



融合文本分类的多任务学习摘要模型

周伟泉, 蓝雯飞

(中南民族大学 计算机科学学院, 武汉 430074)

摘要: 文本摘要应包含源文本中所有重要信息, 传统基于编码器-解码器架构的摘要模型生成的摘要准确性较低。根据文本分类和文本摘要的相关性, 提出一种多任务学习摘要模型。从文本分类辅助任务中学习抽象信息改善摘要生成质量, 使用K-means聚类算法构建Cluster-2、Cluster-10和Cluster-20文本分类数据集训练分类器, 并研究不同分类数据集参与训练对摘要模型的性能影响, 同时利用基于统计分布的判别法全面评价摘要准确性。在CNNDM测试集上的实验结果表明, 该模型在ROUGE-1、ROUGE-2和ROUGE-L指标上相比强基线模型分别提高了0.23、0.17和0.31个百分点, 生成摘要的准确性更高。

关键词: 编码器-解码器架构; 文本摘要; 文本分类; 多任务学习; 聚类算法; 统计分布

开放科学(资源服务)标志码(OSID):



中文引用格式: 周伟泉, 蓝雯飞. 融合文本分类的多任务学习摘要模型[J]. 计算机工程, 2021, 47(4): 48-55.

英文引用格式: ZHOU Weixiao, LAN Wenfei. Summarization model using multi-task learning fused with text classification[J]. Computer Engineering, 2021, 47(4): 48-55.

Summarization Model Using Multi-Task Learning Fused with Text Classification

ZHOU Weixiao, LAN Wenfei

(School of Computer Science, South-Central University for Nationalities, Wuhan 430074, China)

[Abstract] The text summary should include all the important information in the source text, but the summaries generated by traditional summarization models based on encoder-decoder architecture are not accurate. Based on the correlation between text classification and text summarization, this paper proposes a summarization model using Multi-Task Learning (MTL). The model learns abstract information from the auxiliary tasks of text classification to improve the quality of generated summaries. The K-means clustering algorithm is used to construct text classification datasets Cluster-2, Cluster-10 and Cluster-20 to train the classifier. On this basis, the impact of different classification datasets participating in the training on the performance of the summarization model is studied, and a discriminant method based on statistical distribution is proposed to reflect the accuracy of the summary. Experimental results on the CNNDM test set show that the proposed model improves the ROUGE-1, ROUGE-2 and ROUGE-L indexes by 0.23, 0.17 and 0.31 percentage points compared with the strong baseline model, which demonstrates the summaries generated by this model are more accurate.

[Key words] encoder-decoder architecture; text summarization; text classification; Multi-Task Learning (MTL); clustering algorithm; statistical distribution

DOI: 10.19678/j.issn.1000-3428.0057448

0 概述

文本摘要是自然语言处理(Natural Language Processing, NLP)的重要分支^[1], 其将源文本压缩成短文本且包含了源文本中的主要信息。抽取式摘要方法^[2]通过从源文本中抽取句子组成摘要, 具体为对句子重要性打分并按分数排序选取句子^[3]。生成式摘要方法^[4]

通过重新组织源文本的主要内容形成摘要, 摘要过程与人工撰写类似。现有生成式摘要模型本质上都是基于编码器-解码器架构的单任务模型, 文献[5]指出此类模型虽然能够生成流畅的摘要, 但在准确性方面表现较差, 甚至可能与源文本的中心思想相悖。

多任务学习(Multi-Task Learning, MTL)共享相关任务之间的表征, 能使摘要模型更好地概括文本主要

基金项目: 国家自然科学基金(61772562)。

作者简介: 周伟泉(1997—), 男, 硕士研究生, 主研方向为自然语言处理、文本摘要、机器翻译; 蓝雯飞(通信作者), 教授、博士。

收稿日期: 2020-02-21 修回日期: 2020-04-11 E-mail: lanwenfei1@163.com

内容。文献[6]提出一对多、多对一、多对多三种多任务学习方法。文献[7]将情感分类定义为特殊类型的摘要任务,将文本以标签方式进行概括。受文献[6]研究工作的启发,本文根据文本分类和文本摘要的相关性,提出一种多任务学习摘要模型。通过一对多的形式关联主要任务和辅助任务,借助文本分类改善摘要模型的生成质量,使用K-means聚类算法构建文本分类数据集,并利用基于统计分布的判别法全面评价摘要准确性。

1 相关工作

近年来,深度学习技术在文本摘要领域得到广泛应用。文献[8]提出基于注意力机制的端到端模型,并将其应用于文本摘要任务。文献[9]将注意力机制应用于自然语言处理领域。文献[10]提出指针网络,将解码器中固定规模的词汇表扩展至可变规模的词汇表。文献[11]提出 CopyNet模型,并在文本摘要任务上验证了 CopyNet模型相较传统模型具有显著优势。摘要任务与翻译任务不同,其数据集中参考摘要的长度远短于源文本,导致传统模型和 CopyNet模型无法很好地对源端文本以及参考摘要进行对齐。文献[12]指出对齐问题导致解码器容易重复生成冗余的单词或句子,针对此问题,该文献提出在解码器中引入覆盖度机制协助调整未来时间步的注意力。文献[13]提出 Pointer-Generator(Coverage),缓解了集外词(Out of Vocabulary, OOV)和重复生成问题。

随着深度学习技术的发展,研究人员通过多任务学习提高模型生成摘要的准确性。文献[14]将问题生成作为辅助任务,提出基于多层编码器-解码器模型的多任务学习架构。文献[15]指出翻译任务与摘要任务具有较强的相关性,使用机器翻译可提高摘要模型的生成性能。文献[16]提出基于多任务学习的深层神经网络框架,通过同时学习人的注视行为以及文档中的词性和句法属性来预测输入文档中表达的整体情绪。与文献[14-16]研究工作不同,本文将文本分类作为辅助任务,使得摘要模型能够学习到更抽象的信息。文献[17]在文献[13]的基础上定义了一个内容选择器来确定源文本中哪些短语一定包含在摘要模型中,该模型是关键词识别与文本摘要的混合模型。文献[18]将多模态注意力机制引入摘要模型中,输入源文本、参考摘要和参考图片进行训练。多模态注意力机制同时关注摘要和图片中的信息以生成质量更高的摘要,该模型是文本摘要与图片识别的混合模型。此外,文献[19-21]从不同角度研究多模态摘要模型。

针对编码器-解码器架构的改进,文献[22]将单独的编码器划分为多个协作编码器,使用深度通信代理表示这些协作编码器,并在不同代理间实现数据共享,摘要模型通过提高编码质量间接提高生成摘要的质量。文献[23]提出 Transformer模型,其相

比循环神经网络(Recurrent Neural Network, RNN)序列具有更优的并行化处理能力和特征提取能力,逐渐被应用于文本摘要领域。

2 多任务学习摘要模型

2.1 总体框架

多任务学习将相关任务放入同一框架中进行训练,一般分为一个主要任务和若干个辅助任务。文本分类是对不同文本中的重要信息进行区分,文本摘要是从文本中识别出重要的信息并进行提取,类别标签是更抽象的摘要表示。多任务学习摘要模型的总体框架如图1所示,由共享编码器(Shared Encoder)、分类器(Classifier)和摘要解码器(Summarization Decoder)构成。

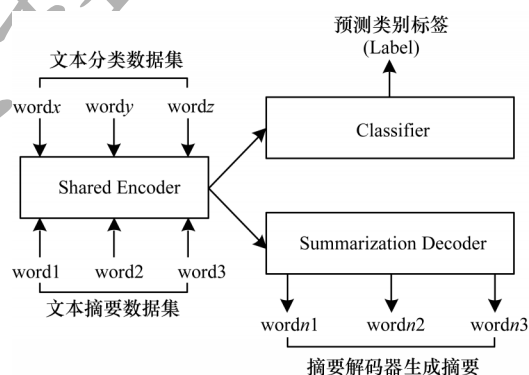


图1 多任务学习摘要模型的总体框架

Fig.1 The overall framework of summarization model with multi-task learning

共享编码器与分类器构成分类模型,与摘要解码器构成摘要模型。编码器采用硬共享机制,来自两个任务的梯度信息直接通过共享参数传递,强制所有任务使用公共空间表示。在训练编码期间,摘要任务与分类任务交替运行,随机输入相应任务数据集中的源文本。在训练解码期间,分类器或摘要解码器进行标签类别预测或摘要生成,相对于真实标签或参考摘要计算损失,反向传播并更新模型参数。在测试编码期间,编码器接收摘要测试集作为输入。在测试解码期间,摘要解码器使用集束搜索预测下一时刻的单词输出并选择概率最高的单词序列作为生成的摘要。

2.2 共享编码器

RNN通常用来处理时间序列数据,能够很好地提取文本单元之间的前后关联信息。RNN变体包括长短时记忆(Long Short-Term Memory, LSTM)网络^[24]和门控循环单元(Gated Recurrent Unit, GRU)^[25]。LSTM在RNN的基础上引入了遗忘门、输入门和输出门,前向传播公式如下:

$$h_t = c_t = \text{LSTM}(x_t, h_{t-1}, c_{t-1}) \quad (1)$$

其中, x_t 为 LSTM 的输入, h_{t-1} 、 c_{t-1} 分别为上一时刻的隐藏状态和细胞状态, h_t 、 c_t 分别为当前时刻的隐藏状态和细胞状态。

GRU将LSTM的3个门简化为更新门和重置门。在逻辑架构中,GRU没有细胞状态 c ,直接将隐藏状态 h 传递给下一个单元,前向传播公式如下:

$$h_t = \text{GRU}(x_t, h_{t-1}) \quad (2)$$

本文使用双向LSTM(Bi-directional LSTM, Bi-LSTM)和双向GRU(Bi-directional GRU, Bi-GRU)作为多任务学习摘要模型的编码器以更好地捕捉双向语义依赖关系。给定源文本 $D = \{w_1, w_2, \dots, w_n\}$,每个单词 w_i 被嵌入 K 维向量, Bi-LSTM或Bi-GRU对其双向编码,输出两个隐藏状态序列 $(\vec{h}_1, \vec{h}_2, \dots, \vec{h}_n)$ 和 $(\overleftarrow{h}_1, \overleftarrow{h}_2, \dots, \overleftarrow{h}_n)$ 。在具体实现中,源文本通过数据块的方式进入编码器,在变长序列中进行补零操作来处理长度变化。

2.3 摘要解码器

2.3.1 Pointer-Generator解码器

摘要解码器与文献[13]提出的Pointer-Generator解码器类似,其生成的单词可以来源于源文本或指定的词汇表。摘要解码器架构如图2所示。

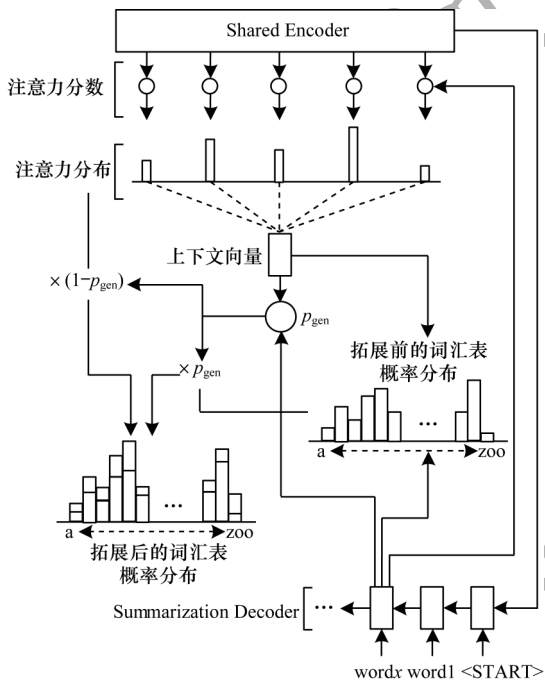


图2 摘要解码器架构

Fig.2 Summarization decoder architecture

Pointer-Generator解码器接收共享编码器输出的隐藏状态序列 $(\vec{h}_1, \vec{h}_2, \dots, \vec{h}_n)$ 和 $(\overleftarrow{h}_1, \overleftarrow{h}_2, \dots, \overleftarrow{h}_n)$,并将源文本表示为两者的级联:

$$h_i = [\vec{h}_i; \overleftarrow{h}_i] \quad (3)$$

在每一时刻 t ,单层LSTM或GRU接收单词的嵌入达到新的解码器状态 s_t ,通过注意力机制计算注意力分数 e'_t 、注意力分布 a'_t 和上下文向量 c_t 。计算公式如下:

$$e'_t = v^T \tanh(W_h h_t + W_s s_t + b_a) \quad (4)$$

$$a'_t = \text{softmax}(e'_t) \quad (5)$$

$$c_t = \sum_i a'_t h_i \quad (6)$$

其中: v, W_h, W_s, b_a 为可优化的参数; \tanh 为激活函数; a'_t 为源文本单词上的概率分布,指导解码器重点关注某些单词。

传统基于注意力机制的编码器-解码器模型从当前解码器状态 s_t 和上下文向量 c_t 中生成词汇表概率分布 P_{vocab} ,而Pointer-Generator解码器定义 p_{gen} ,由当前解码器状态 s_t 、当前时刻输入单词嵌入向量 e_t 和上下文向量 c_t 共同决定,计算公式如下:

$$p_{\text{gen}} = \sigma(w_c^T c_t + w_s^T s_t + w_e^T e_t + b_{p_{\text{gen}}}) \quad (7)$$

其中: $w_c, w_s, w_e, b_{p_{\text{gen}}}$ 表示可优化的参数; σ 表示sigmoid函数; p_{gen} 表示从词汇表中生成单词的概率, $(1-p_{\text{gen}})$ 表示通过从注意力分布 a'_t 中抽取源文本单词的概率。拓展后的词汇表概率分布计算公式如下:

$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1-p_{\text{gen}}) \sum_{i: w_i=w} a'_i \quad (8)$$

其中, w 表示某个单词。与Pointer-Generator相比,传统编码器-解码器模型被限制在容量有限的词汇表中。

2.3.2 覆盖度机制

本文在摘要解码器中引入覆盖度机制^[13]缓解重复生成的问题。覆盖度向量 cov' 表示时刻 t 前所有解码的注意力分布 a^i 的总和,计算公式如下:

$$\text{cov}' = \sum_{i=0}^{t-1} a^i \quad (9)$$

其中, cov' 可解释为源文本单词上的非规范化分布,表示在当前时刻这些单词从注意力机制中获得的覆盖程度。若将 cov' 作为注意力向量的额外输入,则式(4)改写为:

$$e'_t = v^T \tanh(W_h h_t + W_s s_t + W_c \text{cov}'_t + b_a) \quad (10)$$

其中, v, W_h, W_s, W_c, b_a 是可优化的参数。覆盖度机制通过总结已有决策避免重复注意相同的位置。摘要解码器在时刻 t 的损失被定义为复合损失函数,计算公式如下:

$$L_t = -\ln P(w_t^*) + \lambda \sum_i \min(a'_i, c'_i) \quad (11)$$

其中, w_t^* 为解码器在时刻 t 接收的参考摘要的单词, $-\ln P(w_t^*)$ 为Pointer-Generator解码器在时刻 t 的损失, $\sum_i \min(a'_i, c'_i)$ 为覆盖度机制在时刻 t 带来的损失, λ 为超参数。

2.4 分类器

文本分类作为辅助任务帮助摘要模型学习到更加抽象的信息,提高解码器生成摘要的准确性。分类器架构如图3所示。

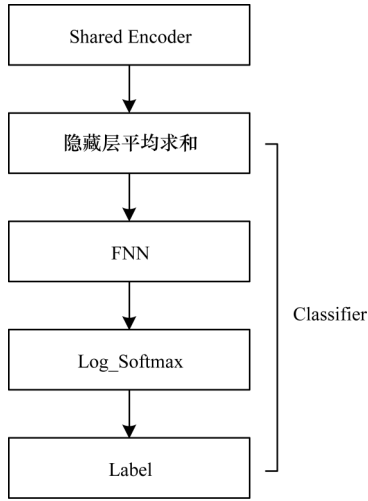


图3 分类器架构

Fig.3 Classifier architecture

共享编码器传递隐藏状态序列 $(\vec{h}_1, \vec{h}_2, \dots, \vec{h}_n)$ 和 $(\vec{h}_1, \vec{h}_2, \dots, \vec{h}_n)$ 进入隐藏层平均求和模块, 计算公式如下:

$$h_f = \frac{\sum_{i=1}^n \vec{h}_i}{n} \quad (12)$$

$$h_b = \frac{\sum_{i=1}^n \vec{h}_i}{n} \quad (13)$$

$$h_s = h_f + h_b \quad (14)$$

其中, h_f 和 h_b 分别表示正向、反向编码的文本表示, h_s 表示最终文本表示。

前馈神经网络 (Feedforward Neural Network, FNN) 将 h_s 映射为与文本类别数量相同的维度。使用 Log_Softmax 函数计算类别标签的概率分布, 计算公式如下:

$$\text{Log_Softmax}(x_i) = \ln \frac{\exp(x_i)}{\sum_j \exp(x_j)} \quad (15)$$

其中, \exp 表示以 e 为底的指数函数。本文结合 Log_Softmax 函数与负对数似然损失计算预测类别与真实类别的损失, 该损失等价于预测类别与真实类别的交叉熵。

3 数据集构建

3.1 文本摘要数据集

CNNDM (CNN/Daily Mail) [26] 是摘要领域的基准数据集之一, 源文本和参考摘要分别来源于在线新闻文章和人工撰写。本文使用 CNNDM 的匿名版本, 其中包含 286 896 组训练集、11 489 组测试集和 13 368 组验证集。

3.2 文本分类数据集

在多任务学习中, 多个数据集原始特征如果有一定相似性, 则可以提高摘要任务的性能。本文通过无监督算法获取 CNNDM 源文本的类别标签, 文本分类数据集的构建流程如图 4 所示, 具体步骤如下:

- 1) 剔除 CNNDM 训练集中的参考摘要, 保留源文本。
- 2) 对抽取出的源文本进行分词、去停用词和去低频词, 避免停用词和低频词对有效信息造成的噪声干扰。
- 3) 使用 TF-IDF 特征提取方法将预处理后的源文本向量化, TF-IDF 特征提取方法的主要思想为评估某个词相对于数据集中某份文件的重要程度。
- 4) 应用 K-means 聚类算法对向量化后的文本进行聚类操作。
- 5) 得到类别数量分别为 2、10、20 的 3 个文本分类数据集, 将其分别称作 Cluster-2、Cluster-10、Cluster-20, 为研究不同类别数量的文本分类数据集参与训练对模型生成摘要准确性的影响提供数据集支撑。

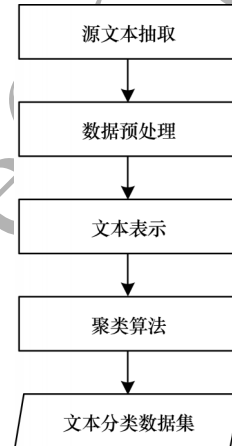


图4 文本分类数据集的构建流程

Fig.4 Construction process of text classification datasets

4 评价指标与判别法

4.1 ROUGE 评价指标

ROUGE [27] 是文本摘要领域的基准评价指标, 基于摘要中 n 元词 (n -gram) 的共现信息来评价摘要的准确性。ROUGE-N 和 ROUGE-L 计算公式如下:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{RS\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \{RS\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \quad (16)$$

其中, n 表示 n -gram 的长度, $\{RS\}$ 表示参考摘要, $\text{Count}_{\text{match}}(\text{gram}_n)$ 表示参考摘要与待测摘要中相同的 n -gram 个数, $\text{Count}(\text{gram}_n)$ 表示参考摘要中出现的 n -gram 个数。

ROUGE-L 中的 L 即为最长公共子序列 (Longest Common Subsequence, LCS), 计算公式如下:

$$R_{LCS} = \frac{LCS(X, Y)}{m} \quad (17)$$

其中, $LCS(X, Y)$ 表示参考摘要与待测摘要最长公共子序列的长度, m 表示参考摘要的长度。

4.2 基于统计分布的判别法

ROUGE 对比参考摘要与待测摘要来判定摘要准确性, 但是忽略了摘要句子在源文本中出现的位置信息。本文提出一种基于统计分布的判别法, 从总体分布的角度判断待测摘要的准确性。基于统计分布的判别流程如图 5 所示。

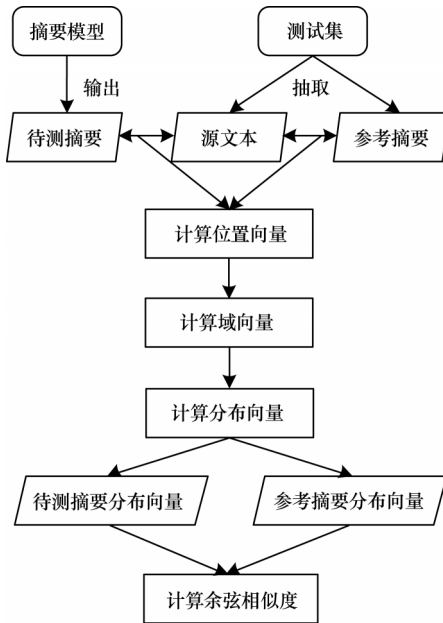


图 5 基于统计分布的判别流程

Fig.5 Discriminant process based on statistical distribution

本文将待测摘要表示为 $S=\{s_1, s_2, \dots, s_n\}$, $s_i=\{e_1, e_2, \dots, e_m\}$, 参考摘要表示为 $G=\{g_1, g_2, \dots, g_n\}$, $g_i=\{t_1, t_2, \dots, t_p\}$, 源文本表示为 $D=\{d_1, d_2, \dots, d_n\}$, $d_i=\{c_1, c_2, \dots, c_q\}$, 其中, s, g, d 分别表示 S, G, D 中的文本, e, t, c 分别表示 s, g, d 中的句子, n 表示测试集数据量, m, p, q 分别表示 s, g, d 的句子数量。计算 s_1 与 d_1 的位置向量 v_{sdp}^1 的具体步骤如下:

1) 定义雅卡尔相似系数, 该系数通常用来判断句子之间是否冗余, 计算公式如下:

$$\text{Jaccard}(r, o) = \max\left(\frac{\text{len}(r \cap o)}{\text{len}(r)}, \frac{\text{len}(r \cap o)}{\text{len}(o)}\right) \quad (18)$$

其中, r, o 分别表示 e_i, c_j 中非冗余单词的集合, \cap 表示交集, len 函数计算集合的大小。

2) 雅卡尔相似系数越高, c_j 替代 e_i 的能力越强, 寻求 e_i 对应最高雅卡尔相似系数的 c_j 。

3) 定义与 d_1 句子数量相等长度的全零位置向量 v_{sdp}^1 , 在 v_{sdp}^1 中将可替代 e_i 的 c_j 在 d_1 中的位置置为 1, 若位置重复则累加。

若 d_1 为 “zhou wei xiao。 yu bo wen。 xu kang。 ni hao。 zhao qiang。 keyi。” 且 s_1 为 “zhou wei。 qiang。 wei xiao。 kang。” 则 $v_{sdp}^1=[2, 0, 1, 0, 1, 0]$ 。计算域向量 v_{sdt}^1 的具体步骤如下: 1) 划分 v_{sdp}^1 为均等的 X 个域; 2) 各个域中元素求和。若 $v_{sdp}^1=[2, 0, 1, 0, 1, 0]$ 且 $X=3$, 则 $v_{sdt}^1=[2, 1, 1]$ 。划分相同数量的域是为了计算分布向量以及同时引入相对位置和绝对位置信息。本文将 S 与 D 的域向量集合表示为 $H_{sdt}=\{v_{sdt}^1, v_{sdt}^2, \dots, v_{sdt}^n\}$, G 与 D 的域向量集合表示为 $H_{gdt}=\{v_{gdt}^1, v_{gdt}^2, \dots, v_{gdt}^n\}$ 。在相同数量域的基础上分别对 H_{sdt}, H_{gdt} 的元素累加得到待测摘要分布向量 v_{sdd} 和参考摘要分布向量 v_{gdd} 。 v_{sdd} 与 v_{gdd} 之间的余弦相似度为待测摘要和参考摘要的总体分布在向量空间中的相似度。余弦相似度越高, 待测摘要越准确。

5 实验与结果分析

5.1 实验设置

本文使用深度学习框架 PyTorch 实现 6 种单任务基线模型和 6 种多任务学习摘要模型, 各个摘要模型基本架构及配置如表 1 所示, 其中, Attention、Coverage 分别表示注意力机制、覆盖度机制, Cluster 表示使用文本分类数据集, Pointer-Generator 默认使用注意力机制和指针网络。

表 1 摘要模型基本架构及配置

Table 1 Basic architecture and configuration of summarization models

类型	序号	基本架构	配置
单任务 基线模型	1	Encoder-Decoder	LSTM, Attention
	2	Encoder-Decoder	GRU, Attention
	3	Pointer-Generator	LSTM
	4	Pointer-Generator	GRU
	5	Pointer-Generator	LSTM, Coverage
	6	Pointer-Generator	GRU, Coverage
多任务学习 摘要模型	1	Multi-task	LSTM, Cluster-2
	2	Multi-task	GRU, Cluster-2
	3	Multi-task	LSTM, Cluster-10
	4	Multi-task	GRU, Cluster-10
	5	Multi-task	LSTM, Cluster-20
	6	Multi-task	GRU, Cluster-20

在训练和测试期间, 截断输入文本至 400 个单词以内, 限制生成摘要长度至 120 个单词以内。所有模型的隐藏状态维度均设置为 256 维, 单词嵌入向量维度设置为 128 维。本文没有使用预训练词向量, 所有模型词汇表大小设置为 50 000, 所有模型的摘要任务采用 Adagrad 优化器^[28], 初始化学习率设置为 0.15, 累加器的起始值设置为 0.1, 梯度剪裁阈值设置为 2。文

本分类任务使用 Adam^[29] 优化器,初始学习率设置为 0.001。所有模型均在单个 GeForce GTX TITAN X 12 GB 显存 GPU 上训练。在训练期间,设置数据块大小为 16,防止多任务学习导致 GPU 显存波动引起显存溢出。在测试期间,设置数据块大小为 100,所有模型的摘要解码器的集束搜索尺寸设置为 4。

5.2 结果分析

本文使用 files2rouge 包测评所有模型生成摘要的 ROUGE-1、ROUGE-2、ROUGE-L 在 95% 置信区间的标准分数,测试文本为 CNNDM 测试集。6 种单任务基线模型的 ROUGE 标准分数如表 2 所示,结果显示 Pointer-Generator 架构的性能优于传统 Encoder-Decoder 架构,特征提取器(LSTM 和 GRU)的选择对基线模型生成摘要准确性的影响较小。

表 2 单任务基线模型性能对比
Table 2 Performance comparison of single task baseline models

序号	基本架构	配置	ROUGE-	ROUGE-	ROUGE-
			1	2	L
1	Encoder-Decoder	LSTM,Attention	32.11	12.52	29.07
2	Encoder-Decoder	GRU,Attention	32.19	12.63	29.16
3	Pointer-Generator	LSTM	36.76	15.84	33.25
4	Pointer-Generator	GRU	36.71	15.85	33.26
5	Pointer-Generator	LSTM,Coverage	39.17	17.30	35.71
6	Pointer-Generator	GRU,Coverage	39.12	17.34	35.64

多任务学习摘要模型的 ROUGE 标准分数如表 3 所示,结果显示:在 ROUGE-1、ROUGE-L 指标上,6 种多任务学习摘要模型均能生成更准确的摘

要,Multi-task(GRU,Cluster-10)相较于强基线模型 Pointer-Generator(GRU,Coverage)分别在 3 个 ROUGE 指标上提高了 0.23、0.17 和 0.31 个百分点;在 ROUGE-2 指标上,Cluster-10 参与训练后的模型优于性能最好的单任务基线模型。使用 GRU 作为特征提取器的模型在部分指标上的提升略高于 LSTM,这可能是数据集较大所致。文本分类数据集的选取对模型的生成性能影响较大。在 ROUGE 指标上,使用 Cluster-10 训练后的模型均生成了最准确的摘要。本文认为出现该结果的主要原因为:过大的类别数量影响文本分类的效果从而影响生成摘要的质量,过小的类别数量使得引入的抽象信息不够从而干扰摘要模型。

表 3 多任务学习摘要模型性能对比

Table 3 Performance comparison of summarization models with multi-task learning

序号	基本架构	配置	ROUGE-	ROUGE-	ROUGE-
			1	2	L
1	Multi-task	LSTM,Cluster-2	39.21	17.24	35.76
2	Multi-task	GRU,Cluster-2	39.26	17.23	35.73
3	Multi-task	LSTM,Cluster-10	39.34	17.45	35.97
4	Multi-task	GRU,Cluster-10	39.35	17.51	35.95
5	Multi-task	LSTM,Cluster-20	39.25	17.21	35.81
6	Multi-task	GRU,Cluster-20	39.29	17.39	35.83

本文将测试集参考摘要的分布向量作为基准,统计并对比 3 种摘要模型生成待测摘要的分布向量,同时计算不同分布向量间的余弦相似度,具体结果如表 4 所示。

表 4 摘要模型的分布向量与余弦相似度统计

Table 4 Distribution vector and cosine similarity statistics of summarization models

摘要模型	分布向量	余弦相似度
测试集参考摘要	[18 476,9 288,5 422,4 126,3 324,3 020]	1.000 000 00
Pointer-Generator(LSTM)	[22 167,7 719,2 377,986,593,437]	0.954 869 59
Pointer-Generator(LSTM,Coverage)	[18 599,6 946,2 337,998,569,434]	0.961 563 97
Multi-task(LSTM,Cluster-10)	[20 117,7 674,2 543,1 106,564,404]	0.961 783 02

本文设置域的数量 $X=6$ 并过滤源文本句子数量小于 X 的样本。分析表 4 数据可知,参考摘要共有 43 656 个句子在源文本中找到替代句,在不同域中的分布呈现急速下降趋势,表明 CNNDM 数据集中源文本头部的信息量较大。3 种模型生成的待测摘要分别有 34 279、29 883、32 408 个句子能够找到替代句。与测试集参考摘要相比,模型生成的摘要更加倾向于头部信息提取。Pointer-Generator(LSTM,Coverage)生成摘要的句子数量少于 Pointer-Generator(LSTM),其主要原因为覆盖度机制的引入缓解了重复生成问题。Multi-task(LSTM,Cluster-10)生成的摘要句子数量介于 Pointer-Generator(LSTM,Coverage)和 Pointer-Generator

(LSTM)之间,本文认为这是因为多任务学习摘要模型学习到了分类数据集中的抽象信息,其在保证句子数量的前提下更加集中于第 3 个和第 4 个域中信息的提取,其生成摘要与测试集参考摘要最接近。

5.3 样例分析

本文通过样例分析以验证多任务学习摘要模型学习到了分类数据集中的抽象信息。测试集参考摘要与模型生成摘要的样例如下:

1)测试集参考摘要:

“shelley dufresne was arrested in september when a student at the high school she taught at started bragging about sleeping with two teachers.it was later revealed that

the 16-year-old had sex with both dufresne and his former english teacher, 24-year-old rachel resposs. dufresne, 32, pleaded not guilty to charges in november, but changed course on thursday when she admitted having sex with the teen. in a forgiving plea deal, dufresne will only have to attend a 90-day therapy program, stay away from the victim and turn in her teacher's license. in exchange, the charge of carnal knowledge of a child will be dropped after her probation and she won't have to register as a sex offender. however, dufresne is still awaiting an arraignment on charges for having a threesome with the same student and resposs in a different parish."

2) Pointer-Generator(LSTM):

"shelley dufresne, 32, was arrested in september when a student at destrehan high school started bragging to friends at the st charles parish courthouse on thursday. she was arrested in september when a student at destrehan high school started bragging that he was sleeping with both of them. dufresne, 32, was arrested in september when a student at destrehan high school started bragging that he was sleeping with both of them."

3) Pointer-Generator(LSTM, Coverage):

"shelley dufresne, 32, was arrested in september after the teen in question started bragging to friends at the st charles parish, louisiana school that he had slept with two teachers. an investigation later revealed that the unidentified teen had sex with both his current english teacher at the time, dufresne, and his english teacher from the prior year, 24-year-old rachel resposs-including an alleged a threesome with both at resposs's house."

4) Multi-task(LSTM, Cluster-10):

"shelley dufresne confessed in court thursday to having sex with a 16-year-old student in exchange for a plea deal that gets her out of prison time. an investigation later revealed that the unidentified teen had sex with both his current english teacher at the time, dufresne, and his english teacher from the prior year. resposs taught the same student english the year prior. though she originally pleaded not guilty to a count of obscenity, the original charge of carnal knowledge of a juvenile will be held for the three years dufresne is under probation for obscenity."

对比样例显示, Pointer-Generator(LSTM)存在重复生成的问题("dufresne, 32, was arrested in september when a student at destrehan high school started bragging"出现了两次), Pointer-Generator(LSTM, Coverage)避免了此问题。但上述两种模型忽略了测试集参考摘要中“交换条件”的基本事实("in exchange, the charge of

carnal knowledge of a child will be dropped after her probation and she won't have to register as a sex offender"), Multi-task(LSTM, Cluster-10)生成的摘要包括“交换条件”的事实("in exchange for a plea deal that gets her out of prison time")。在对部分研究样例进行对比分析后, 本文发现多任务学习摘要模型更容易学习到时间点信息以及比较隐蔽的逻辑信息。

6 结束语

本文结合文本分类辅助任务, 提出一种多任务学习摘要模型, 使用 K-means 聚类算法构建 Cluster-2、Cluster-10 和 Cluster-20 文本分类数据集, 利用基于统计分布的判别法计算待测摘要与测试集参考摘要的分布向量在向量空间中的余弦相似度, 从总体分布的角度判断待测摘要的准确性。实验结果表明, 与现有摘要模型相比, 该模型生成的摘要更准确。由于 Transformer 模型架构具有优越的并行化序列处理能力以及特征提取能力, 因此后续可将 Transformer 模型架构和其他相关任务引入多任务学习摘要模型中, 进一步提升其在不同摘要数据集中的摘要生成质量。

参考文献

- [1] EVANGELOPOULOS G, ZLATINTSI A. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention[J]. IEEE Transactions on Multimedia, 2014, 15(7): 1553-1568.
- [2] JAYANTH J, SUNDARARAJ J, BHATTACHARYYA P. Monotone submodularity in opinion summaries [C]// Proceedings of Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: ACL Press, 2015: 169-178.
- [3] ZHANG Jiajun, ZHOU Yu, ZONG Chengqing. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(10): 1842-1853.
- [4] LIN H, BILMES J. Multi-document summarization via budgeted maximization of submodular functions [C]// Proceedings of 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. New York, USA: ACM Press, 2010: 912-920.
- [5] TAN Jiwei, WAN Xiaojun, XIAO Jianguo. Abstractive document summarization with a graph-based attentional neural model[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: ACL Press, 2017: 1171-1181.
- [6] LUONG M, LE Q, SUTSKEVER I, et al. Multi-task sequence to sequence learning[EB/OL]. [2020-01-06]. <https://arxiv.org/pdf/1511.06114.pdf>.

- [7] MA Shuming, SUN Xu, LIN Junyang, et al. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification [EB/OL]. [2020-01-06]. <https://arxiv.org/pdf/1805.01089.pdf>.
- [8] CHOPRA S, AULI M, RUSH A. Abstractive sentence summarization with attentive recurrent neural networks[C]//Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Philadelphia, USA: ACL Press, 2016: 93-98.
- [9] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[EB/OL]. [2020-01-06]. <https://arxiv.org/pdf/1409.0473.pdf>.
- [10] VINYALS O, FORTUNATO M, JAITLY N. Pointer networks[C]//Proceedings of Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2015: 2692-2700.
- [11] GU Jiatao, LU Zhengdong, LI Hang, et al. Incorporating copying mechanism in sequence-to-sequence learning [EB/OL]. [2020-01-06]. <https://arxiv.org/abs/1603.06393>.
- [12] TU Zhaopeng, LU Zhengdong, LIU Yang, et al. Modeling coverage for neural machine translation[EB/OL]. [2020-01-06]. <https://arxiv.org/pdf/1601.04811.pdf>.
- [13] SEE A, LIU P, 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. Philadelphia, USA: ACL Press, 2017: 1073-1083.
- [14] GUO Han, PASUNURU R, BANSAL M. Soft layer-specific multi-task summarization with entailment and question generation [EB/OL]. [2020-01-06]. <https://arxiv.org/pdf/1805.11004.pdf>.
- [15] ZHU Junnan, WANG Qian, WANG Yining, et al. NCLS: neural cross-lingual summarization [C]//Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Philadelphia, USA: ACL Press, 2019: 3045-3055.
- [16] MISHRA A, TAMILSELVAM S, RIDDHIMAN D, et al. Cognition-cognizant sentiment analysis with multitask subjectivity summarization based on annotators' gaze behavior[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2018: 1-8.
- [17] GEHRMANN S, DENG Y T, RUSH A. Bottom-up abstractive summarization [C]//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: ACL Press, 2018: 4098-4109.
- [18] ZHU Junnan, LI Haoran, LIU Tianshang, et al. MSMO: multimodal summarization with multimodal output[C]//Proceedings of 2018 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: ACL Press, 2018: 4154-4164.
- [19] LI Haoran, ZHU Junnan, MA Cong, et al. Multi-modal summarization for asynchronous collection of text, image, audio and video [C]//Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing. Philadelphia, USA: ACL Press, 2017: 1103-1113.
- [20] LI Haoran, ZHU Junnan, MA Cong, et al. Read, watch, listen, and summarize: multi-modal summarization for asynchronous text, image, audio and video [J]. IEEE Transactions on Knowledge and Data Engineering, 2018, 31(5): 996-1009.
- [21] ZHU Junnan, ZHOU Yu, ZHANG Jiajun, et al. Multimodal summarization with guidance with multimodal reference [C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI Press, 2020: 63-70.
- [22] CELIKYILMAZ A, BOSSELT A, HE X D, et al. Deep communicating agents for abstractive summarization [C]//Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Philadelphia, USA: ACL Press, 2018: 1662-1675.
- [23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2017: 5998-6008.
- [24] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
- [25] CHO K, MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [EB/OL]. [2020-01-06]. <https://arxiv.org/pdf/1406.1078.pdf>.
- [26] HERMANN M, KOCISKY T, GREFFENSTETTE E, et al. Teaching machines to read and comprehend [EB/OL]. [2020-01-06]. <https://arxiv.org/pdf/1506.03340.pdf>.
- [27] LIN C. ROUGE: a package for automatic evaluation of summaries [C]//Proceedings of Workshop on Text Summarization Branches Out. Philadelphia, USA: ACL Press, 2004: 74-81.
- [28] DUCHI J, HAZAN E, SINGER Y. Adaptive subgradient methods for online learning and stochastic optimization [J]. Journal of Machine Learning Research, 2011, 12(7): 2121-2159.
- [29] KINGMA D, BA J. Adam: a method for stochastic optimization [EB/OL]. [2020-01-06]. <https://arxiv.org/pdf/1412.6980v8.pdf>.