基于子词单元的深度学习摘要生成方法

陈雪雯

（中国科学技术大学计算机科学与技术学院中国科大-伯明翰大学智能计算与应用联合研究所 安徽 合肥 230027）

摘 要 现有的生成式文本摘要方法存在一些局限，包括难以产生可靠的源文本表示，产生的摘要句与源文本的语义相似度较低，存在集外词问题等。对此提出一种混合神经网络编码器结构，对源文本的长距依赖和上下文信息进行捕捉，得到高质量的文本表示；提出一种基于关键短语的重排序机制，利用源文本中抽取的关键短语对集束搜索生成的候选序列进行重新排序，以减小其与源文本语义上的距离；对文本进行子词单元提取，利用更细粒度的单元对文本进行表示。该方法在不同长度的摘要数据集上进行实验，均取得了良好的效果。

关键词 生成式文本摘要 字节对编码 集束搜索 深度学习

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.03.034

ABSTRACT GENERATION METHOD OF DEEP LEARNING BASED ON SUBWORD UNITS

Chen Xuewen

(*USTC-Birmingham Joint Research Institute in Intelligent Computation and Its Application (UBRI)*, *School of Computer Science and Technology*, *University of Science and Technology of China*, *Hefei 230027*, *Anhui*, *China*)

Abstract There are some limitations in the existing generative text abstract methods, including the difficulty in generating reliable source text representation, the low semantic similarity between the generated abstract sentences and the source text, and the existence of the problem of out-of-vocabulary (OOV) words. A hybrid neural network encoder structure was proposed to capture the long-distance dependence and context information of the source text and obtain high-quality text representation. Then we proposed a reordering mechanism based on the key phrases extracted from the source text to reorder the candidate sequences generated by beam search, so as to reduce the semantic distance between them and the source text. And we extracted the subword units of the text and express the text with more fine-grained units. This method has been tested on different length summary datasets and achieves good results.

Keywords Generative text abstract Byte pair encoding Beam search Deep learning

0 引 言

当今互联网技术快速发展更迭，人类获取信息的途径变得更加丰富，世界各地每时每刻发生的事件都能通过计算机、各种手持终端以及其他传统媒体传送到我们面前，传统的时间和空间障碍被现代通信和传播技术所克服，全球的信息共享和交互已经实现，世界被更进一步地联结为一体。然而，与之俱来的问题不容忽视：人们正面临着前所未有的“信息爆炸”带来的挑战。如何更有效地利用信息资源，帮助人们快速处理信息和准确获取信息已经成为一个亟待解决的问题。

自动文摘是指从一个或者多个文档中，自动生成高度浓缩、通顺流畅并忠实于原文关键信息的摘要，从而帮助读者快速浏览和理解文档内容[1]。一个优秀的文摘生成系统，一方面要能够对输入的原始文本进行深入理解，捕获其核心语义信息并加以合理表示，另一方面需要具备根据原始文本表示，生成信息丰富、简明扼要、通顺流畅的摘要的能力。除了给用户提供简洁的文本，方便用户快速获取信息之外，文档摘要技术也可以用于问答系统中的答案后处理[2]等自然语言处理

收稿日期：2019 - 04 - 15。**陈雪雯**，硕士生，主研领域：自然语言理解。

任务中。

1　相关工作

自动文摘的研究始于 20 世纪 50 年代初期,自其首次被提出至今,在六十多年来的研究与发展中诞生了许多摘要方法,这些方法大致分为两类。第一种是抽取式方法[3],基本过程通常包括对原始文档中句子的重要程度进行评估并利用各种打分排序算法选出候选句,最后按照一定的组合策略连接形成摘要。另一种方法被称为生成式方法,最大的特点是可以使用原始文本中没有出现过的新词,这些词通常是对原始文本的改写。相较之下,抽取式方法生成的摘要通常是文档中一些重要句子的简单拼凑,不仅包含了大量的冗余信息,且句子与句子之间没有一定的关联性,从而导致了信息的碎片化和歧义性,而生成式方法则能够很好地克服这些缺点。

生成式摘要方法的发展得益于近年来深度学习研究的逐渐深入,尤其是机器翻译领域首创的序列到序列模型[4],其意义在于无须依赖人工先验知识,特征完全从数据中学习出来,并且可以得到更好的效果。受到神经网络机器翻译技术的启发,2015 年 Rush 等[5]将神经语言模型和基于上下文的输入编码器相结合,提出一种基于编码器-解码器框架的句子摘要模型 ABS,在给定输入句子的情况下,逐字生成摘要的每个词。该工作首次实现了将注意力机制应用到文本摘要任务中并提出利用 Gigaword 构建大量平行句对的方法,使得利用神经网络训练成为可能。次年,同组的 Chopra 等[6]在 Rush 等工作的基础上更进一步,使用一种有条件的卷积注意力模型作为编码器,将 ABS 模型中作为解码器的前向神经网络语言模型替换为循环神经网络,在 Gigaword 语料上和 DUC-2004 任务中取得了更好的效果。Nallapti 等[7]将自动文摘问题作为一个序列到序列问题,将传统方法中的特征显式地作为神经网络的输入,在两个数据集上均取得了超越 ABS 模型的结果。然而,上述模型多采用单一结构的编码器,其文本表示质量存在提升的空间[8]。这些系统生成的摘要可能存在与原始文本在语义上相似不高的问题[9]。

另外,为解决文本生成中普遍存在的集外词问题,文献[10]提出对词语进行字符级或子词级别的建模,目前这种细粒度文本表示主要应用于机器翻译任务中。

2　问题形式化描述

为了阐述方便,本文首先给出文本摘要问题的形式化描述。给定原始文本 $X=(x_1,x_2,\cdots,x_n)$,生成的摘要可以表示为 $Y=(y_1,y_2,\cdots,y_m)$,分别由其各自的单词序列组成,n 和 m 分别表示两者的长度,满足约束条件 $n>>m$。用 V 表示原始文本的词典,该词典由语料库中频度最高的 N 个单词组成。文本自动摘要问题的目标函数可以表示为:

$$\arg\max_{y\in V} P(Y\mid X)$$

$$P(Y\mid X)=\prod_{t=1}^{m}P(y_t\mid(y_1,y_2,\cdots,y_{t-1}),X,\Theta)$$

式中:Θ 表示训练过程中学习到的参数。上式表示摘要句中 t 时刻的单词 y_t 是基于原始文本 X 以及 t 时刻之前生成的所有单词$(y_1,y_2,\cdots,y_{t-1})$而产生的。

3　算法设计与实现

本文提出的自动文摘模型如图 1 所示。首先对数据集中的文本进行子词单元切分,得到该数据集的词表。给定一段处理过的原始文本作为输入,将该序列中的单词映射到连续的词向量空间得到其向量表示,并使用结合卷积神经网络和循环神经网络的混合神经网络编码器编码得到其隐层状态。在原始文本被编码完成后,利用结合注意力机制的解码器逐字生成摘要。在解码过程中,本模型引入基于关键短语的重排序模块对集束搜索生成的多个候选句进行处理,同时考虑其原始得分和所包含的关键短语的重要性得分,选择得分最高的候选句子作为最终生成的摘要。

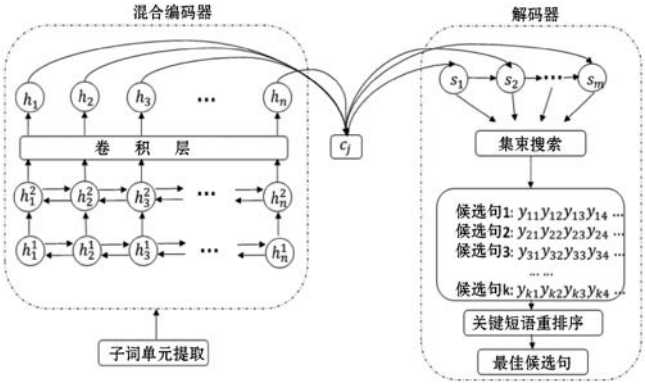

图 1　自动摘要模型结构

3.1　基于子词单元的文本表示

本文模型采用比词语级更细粒度的方法对原始文本中的词语进行表示,以更好地利用单词内部信息,在

一定程度上缓解文本摘要任务上的罕见词及集外词问题。

字节对编码算法迭代地使用一个未使用的字节将序列中出现次数最多的字节对进行替换。源端文本中最频繁的字符或字符序列被合并,在解码时,每个单词首先被分成字符序列,接着训练过程中学习得到的操作将字符合并成为更长的已知符号。该算法对词表进行子词提取的具体步骤描述如下:

① 对符号词表进行初始化,将单词拆分后的字符添加到符号词表中并对所有单词的词尾加入结束符</w>。该标识符用于解码时单词的扩展还原。

② 对词表中所有的符号进行迭代计数,获得词汇中所有的字符对,找出其中出现最频繁的连字序列,如('a','b'),用'ab'替换。

③ 每次的合并过程均会产生新的符号,该符号代表着单词中出现次数较多的子词,即 n-gram。

④ 合并过程最终产生的子词(或者完整的单词),将被加入到词表中。词表的大小则为初始大小与合并次数之和。

假设原始词表为{'h i g h e r </w>':2,'h o t t e s t </w>':5,'h i g h </w>':4,'b e s t </w>':3},其中的关键字是词表的单词分割成字符加上结尾符,键值则是该单词出现的频数,则整个过程模拟如图 2 所示。

```
原始词表:{'h i g h e r </w>':2, 'h o t t e s t </w>':5, 'h i g h </w>':4, 'b e s t </w>':3}
出现最频繁的序列:('s', 't'):8
合并该序列后的词表:{'h i g h e r </w>':2, 'h o t t e s t </w>':5, 'h i g h </w>':4, 'b e s t </w>':3}
出现最频繁的序列:('e', 'st'):8
合并该序列后的词表:{'h i g h e r </w>':2, 'h o t t e st </w>':5, 'h i g h </w>':4, 'b e st </w>':3}
出现最频繁的序列:('h', 'i'):6
合并该序列后的词表:{'h i g h e r </w>':2, 'h o t t e st </w>':5, 'h i g h </w>':4, 'b e st </w>':3}
出现最频繁的序列:('hi', 'g'):6
合并该序列后的词表:{'hig h e r </w>':2, 'h o t t e st </w>':5, 'hig h </w>':4, 'b e st </w>':3}
出现最频繁的序列:('hig', 'h'):6
合并该序列后的词表:{'high e r </w>':2, 'h o t t e st </w>':5, 'high </w>':4, 'b e st </w>':3}
```

图 2 BPE 算法在给定词表上的合并操作

至此,找出了该词表中出现频次最高的相邻字符,并将其合并得到子词单元,最终形成了更为合理的词表。通过子词单元的提取,将词语的含义与其形态分开,能够有效减少词表大小。经过子词单元处理的文本如图 3 所示,被切分成多个子词单元的词语中,前一个子词单元后会附加一个特殊的后缀"@@",这些文本被输入到编码器-解码器框架中进行端到端的处理,生成的摘要结果中也会包含该后缀,这有利于模型对这个单词进行恢复。

原始文本:The failure rate of a system usually depends on time, with the rate varying over the life cycle of the system.
子词处理后的文本:The fail@@ ure rate of a system usually depend@@ s on time, with the rate vary@@ ing over the life cycle of the system.

图 3 子词切分处理前后的文本实例

3.2 结合注意力机制的编码器-解码器结构

为了解决文档表示问题,本文提出了一种全新的编码器形式,该编码器结合了卷积神经网络和循环神经网络的结构特点,显式地利用卷积层来捕获目标词汇单元及其邻近单词之间的上下文关系,强化了语境的作用,是对传统的基于循环神经网络的编码器结构的一个很好的补充。基于此,本编码器既能学习到循环神经网络所擅长的序列信息和长距离依赖,同时能检测到局部时序无关的特征,从而得到高质量的原始文本的表示,作为摘要生成过程的基础。

如图 4 所示,本文采用双向 LSTM 作为编码器的基本计算单元,因为双向建模方法可以更完整地捕获序列中各单元之间的关联。具体来说,输入的原始文本序列$(x_1,x_2,\cdots,x_n)$中字符的词向量表示$e_x(x_t)$根据式(1)、式(2)被映射为前向隐层状态向量$(\overrightarrow{\boldsymbol{h}}_1,\overrightarrow{\boldsymbol{h}}_2,\cdots,\overrightarrow{\boldsymbol{h}}_n)$和后向隐层状态向量$(\overleftarrow{\boldsymbol{h}}_1,\overleftarrow{\boldsymbol{h}}_2,\cdots,\overleftarrow{\boldsymbol{h}}_n)$,每一时刻的前后向隐层状态按照式(3)进行拼接作为该时刻的整体隐层状态表示,其中ϕ表示循环神经网络中的激活函数。

$$\overrightarrow{\boldsymbol{h}}_t=\overrightarrow{\boldsymbol{\phi}}(e_x(x_t),\overrightarrow{\boldsymbol{h}}_{t-1}) \tag{1}$$

$$\overleftarrow{\boldsymbol{h}}_t=\overleftarrow{\boldsymbol{\phi}}(e_x(x_t),\overleftarrow{\boldsymbol{h}}_{t-1}) \tag{2}$$

$$\boldsymbol{h}_t=[\overrightarrow{\boldsymbol{h}}_t;\overleftarrow{\boldsymbol{h}}_t] \tag{3}$$

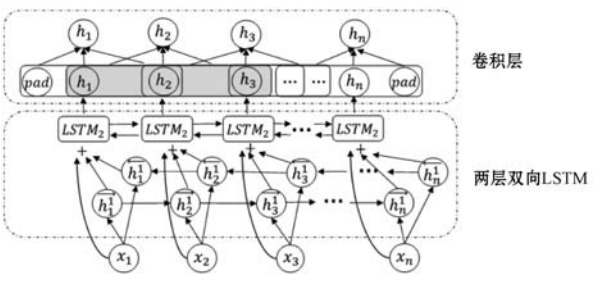

图 4 混合神经网络编码器结构

为了实现多层神经网络间的梯度传递,两层循环神经网络之间使用了残差连接。具体地,每一时刻底

层 LSTM 的输入被添加到其输出,其总和作为输入被馈送到第二层 LSTM。事实上,残差连接能够帮助构建更深的网络,缓解梯度消失等问题。由于硬件条件所限,本文模型选择了两层循环神经网络。在循环神经网络之上级联了一层无池化操作的卷积神经网络,其目的是利用卷积层对相邻状态之间的局部特征进行抽取,如下式所示,则第 i 个输入的隐层状态由其前后窗口内的相邻隐层状态共同决定。

$$h_i = \sigma(\theta \cdot h_{(i-(\omega-1)/2):(i+(\omega-1)/2)} + b) \tag{4}$$

式中:ω 表示表示卷积神经网络的滤波器窗口大小,我们将其设置为 3;σ 表示 Sigmoid 激活函数;b 则表示公式的偏置项。

编码完成后可以得到整个文本的上下文表示集合 $C = \{h_1, h_2, \cdots, h_n\}$,模型的解码器由一层单向 LSTM 组成,其中 j 时刻的隐层状态 s_j 由其前一时刻的隐层状态 s_{j-1} 和前一时刻的输出 y_{j-1} 共同更新,其计算式为:

$$s_j = \phi(e_y(y_{j-1}), s_{j-1}, c_j) \tag{5}$$

式中:$e_y(y_{j-1})$ 表示目标词语的词向量表示,上下文向量 c_j 可以根据注意力机制计算得到,具体表示为:

$$c_j = f(e_y(y_{j-1}), s_{j-1}, C) \tag{6}$$

注意力机制按照上下文表示集合 C 中的各个向量与已经生成的文本的相关性,赋予其一定的权值,每个 h_t 的权值可以根据下式计算得到:

$$\alpha_{t,j} = \frac{1}{Z} e^{g(e_y(y_{j-1}), s_j, h_t)} \tag{7}$$

$$Z = \sum_{k=1}^{T_x} e^{g(e_y(y_{j-1}), s_{j-1}, h_k)} \tag{8}$$

式中:$g(\cdot)$ 的作用是在给定 y_{j-1} 和 s_{j-1} 的情况下计算原始文本中隐层状态 h_t 的原始得分,而 Z 则用于对其进行归一化,计算方法如式(8)所示,在这里使用一个前馈神经网络对其进行模拟。该过程可以理解为利用注意力机制对原始文本中 t 时刻的输入词语及摘要中 j 时刻的目标词语关联程度概率的计算。

根据解码器端的隐层状态 s_j 及前一时刻生成的词语 y_{j-1} 和动态计算得到的上下文向量 c_j,模型可以对 j 时刻解码器循环神经网络的条件概率进行计算,表示为:

$$p(y_j \mid y_{<j}, X) \propto e^{g(e_y(y_{j-1}), s_j, c_j)} \tag{9}$$

3.3　基于关键短语的重排序机制

模型在预测阶段没有参考摘要的指导,解码器的工作过程是:接收编码器的 < EOS > (End of Sentence) 符号作为开始解码的信号,生成一个字符后将其作为下一时刻的输入,重复这个过程直到生成的句子达到设定长度或生成 < EOS > 符号则解码结束。预测的目标是根据当前模型选择概率最大的字符作为输出,然而得到这个最优解的复杂度一般非常高,为了减小搜索空间,本文采用集束搜索的策略进行近似求解。

集束搜索为了减少搜索范围降低问题复杂度,在每一步深度扩展的时候,仅保留 B 个最高得分的输出,而对质量较差的结点进行剪枝,最后从 B 个输出结果中选择得分最高的句子作为最终的输出。这里的 B 被称为集束宽度,而每一步扩展的评分函数则是当前时刻为止生成的各个单词的对数似然的总和,计算式如下:

$$score(Y_{t-1}, y_t \mid x) = \sum_{t=1} \log p(y_t \mid y_{<t}, x) \tag{10}$$

式中:x 表示原始文本中的字符,y_t 表示当前时刻生成的单词,Y_t 表示到 t 时刻为止扩展得到的候选句子序列,即 $Y_t = \{y_1 y_2 \cdots y_t\}$。

本文提出通过一种基于关键短语的重排序机制对集束搜索得到的候选序列进行选择最佳摘要句子。具体来说,这种机制根据序列的原始得分 $score(Y_{t-1}, y_t \mid x)$ 与其和原始文本中重合的关键词的重要性得分对生成的候选序列进行重新排序,选择其中得分最高的作为最后的摘要。直观上,关键短语中包含了可用于构建简明摘要的代表性实体,能够很好地捕获原始文本中的要点,因此,可以认为候选序列与原始文本中关键词的重叠越大,其包含的信息量越大,实验部分的定性分析也证实了这一假设。

总体而言,基于关键短语的重排序机制分为三个步骤。首先,使用基于图的无监督排名模型 TextRank 算法[11]从原始文本中提取出关键短语。具体地,原始文本的每个词汇单元被视为图的节点,而图中的连边则指示了预先设定好的窗口大小内的词汇单元之间的共现关系,连接节点 V_i 和 V_j 的边的初始权重 w_{ij} 被随机赋初始值,然后根据下式迭代计算节点 V_i 的重要性得分直到收敛。

$$s(V_i) = (1-d) + d \times \sum_{V_j \in adj(V_i)} \frac{w_{ij}}{\sum_{V_k \in adj(V_j)} w_{jk}} s(V_j) \tag{11}$$

式中:d 表示阻尼因子,通常被设置 0.85,$adj(V)$ 表示结点 V 的邻居节点。

其次,在图模型构建完毕后,将通过一个语法过滤器提取原始文本中的关键短语。本文设置语法过滤器为(JJ) * (NNP | NNPS | NNS | NN),其中 JJ 表示形容词或序数,NNP 和 NN 分别表示专有名词和普通名词,NNPS 和 NNS 是它们的复数形式。可以看到,该过滤

器主要由名词组成,这是因为名词短语在捕获文本主题的能力上较其他词性的短语有更强的优势,而其他词性的短语在神经网络端到端模型更易生成。这里设定符合条件的关键短语 kp 的得分是它所包含的词汇单位 V 的重要性得分的总和,可根据下式计算得到:

$$score(kp) = \frac{\sum_{V \in kp} s(V)}{len(kp) + 1} \tag{12}$$

最后,给定一个候选序列 Y_t,利用其原始分数及其对应的关键短语得分共同表示其与原始文本的相关性。另外,为了避免此得分函数对长句的偏好,将该得分除以候选序列的长度以达到归一化的目的。因此,最终得分 $score(Y_t)$ 由下式给出:

$$score(Y_t) = \frac{score(Y_{t-1}, y_t \mid x) \times \sum score(kp)}{len(Y_t) + 1} \tag{13}$$

经过关键短语重排序模块处理,将得到各候选序列的分数,选择得分最高的候选序列添加到生成的摘要句子中。

4 实 验

本节介绍了上文所提出文本摘要模型在包括句子摘要和标题生成在内的两个任务上的表现,并将其与多个当前最先进的系统在常用数据集上的摘要生成结果进行对比。本文使用 subword 表示子词单元的文本表示方法,与之对应,实现了一种词语级处理的传统方法 seq2seq(word level),subword-keyphrase 则表示本文完整的模型实现。这三种模型中的编码器均采用上文所提出的混合神经网络结构。

4.1 数据集及实验设置

首先,对本实验两个任务中所采用的数据集分别作出介绍,其详细信息如表 1 所示。

表 1 数据集详细信息

数据集	文章数目	平均字长	
		文章	摘要
CNN/Daily Mail	312 084	781	56
BBC	2 225	258	8
Inspec	2 000	121	12

CNN/Daily Mail 数据集[7]中包含了大量长篇新闻文章及其由多句话组成的摘要句,其训练集、验证集和测试集中的文章-摘要对的数量分别为 287 226、13 368、11 490。BBC[12]包含来自 BBC 新闻网站的 2 225 篇中等长度的文章及其摘要,将这些文档随机打乱并分为三个部分:训练集(1 100 篇),验证集(625 篇)和测试集(500 篇)。Inspec[13]是一个科学期刊摘要数据集,由 2 000 个简短文档组成,其标题被视为摘要,其中训练集、测试集、验证集的数目分别为 1 000、500、500。

本文使用深度学习框架 tensorflow 实现此模型。在编码器部分,选择两层双向 LSTM,其中每个门的隐藏单元设置为 200,解码器部分则是一层隐藏单元为 400 的单向 LSTM。词向量维度设置为 200,没有使用预先训练的词向量对其进行初始化,而是在训练过程中进行学习。关于优化器的选择,本模型使用了 Adam 优化器并采用其默认超参数设置:学习率 $\alpha = 0.01$,$\beta_1 = 0.9$,$\beta_2 = 0.999$,$\varepsilon = 1e-8$。其他参数在[-0.1, 0.1]的范围内进行了随机初始化。为了减轻解码器在训练阶段和生成阶段所输入数据的概率分布不一致的问题,以 0.2 的概率采样模型的输出并返回,以用作同一批次的第二次训练迭代的输入。此外,验证集上的损失被用于实现早停以防止过拟合。

4.2 实验结果与分析

4.2.1 句子摘要任务

本文采用 ROUGE 指标[15]评估所生成摘要的质量,其变体用于衡量参考摘要和生成摘要的 n-gram 同现情况,ROUGE-1 得分代表了自动摘要的信息量,ROUGE-2 得分评估了自动摘要的流畅性,而 ROUGE-L 可视为摘要对原文的涵盖率。表 2 展示了不同模型在 CNN/Daily Mail 数据集中上述指标的得分,其中基准模型的得分均来自其对应研究工作中的报告结果。

表 2 CNN/Daily Mail 数据集上的 ROUGE 分数

模型	R-1	R-2	R-L
ABS[5]	30.49	11.17	28.08
ABS +[5]	31.33	11.81	28.83
words-lvt2k-temp-att[7]	35.46	13.30	32.65
PGNet[14]	36.44	15.66	33.42
seq2seq(word level)	32.14	11.95	29.34
subword	37.61	16.54	34.57
subword-keyphrase	**38.98**	**17.46**	**36.63**

如表 2 所示,本文实现的 subword-keyphrase 模型在三个指标上均获得了最高分。与目前最优的 PGNet 模型相比,三个指标均有不同程度的提高(+2.54 ROUGE-1, +1.80 ROUGE-2, +3.21 ROUGE-L)。注意到 ROUGE-1 和 ROUGE-L 分数的增加比 ROUGE-2 更显著,这主要是因为关键短语重排序模块中语法过滤器模式的设置使得从 CNN/Daily Mail 数据集中提取出的关键短语鲜

有二元词组,而其中大量存在的 *n-gram* ($n \geqslant 3$)在很大程度上提升了最长公共子序列同现的性能。此外,通过比较 subword 和 seq2seq(word level)模型的得分,可以验证本文提出的子词处理方法的有效性。值得注意的是,seq2seq(word level)模型的指标与 ABS 模型相比亦有显著的提升(+1.65 ROUGE-1,+0.78 ROUGE-2,+1.26 ROUGE-L),这组对照实验中 ROUGE 得分的提升应归功于将 ABS 模型中的 RNN 替换为 seq2seq(word level)编码器的 LSTM 和 CNN 级联结构。可以推断,这种混合神经网络结构捕捉到了原始文本中更多的特征,得到了更好的原始文本表示,从而提升了摘要句的生成质量。

4.2.2　标题生成任务

与句子摘要任务不同,标题生成中的目标句一般不超过 20 字。本文在 BBC 和 Inspec 数据集上对模型进行了训练和评估。

实验结果如表 3 所示,展示了本文模型和基准模型在 ROUGE 指标上的得分比较。在 BBC 数据集上,我们观察到,本文模型的最佳表现 subword-keyphrase 在三个指标上分别超过基准模型 PGNet 得分 2.49、0.82 和 2.42;在 Inspec 数据集中,subword-keyphrase 模型亦表现出相同的优势,以平均高出 5 分的优势击败了 PGNet。另外,subword-keyphrase 与 subword 模型对比指标的提高验证了关键词重排序模块的有效性,subword 与 seq2seq 模型的对比则验证了子词单元提取处理对摘要句质量的提升,同样,seq2seq 与 ABS 间的编码器结构及表现差异则验证了混合神经编码器对模型性能的助益。

表 3　BBC 和 Inspec 数据集上的 ROUGE 分数

模型	BBC			Inspec		
	R-1	R-2	R-L	R-1	R-2	R-L
ABS[5]	12.02	0.64	11.32	18.68	2.97	14.53
ABS +[5]	12.76	0.76	13.04	19.41	3.53	16.21
PGNet[13]	14.33	0.71	13.47	21.13	3.07	17.65
seq2seq (word level)	13.60	0.73	13.16	19.36	3.14	16.27
subwod	16.51	1.03	14.72	24.74	4.23	20.12
subword-keyphrase	**16.82**	**1.53**	**15.89**	**26.74**	**7.44**	**22.34**

4.3　实例分析

图 5 展示了 subword-keyphrase 方法和基准模型 PGNet 在 Inspec 数据集上的生成的摘要样例,将它们与参考摘要进行比较。

原始文本:The spyware tool was only released by Microsoft in the last few weeks and has been downloaded by six million people. Stephen Toulouse, a security manager at Microsoft, said the malicious program was called Bankash-A Trojan and was being sent as an e-mail attachment. The program attempts to disable or delete Microsoft's anti-spyware tool and suppress warning messages given to users. Microsoft said in a statement it is investigating what it called a criminal attack on its software. Earlier this week, Microsoft said it would buy anti-virus software maker Sybari Software to improve its security in its Windows and e-mail software. Microsoft has said it plans to offer its own paid-for anti-virus software but it has not yet set a date for its release. The anti-spyware program being targeted is currently only in beta form and aims to help users find and remove spyware-programs which monitor internet use, causes advert pop-ups and slow a PC's performance.
关键词:Microsoft, spyware, investigate, program, software
参考摘要:Microsoft is investigating a trojan program to improve its security.
PGNet 模型摘要:Microsoft's anti-spyware help remove remove spyware.
subword-keyphrase.
模型摘要:Microsoft seeking spyware trojan program.

图 5　本模型和基准模型生成的摘要样例

由于 Inspec 文档长度是所使用的三个数据集中最短的,因此我们将其关键短语的过滤模式放松到重要的单词,从原始文本中提取的关键短语数上限被设置为 5,这就对生成的摘要与原始文本中关键短语匹配提出了更高的要求。但从实验结果来看,本文模型很好地命中了从原始文本中提取的多数关键短语,且表现出了一定的改写能力,将原始文本中的“investigating”转化成“seeking”,这种改写能力正是生成式摘要方法的重要特征。另一方面,尽管 PGNet 模型生成的摘要中包含了与原始文本出现过的短语“Microsoft”“spyware”等,但整体语义上与原始文本的主旨相去甚远。PGNet 摘要中有一些重复的单词,破坏了整个句子的连贯性和可读性。相比之下,本模型摘要由于关键短语的指导,表现出了良好的可读性。

5　结　语

本文提出了一种基于子词单元的生成式摘要方法。针对摘要生成中的罕见词和集外词问题,提出一种基于字节对编码算法的子词处理方法对原始文本中的词语进行表示,利用单词内部信息的同时有效减小词表规模。针对语句表示学习问题,本文设计了一种全新的基于深度混合神经网络的编码器结构,以提高文本表示质量,作为解码的基础。此外,本模型在搜索算法中集成了一个基于关键短语的重排序模块,能够指导摘要句的选择,有助于提高生成的摘要与原始文本之间的语义相关性。实验结果表明本文方法在不同

文档长度的数据集上的表现都优于目前最先进的文摘系统。

参 考 文 献

[1] 王萌,唐新来,何婷婷. 一种文本分割技术的多文档文摘方法研究[J]. 计算机应用与软件,2014,31(9):40 - 44.

[2] 胡迁,黄青松,刘利军,等. 基于自动文摘的答案生成方法研究[J]. 计算机应用与软件,2018,35(12):187 - 192,307.

[3] 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.

[4] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014: 3104 - 3112.

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

[6] 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.

[7] Nallapati R, Zhou B, Dos Santos C N, et al. Abstractive text summarization using sequence-to-sequence RNNs and beyond[C]//Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. 2016:280 - 290.

[8] Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[C]//29th AAAI Conference on Artificial Intelligence. 2015.

[9] Ma S, Sun X. A semantic relevance based neural network for text summarization and text simplification[EB]. arXiv preprint arXiv:1710.02318, 2017.

[10] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units[EB]. arXiv preprint arXiv:1508.07909, 2015.

[11] Mihalcea R, Tarau P. Textrank: Bringing order into text[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004.

[12] Greene D, Cunningham P. Practical solutions to the problem of diagonal dominance in kernel document clustering[C]// Proceedings of the 23rd International Conference on Machine Learning. Association for Computing Machinery, 2006: 377 - 384.

[13] Hulth A. Improved automatic keyword extraction given more linguistic knowledge[C]//Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2003: 216 - 223.

[14] See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1073 - 1086.

[15] Lin C Y. ROUGE: A package for automatic evaluation of summaries[C]//Proceedings of the Workshop on Text Summarization Branches Out(WAS 2004). 2004.

(上接第 197 页)

[46] Premaratne P, Yang S, Vial P, et al. Centroid tracking based dynamic hand gesture recognition using discrete Hidden Markov Models[J]. Neurocomputing, 2017, 228: 79 - 83.

[47] Wang X, Xia M, Cai H, et al. Hidden-Markov-Models-based dynamic hand gesture recognition[J]. Mathematical Problems in Engineering, 2012, 2012:986134.

[48] Neverova N, Wolf C, Taylor G W, et al. Hand segmentation with structured convolutional learning[C]//Asian Conference on Computer Vision. Springer, Cham, 2014:687 - 702.

[49] Oyedotun O K, Khashman A. Deep learning in vision-based static hand gesture recognition[J]. Neural Computing and Applications, 2017, 28(12): 3941 - 3951.

[50] Molchanov P, Gupta S, Kim K, et al. Hand gesture recognition with 3D convolutional neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2015: 1 - 7.

[51] Huang J, Zhou W, Li H, et al. Sign language recognition using 3d convolutional neural networks[C]//2015 IEEE international conference on multimedia and expo(ICME). IEEE, 2015: 1 - 6.

[52] Molchanov P, Yang X, Gupta S, et al. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4207 - 4215.

[53] Nunez J C, Cabido R, Pantrigo J J, et al. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition[J]. Pattern Recognition, 2018, 76: 80 - 94.

[54] O'Hara K, Gonzalez G, Sellen A, et al. Touchless interaction in surgery[J]. Communications of the ACM, 2014, 57(1): 70 - 77.

[55] Strickland M, Tremaine J, Brigley G, et al. Using a depth-sensing infrared camera system to access and manipulate medical imaging from within the sterile operating field[J]. Canadian Journal of Surgery, 2013, 56(3): E1 - E6.

[56] Fischinger D, Einramhof P, Papoutsakis K, et al. Hobbit, a care robot supporting independent living at home: First prototype and lessons learned[J]. Robotics and Autonomous Systems, 2016, 75: 60 - 78.