题关键词. TextRank[33] 是一种基于图模型的主题关键词抽取

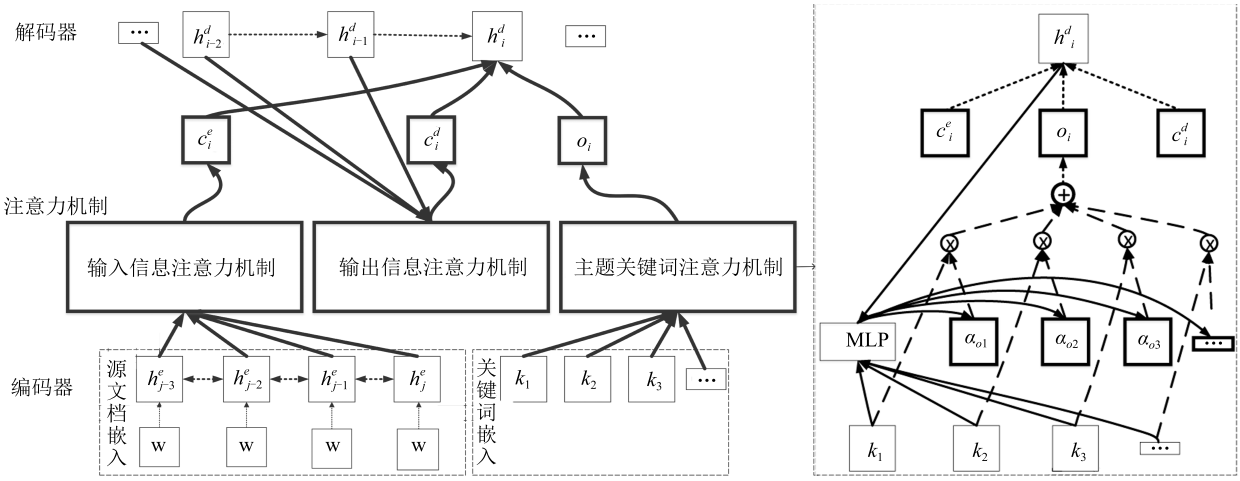

图 3　主题关键词信息融合的多注意力序列到序列模型

Fig. 3　The multi-attention sequence-to-sequence model based on keywords information

[2]http://hanlp.linrunsoft.com/doc/_build/html/extract.html#extract-keyword

算法, 基本思想源自谷歌的 PageRank 算法, 核心是利用投票机制迭代计算图中每个结点的全局得分, 然后取出得分最高的若干词作为主题关键词. 与 LDA 和隐马尔科夫模型 (Hidden Markov model, HMM) 等模型不同, TextRank 不需要事先对多篇文本进行学习训练, 因简洁有效获得了较广泛的应用.

2) 主题关键词注意力机制实现

人类撰写文章或摘要, 都会预先设定一些内容框架并提取重要的实体信息, 然后根据框架和实体信息构建语言. 受此启发, 本文通过自动提取文本的主题关键词组成一个文本的框架, 然后将模型对文本的注意力引到这些预先提取的主题关键词信息上, 由此生成基于主题信息的摘要.

图 3 右半部分对主题关键词注意力机制的基本结构进行了直观呈现. 该机制将提取出的主题关键词通过注意力机制融入到模型中, 通过主题关键词中蕴含的语义信息来引导模型生成更完善的摘要.

在编码阶段, 由于原文本的输入形式是使用 word2vec 训练得到的词向量, 因此为了保持词嵌入信息的一致性, 对从原文中抽取出的主题关键词, 直接利用 word2vec 训练出的词向量 $\boldsymbol{k} = \{\boldsymbol{k}_1, \boldsymbol{k}_2, \cdots, \boldsymbol{k}_n\}$ 作为输入, 其中 n 为主题关键词的数量.

在解码阶段, 主题关键词注意力机制通过解码器解码当前输出的 LSTM 隐藏层状态向量中的信息来获得对所有主题关键词信息的不同注意程度. 通过此机制, 该模型在生成摘要的过程中能自然融入文本中的主题信息来生成基于主题引导式摘要. 本文提出的主题关键词注意力机制中注意力系数的具体计算方法为

$$\alpha_{ij}^{k} = \frac{\exp(\eta_o(\boldsymbol{h}_i^d, w_k, \boldsymbol{k}_j))}{\sum\limits_{a=1}^{n} \exp(\eta_o(\boldsymbol{h}_i^d, w_k, \boldsymbol{k}_a))} \tag{7}$$

其中, α_{ij}^{k} 表示在解码器时间点 i, 主题关键词注意力机制根据解码器的隐藏层向量 $\boldsymbol{h}_i^d$ 对第 j 个主题关键词分配的注意力系数, η_o 是一个多层感知器, 采用 tanh 作为激活函数, 此多层感知器由解码器解码当前的隐藏层状态向量作为输入来对编码器嵌入的主题关键词的重要性进行感知, 然后通过一个 softmax 函数获得当前解码器时间点 i 对主题关键词的注意力系数, n 表示主题关键词个数, $\boldsymbol{k}_j$ 表示第 j 个主题关键词的向量表示, w_k 表示注意力权重矩阵.

通过式 (7) 得到当前解码器时间点 i 对主题关键词的注意力系数后, 便可结合主题关键词的嵌入向量 $\boldsymbol{k} = \{\boldsymbol{k}_1, \boldsymbol{k}_2, \cdots, \boldsymbol{k}_n\}$ 生成上下文语义向量 $\boldsymbol{o}_j$, 具体为

$$\boldsymbol{o}_i = \sum_{j=1}^{n} \alpha_{ij}^{k} \boldsymbol{k}_j \tag{8}$$

其中, n 表示主题关键词个数, $\boldsymbol{k}_j$ 表示第 j 个主题关键词的向量表示.

3.2 输入输出注意力机制

图 2 中的输入输出注意力机制是将输入序列和输出序列的注意力结合起来共同嵌入到解码器当前时间点的输出序列中, 这样既能考虑输入序列的信息, 又可以通过对输出序列信息的回顾来避免信息的冗余和重复.

输入序列的注意力机制将原文中隐含的信息提取出来嵌入到输出序列中, 其上下文语义向量表示为 $\boldsymbol{c}_i^e$.

输出序列注意力机制与输入序列注意力机制的实现方式类似, 但意义不同, 解决的问题也不同. 由于注意力机制的序列到序列模型在生成摘要的过程中存在重复信息的问题, 而在该模型中加入对输出序列的注意力机制可在一定程度上缓解此问题, 因此, 本模型也一并加入了输出序列的注意力机制来优化摘要的生成结果, 具体实现方法为

$$\alpha_{ij}^{d} = \frac{\exp(\eta(\boldsymbol{h}_i^d, w_d, \boldsymbol{h}_j^d))}{\sum\limits_{k=1}^{i-1} \exp(\eta(\boldsymbol{h}_i^d, w_d, \boldsymbol{h}_k^d))} \tag{9}$$

$$\boldsymbol{c}_i^d = \sum_{j=1}^{i-1} \alpha_{ij}^{d} \boldsymbol{h}_j^d \tag{10}$$

其中, $\boldsymbol{h}_i^d$ 表示解码器当前时间点 i 的隐藏层向量, w_d 表示注意力权重矩阵.

3.3 多种注意力融合

在获得主题关键词注意力和输入, 输出注意力之后, 我们将这两种注意力联合嵌入当前解码器的输出向量中以获得输出词的条件概率. 通过此方法, 输出向量中不仅包含输出序列的信息, 也自然融入了原文本中的语义信息以及主题关键词信息, 结合这些信息有望输出更优质的摘要. 为了不加重网络的训练负担, 本文仅采用线性加和的方式将多种注意力机制获得的上下文语义向量融合到一起, 实验证明该种融合方式有效. 具体融合方式为: 先利用线性组合将三个注意力机制获得的上下文语义向量联合嵌入到解码器的第 i 个时间点隐藏状态中, 然后使用 softmax 层得出词表中词的输出概率, 具体计算方法为

$$p(\boldsymbol{y}_i | \boldsymbol{y}_{<i}, \boldsymbol{X}) = \text{softmax}(L(\boldsymbol{h}_i^d, \boldsymbol{o}_i, \boldsymbol{c}_i^d, \boldsymbol{c}_i^e)) \tag{11}$$

其中, L 表示线性组合 linear, $\boldsymbol{h}_i^d$ 表示解码器当前时间点 i 的隐藏层向量, $\boldsymbol{o}_i$ 表示文章主题关键词通过主题关键词注意力机制计算得出的上下文语义向量, $\boldsymbol{c}_i^d$ 表示之前所有输出向量通过输出信息注意力机制计算得出的上下文语义向量, $\boldsymbol{c}_i^e$ 表示输入向量通过输入信息注意力机制计算得出的上下文语义向量.

4　实验

4.1　数据集

本研究的实验语料采用 NLPCC 2017 的中文单文档摘要评测数据集, 此数据集是今日头条提供的公开新闻数据, 包括 50 000 个文本–摘要对, 每篇文章的长度从 10～10 000 个中文字符不等, 每篇摘要的长度不超过 60 个中文字符. 在实验中, 将其中 49 500 个文本–摘要对作为训练集和验证集, 另外 500 个作为测试集.

4.2　评价标准

评价方法采用自动摘要领域常用的基于召回率统计的摘要评价工具 ROUGE (Recall-oriented understudy for gisting evaluation)[34]. ROUGE 由 ISI 的 Lin 和 Hovy 提出, 基于机器摘要和人工标准摘要中的 n 元词 (即 n-gram) 匹配情况来生成量化的评价结果. ROUGE 指标由一系列具有细微差别的计算方法组成, 包括 ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L 等. ROUGE-1.5.5 工具包已被 DUC 和 TAC 等国际著名的文本摘要评测会议作为标准的评价工具采用.

本实验使用了 ROUGE 的五类评价指标, 分别为 ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4 和 ROUGE-L. 直观看, ROUGE-1 可以代表自动摘要的信息量, ROUGE-2、ROUGE-3 以及 ROUGE-4 则侧重于评估摘要的流畅性, 而 ROUGE-L 可看成是摘要对原文信息的涵盖程度的某种度量. 其中 ROUGE-N 的计算方法为

$$\text{ROUGE-}N = \frac{\sum\limits_{S\in RS}\sum\limits_{g_n\in S} Count_m(g_n)}{\sum\limits_{S\in RS}\sum\limits_{g_n\in S} Count(g_n)} \tag{12}$$

其中, RS 表示参考摘要, 该摘要为人工生成的标准摘要. g_n 表示 n 元词, $Count_m(g_n)$ 表示系统生成的摘要和标准摘要中同现的相同 n-gram 的最大数量, $Count(g_n)$ 表示标准摘要中出现的 n-gram 个数.

ROUGE 为每类评价指标分别计算了准确率 P、召回率 R 和 F 值 (其中, $F = 2PR/(P+R)$), 由于 F 值综合考虑了评价指标的准确率和召回率, 因此本文统一将 F 值[3] 作为实验的最终结果汇报.

4.3　实验步骤

在数据预处理过程中, 使用 jieba[4] 开源分词工具对文本进行分词, 再用 subword 模型[5] 对分词后的数据进行更细致的切分. 通过这些操作, 最终形成包含 28 193 个中文词的词典. 实验中采用 subword 模型可以减小词表的大小, 同时解决序列到序列模型中常遇到的罕见词问题 (即 UNK 问题). 为了使词内信息得到合理的保存, 本文使用的 subword 模型仅对词内信息进行切分和重组而不组合词间信息, 因而先将分词后的词语以每个词为单位切分成字, 然后使用 subword 模型将该结果使用 2-gram 的方法抽取出频率较大的词内组合, 将此组合从之前的词中分离出来独立变为一个词. 采用此方法可以极大地减少字典的冗余度, 同时保留部分词信息, 最终的分词结果为词组和字的混合文本.

接下来, 利用 gensim 工具包中的 word2vec 对词典中的每个词进行词嵌入训练, 训练集为 NLPCC 2017 的中文单文档摘要评测任务分享的全部数据集, 每个词的向量维度均设置为 256 维, 通过预训练可以在一定程度上优化模型的效果.

本文使用 tensorflow 实现了基于主题关键词注意力的序列到序列模型, 编码器层为一层双向的 LSTM, 解码器为一层单向的 LSTM. LSTM 隐藏层维度设为 128. 在训练阶段, 本文使用的优化函数为 Adam[35], 学习率设置为 0.001, 并在训练过程中利用梯度的一阶矩估计和二阶矩估计动态调整每个参数的学习率, 最小学习率设置为 0.0001, 在训练过程中, 损失开始下降很快, 训练几轮后, 损失逐渐趋于平稳, 且在几百个 batch 内损失值固定在 1 点多的范围内, 数值不继续降低, 模型趋于收敛.

在测试阶段, 使用 beamsearch 方法生成最终的结果. beamsearch 是一种启发式搜索算法, 是对优先搜索策略的一种优化, 能降低内存需求, 根据启发式规则对所有局部解进行排序, 以找到与全局解最近的局部解, 此方法常应用于序列到序列模型中优化解的生成.

4.4　对比模型

选取 5 种基准模型与本文提出的模型进行比较, 5 种方法中的 3 种方法 (即 LexPageRank[6], MEAD[36] 和 Submodular[15]) 是目前最典型的抽取式摘要方法, 由开源工具包 PKUSUMSUM[37] 提供, 另外两种方法是生成式摘要方法的代表.

[3] ROUGE 参数: -n 4 -U -z SPL -l 60
[4] https://pypi.python.org/pypi/jieba/
[5] https://github.com/rsennrich/subword-nmt

LexPageRank[6] 是一个基于图模型的摘要算法, 它将 PageRank 算法应用到文本句子关系表示及摘要抽取中.

MEAD[36] 方法则通过联合考虑句子的 4 种常用特征来为其打分, 包括质心、句子的位置、公共子序列及关键词.

Submodular[15] 方法利用子模函数的收益递减特性来挑选重要句子生成摘要.

UniAttention[20] 是基本的注意力序列到序列模型, 实现了对原文本输入信息的注意力机制考虑及摘要生成.

NLP_ONE[38] 是在 NLPCC 2017 的中文单文档摘要评测任务中获得第一名的参赛模型, 包含了输入序列的注意力机制和输出序列的注意力机制, 但它没有对主题关键词信息进行融合考虑.

pointer-generator[29] 是 ACL 2017 公开发表的一个最新的同类模型, 使用 pointer 机制解决了输出信息错误和罕见词的问题.

4.5 实验结果分析

第 4.4 节中的 5 种模型与本文提出模型的具体实验结果比较如表 1 所示. 由表 1 的结果可见:

1) 生成式摘要方法在 ROUGE 的 F 值比较中比抽取式摘要方法平均高 4～10 个百分点, 这说明在自动生成短文本的摘要任务中, 生成式方法更有效.

2) 由 UniAttention 模型与本文模型的对比结果可见, 将文本关键词的注意力信息和输入输出序列的注意力信息共同融入到序列到序列模型中可以显著地提高模型的摘要效果 (具体可提升 3～4 个百分点).

3) 本文对 NLP_ONE, pointer-generator 和本文模型的实验数据进行了统计显著性分析, 发现结合主题关键词信息和原文本中多维信息来引导摘要生成能有效地提高现有基于 RNN 注意力机制的生成式摘要模型的摘要效果, 充分说明主题关键词信息在生成式摘要中发挥了积极的引导作用.

4) 本文所提模型产生其摘要的实际效果举例如表 2 所示. 表 2 展示了从 3 个序列到序列模型生成的摘要中抽取的 5 例摘要, 从表 2 可以看出, 生成式摘要技术尽力去学习和模拟人类撰写摘要的方法, 生成的摘要根据需要表达的主题信息和语义信息引导词语组合而成, 而不仅仅由抽取的句子简单拼凑而成, 因而在生成短文本摘要时, 相比抽取式摘要, 生成式摘要的文本流畅性、句间连贯性以及信息丰富性均更胜一筹.

5) 对比表 2 中的机器自动生成摘要的内容可以发现: 本文提出的模型在学习摘要的生成过程中, 更注重内容信息的表达, 同时也抓住了文本中的关键主题信息, 使生成的摘要的信息量更充足. 在同等数据集的条件下, 相比未融入主题信息的序列到序列模型, 本文提出的模型效果更优, 因为该模型将更多的主题信息显式提取出来用于指导摘要的生成, 特别是主题关键词信息协助模型更有针对性地选择与主题相关的词语来构成摘要.

4.6 存在的问题

根据实验结果, 尽管生成式摘要相比抽取式摘要在中文短文本摘要生成任务中效果较好, 但仍需相对较大的数据来协助训练以生成高质量的摘要. 通过对实验数据的细致分析可以发现: 由于数据分布不均匀使得模型对训练样本较多的内容其学习效果比数量较少的内容学习效果好. 虽然主题关键词的融入在内容上对文本的信息进行了补充, 使得生成的摘要可以抓住文章的重点信息, 但在表达的流畅度方面, 样本量越大往往效果越好. 例如表 2 中原文为天气和受贿内容的生成摘要比其他类型的摘要生成效果好, 若训练样本充足, 则生成的摘要和原标

表 1 摘要评价结果

Table 1 The results of summaries

方法	ROUGE-1	ROUGE-2	ROUGE-3	ROUGE-4	ROUGE-L
LexPageRank	0.23634	0.10884	0.05892	0.03880	0.17578
MEAD	0.28674	0.14872	0.08761	0.06124	0.22365
Submodular	0.29704	0.15283	0.08917	0.06254	0.21668
UniAttention	0.33752	0.20067	0.13215	0.10178	0.29462
NLP_ONE	0.34983	0.21181	0.14490	0.11266	0.30686
pointer-generator	0.36022	0.21978	0.14783	0.11458	0.29888
本文模型	**0.37667**	**0.24077**	**0.16665**	**0.12914**	**0.32886**

表 2　生成摘要对比示例

Table 2　The examples of the generative summaries

对比模型	生成结果
标准摘要:	昨天下午, 山西平遥县 6 名儿童结伴滑冰玩耍时, 不慎溺水身亡, 其中年龄最大的 11 岁, 最小的为 5 岁.
UniAttention:	今日下午, 山西平遥县发生一起溺水事件, 6 名儿童玩耍不慎溺水身亡.
NLP_ONE:	今晨, 山西平遥县发生意外溺水身亡, 最小为 5 岁, 最小为 5 岁, 最小为 5 岁.
pointer-generator:	快讯: 平遥县发生一起意外溺水事件, 已致 1 死 1 伤, [UNK] 最小的岁, 最小为 5 岁 (图)
本文模型:	组图: **平遥县 6 名儿童结伴滑冰玩耍不慎**落水, **其中年龄最大的 11 岁**, 最小的为 5 岁, **最小的为 5 岁**.
标准摘要:	石嘴山市发布雷电黄色预警: 预计未来 6 小时, 石嘴山市有雷阵雨, 并伴有雷电和短时强降水. 提请相关部门和人员做好防范工作 · · ·
UniAttention:	石嘴山市发布雷电黄色预警: 预计未来 6 小时, 石嘴山市有雷阵雨, 并伴有雷电和短时强降水, · · ·
NLP_ONE:	石嘴山市发布雷电黄色预警: 预计未来 6 小时, 石嘴山市有雷阵雨, 并伴有雷电和短时强降水. 提请相关部门和人员做好防范工作
pointer-generator:	石嘴山市发布雷电黄色预警: 预计未来 6 小时石嘴山市有雷阵雨, 并伴有雷电和短时强降水. 提请相关部门和人员做好防范工作. · · · (· · ·)
本文模型:	**石嘴山市发布雷电黄色预警: 预计未来 6 小时, 石嘴山市**伴**有雷电和短时强降水. 提请相关部门和人员做好防范工作** · · ·
标准摘要:	今日上午 10 时, 郑州一个汽车销售点停车场发生火灾, 70 余辆新车被烧毁. (图)
UniAttention:	郑州: 70 余辆汽车结算着火, 70 余辆汽车被烧毁,70 余辆汽车被烧毁.
NLP_ONE:	组图: 郑州一汽车购点停车场起火, 包括名媛 43 辆, 包括名媛 43 辆, 或因电线破损打火, 火星沉地面干草中所致.
pointer-generator:	组图: 郑州一汽车销售点停车场失火, 70 余辆汽车被烧毁, 70 余辆汽车被烧毁, 郑州一汽车销售点停车场失火 (图)
本文模型:	组图: 今日 10 时许, **郑州一汽车销售时停车场**失火, 70 **余辆汽车被烧毁**.
标准摘要:	呼和浩特原市委副书记, 市长汤爱军涉嫌受贿犯罪被立案侦查, 案件侦查工作正在进行中.
UniAttention:	滁州市原市委书记汤爱军涉嫌受贿, 受贿犯罪被立案侦查, 并采取强制措施.
NLP_ONE:	原海东工业园区原市委书记江山涉嫌受贿犯罪被立案侦查, 并采取强制措施.
pointer-generator:	中纪委网站消息, 呼和浩特原市委副书记, 管委会常务副主任, 管委会市长张丽菊等四名厅级干部受贿犯罪立案侦查, 被证监会涉嫌滥用职权, 受贿犯罪被逮捕.
本文模型:	**呼和浩特原市委**书记江山**涉嫌受贿犯罪**, 已**被立案侦查 案件侦查正在进行中**.
标准摘要:	今日 6 时 50 分许, 西宁城北区一 12 岁男孩坠楼身亡; 3 小时后, 城中区一 46 岁男子坠楼身亡.
UniAttention:	西宁: 12 岁男孩从 21 层楼坠下身亡, 警方正在对两人死因展开调查.
NLP_ONE:	今晨 6 时 50 分, 城北区一 12 岁男孩坠楼身亡 (图).
pointer-generator:	西宁一 12 岁男孩坠楼身亡, 一名 12 岁男孩城中区小区 14 号楼坠楼者死因展开调查; 此前 12 岁男孩 20 岁男生是从 20 层的家中坠落.
本文模型:	组图: **今晨 6 时 50 分许**, **城北区**民惠城内12 **岁男孩坠楼身亡**, 仅3 **小时后**, 其车速 3 小时后**坠楼身亡**.
标准摘要:	达州一煤矿发生瓦斯爆炸事故 4 人被困井下, 1 人受伤, 相关部门正在全力救援被困人员.
UniAttention:	组图: 达州茶园煤矿发生爆炸事故, 造成 4 人被困井下, 伤者已送救援人员.
NLP_ONE:	今日下午发生瓦斯爆炸事故, 致 4 人被困井下,1 人被困井下, 无生命危险.
pointer-generator:	成都: 境内境内境内茶园煤矿生产系统工程瓦斯爆炸事故, 造成 4 人被困井下, 1 人被困井下, 1 人受伤, 1 人受伤 (图)
本文模型:	组图: **达川**区**发生瓦斯爆炸事故**, 4 **人被困井下**, 1 **人受伤**, 伤者已送达州医院救治.

注: 粗体是本文模型与标准摘要可完全匹配的词

准摘要在内容和表达上均能达到 90 % 以上的匹配度. 因而在训练数据量有限的情况下, 如何更好地生成拟人式高质量摘要仍是需要进一步深入探索的问题.

5　结束语

本文提出了一种新的基于神经网络的生成式中文自动摘要方法, 不仅融入了对输入序列的注意力及输出序列的注意力的区分性考虑, 还自然嵌入了文本中的关键主题信息下的注意力, 最终的实验及评价结果证实了引入关键词信息对提升中文生成式摘要模型的显著效果. 未来尚有很多可以拓展的工作, 例如在 LCSTS 等中文大规模文摘数据集上进行实验, 将神经网络模型应用到多文档多句子式的生成摘要中, 以及如何更有效地提取文本中全局和局部的不同粒度或不同模态的关键主题信息.

References

1 Chen Wei-Hong, An Ji-Yao, Li Ren-Fa, Li Wan-Li. Review on deep-learning-based cognitive computing. *Acta Automatica Sinica*, 2017, **43**(11): 1886−1897
(陈伟宏, 安吉尧, 李仁发, 李万里. 深度学习认知计算综述. 自动化学报, 2017, **43**(11): 1886−1897)

2 Xi Xue-Feng, Zhou Guo-Dong. A survey on deep learning for natural language processing. *Acta Automatica Sinica*, 2016, **42**(10): 1445−1465
(奚雪峰, 周国栋. 面向自然语言处理的深度学习研究. 自动化学报, 2016, **42**(10): 1445−1465)

3 Liu Kang, Zhang Yuan-Zhe, Ji Guo-Liang, Lai Si-Wei, Zhao Jun. Representation learning for question answering over knowledge base: an overview. *Acta Automatica Sinica*, 2016, **42**(6): 807−818
(刘康, 张元哲, 纪国良, 来斯惟, 赵军. 基于表示学习的知识库问答研究进展与展望. 自动化学报, 2016, **42**(6): 807−818)

4 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint, arXiv: 1409.0473, 2014.

5 Radev D R, Jing H Y, Styś M, Tam D. Centroid-based summarization of multiple documents. *Information Processing & Management*, 2004, **40**(6): 919−938

6 Erkan G, Radev D R. LexPageRank: prestige in multi-document text summarization. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: DBLP, 2004. 365−371

7 Wan X J, Yang J W, Xiao J G. Manifold-ranking based topic-focused multi-document summarization. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad, India: Morgan Kaufmann Publishers Inc, 2007. 2903−2908

8 Ji Wen-Qian, Li Zhou-Jun, Chao Wen-Han, Chen Xiao-Ming. Automatic abstracting system based on improved lexRank algorithm. *Computer Science*, 2010, **37**(5): 151−154
(纪文倩, 李舟军, 巢文涵, 陈小明. 一种基于 LexRank 算法的改进的自动文摘系统. 计算机科学, 2010, **37**(5): 151−154)

9 Titov I, McDonald R. A joint model of text and aspect ratings for sentiment summarization. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics. Columbus, Ohio, USA: Association for Computational Linguistics, 2008. 308−316

10 Hirao T, Yoshida Y, Nishino M, Yasuda N, Nagata M. Single-document summarization as a tree knapsack problem. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, Washington, USA: Association for Computational Linguistics, 2013. 1515−1520

11 Li S J, Ouyang Y, Wang W, Sun B. Multi-document summarization using support vector regression. In: Proceedings of the 2007 Document Understanding Workshop (Presented at the HLT/NAACL). Rochester, New York, USA, 2007.

12 Nishikawa H, Arita K, Tanaka K, Hirao T, Makino T, Matsuo Y. Learning to generate coherent summary with discriminative hidden Semi-Markov model. In: Proceedings of the 25th International Conference on Computation Linguistics. Dublin, Ireland: Association for Computational Linguistics, 2014. 1648−1659

13 Gillick D, Favre B. A scalable global model for summarization. In: Proceedings of the 2009 NAACL HLT Workshop on Integer Linear Programming for Natural Language Processing. Boulder, Colorado, USA: Association for Computational Linguistics, 2009. 10−18

14 Li J X, Li L, Li T. Multi-document summarization via submodularity. *Applied Intelligence*, 2012, **37**(3): 420−430

15 Lin H, Bilmes J. Multi-document summarization via budgeted maximization of submodular functions. In: Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Los Angeles, CA, USA: Association for Computational Linguistics, 2010. 912−920

16 Ren Zhao-Chun, Ma Jun, Chen Zhu-Min. Web forum thread summarization based on dynamic topic modeling. *Journal of Computer Research and Development*, 2012, **49**(11): 2359−2367
(任昭春, 马军, 陈竹敏. 基于动态主题建模的 Web 论坛文档摘要. 计算机研究与发展, 2012, **49**(11): 2359−2367)

17 Mo Peng, Hu Po, Huang Xiang-Ji, He Ting-Ting. A hypergraph based approach to collaborative text summarization and keyword extraction. *Journal of Chinese Information Processing*, 2015, **29**(6): 135−140
(莫鹏, 胡珀, 黄湘冀, 何婷婷. 基于超图的文本摘要与关键词协同抽取研究. 中文信息学报, 2015, **29**(6): 135−140)

18 Peng Di-Chao, Liu Lin, Chen Guang-Yu, Chen Hai-Dong, Zuo Wu-Heng, Chen Wei. A novel approach for abstractive video visualization. *Journal of Computer Research and Development*, 2013, **50**(2): 371−378
(彭帝超, 刘琳, 陈广宇, 陈海东, 左伍衡, 陈为. 一种新的视频摘要可视化算法. 计算机研究与发展, 2013, **50**(2): 371−378)

19 Rush A M, Chopra S, Weston J. A neural attention model for abstractive sentence summarization. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015. 379−389

20 Chopra S, Auli M, Rush A M. Abstractive sentence summarization with attentive recurrent neural networks. In: Proceedings of the 2016 NAACL-HLT. San Diego, California, USA: Association for Computational Linguistics, 2016. 93−98

21 Nallapati R, Zhou B W, Santos C N D, Gùlçehre C, Xiang B. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Berlin, Germany: Association for Computational Linguistics, 2016. 280−290

22 Paulus R, Xiong C M, Socher R. A deep reinforced model for abstractive summarization. arXiv preprint, arXiv: 1705.04304, 2017.

23 Ma S M, Sun X, Xu J J, Wang H F, Li W J, Su Q. Improving semantic relevance for sequence-to-sequence learning of Chinese social media text summarization. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017. 635−640

24 Tan J W, Wan X J, Xiao J G. Abstractive document summarization with a graph-based attentional neural model. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017. 1171−1181

25 Li P J, Lam W, Bing L D, Wang Z H. Deep recurrent generative decoder for abstractive text summarization. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017. 2091−2100

26 Chen Q, Zhu X D, Ling Z H, Wei S, Jiang H. Distraction-based neural networks for document summarization. arXiv preprint, arXiv: 1610.08462, 2016.

27 Nema P, Khapra M M, Laha A, Ravindran B. Diversity driven attention model for query-based abstractive summarization. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017. 1063−1072

28 Zhou Q Y, Yang N, Wei F R, Zhou M. Selective encoding for abstractive sentence summarization. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017. 1095−1104

29 See A, Liu P J, Manning D C. Get to the point: summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: Association for Computational Linguistics, 2017. 1073−1083

30 Hsieh Y L, Liu S H, Chen K Y, Wang H M, Hsu W L, Chen B. Exploiting sequence-to-sequence generation framework for automatic abstractive summarization. In: Proceedings of the 28th Conference on Computational Linguistics and Speech Processing. Tainan, China: ACLCLP, 2016. 115−128

31 Baetens J. Conversations on cognitive cultural studies: literature, language, and aesthetics. *Leonardo*, 2015, **48**(1): 93−94

32 Zhao Jing-Sheng, Zhu Qiao-Ming, Zhou Guo-Dong, Zhang Li. Review of research in automatic keyword extraction. *Journal of Software*, 2017, **28**(9): 2431−2449
(赵京胜, 朱巧明, 周国栋, 张丽. 自动关键词抽取研究综述. 软件学报, 2017, **28**(9): 2431−2449)

33 Mihalcea R, Tarau P. TextRank: bringing order into texts. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain: UNT Scholarly Works, 2004. 404−411

34 Lin C Y. Rouge: a package for automatic evaluation of summaries. Text summarization branches out. In: Proceedings of the ACL-04 Workshop. East Stroudsburg, USA: Association for Computational Linguistics, 2004. volume 8

35 Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv preprint, arXiv: 1412.6980, 2014.

36 Radev D, Allison T, Blair-Goldensohn S, Blitzer J, Çelebi A, Dimitrov S, et al. MEAD — a platform for multidocument multilingual text summarization. In: Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal: ELRA, 2004. 699−702

37 Zhang J M, Wang T M, Wan X J. PKUSUMSUM: a java platform for multilingual document summarization. In: Proceedings of the 26th International Conference on Computational Linguistics. Osaka, Japan: The COLING 2016 Organizing Committee, 2016. 287−291

38 Hou L W, Hu P, Bei C. Abstractive document summarization via neural model with joint attention. In: Proceedings of the 2018 Natural Language Processing and Chinese Computing, Lecture Notes in Computer Science, vol. 10619. Dalian, China: Springer, 2018. 329−338

侯丽微 华中师范大学计算机学院硕士研究生. 主要研究方向为自然语言处理.
E-mail: houliwei@mails.ccnu.edu.cn
(**HOU Li-Wei** Master student at the School of Computer Science, Central China Normal University. Her main research interest is natural language processing.)

胡 珀 华中师范大学计算机学院副教授. 主要研究方向为自然语言处理, 机器学习, 本文通信作者.
E-mail: phu@mail.ccnu.edu.cn
(**HU Po** Associate professor at the School of Computer Science, Central China Normal University. His research interest covers natural language processing and machine learning. Corresponding author of this paper.)

曹雯琳 华中师范大学计算机学院硕士研究生. 主要研究方向为自然语言处理.
E-mail: caowenlin@mails.ccnu.edu.cn
(**CAO Wen-Lin** Master student at the School of Computer Science, Central China Normal University. Her main research interest is natural language processing.)