计算机科学
COMPUTER SCIENCE

基于知识蒸馏的抽取式自动摘要模型

赵江江，王洋，许楹楹，高扬

引用本文

赵江江，王洋，许楹楹，高扬.基于知识蒸馏的抽取式自动摘要模型[J]．计算机科学, 2023, 50(6A): 210300179-7.

ZHAO Jiangjiang, WANG Yang, XU Yingying, GAO Yang. Extractive Automatic Summarization Model Based on Knowledge Distillation [J]. Computer Science, 2023, 50(6A): 210300179-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300179

基于知识蒸馏的抽取式自动摘要模型

赵江江[1]　王　洋[2]　许楹楹[1]　高　扬[2]
1 中移在线服务有限公司　北京 100033
2 北京理工大学计算机学院　北京 100081
(zhaojiangjiang@cmos.chinamobile.com)

摘　要　抽取式自动摘要任务的目标是通过抽取原文中重要的句子来构成简短的摘要,同时保留原文中重要的内容。查询导向的抽取式摘要模型则可以进一步满足用户对摘要内容的不同需求。抽取式摘要模型具有能保证摘要内容正确性和句子可读性的天然优势,在此基础上确保摘要内容的相关性和显著性则成为了模型摘要目标的关键。为了实现抽取式摘要模型既满足查询的相关性又能保证摘要内容的显著性的目的,将查询信息作为模型学习的目标,利用摘要数据集的标题和图片信息额外构建了基于查询的扩展摘要数据集,并结合知识蒸馏方法提出了基于知识蒸馏的抽取式摘要模型。在实验中采用预训练语言模型 BERT 作为编码器,并结合知识蒸馏理论提出了两种模型训练策略:引导训练和蒸馏训练。在公开的新闻摘要数据集 CNN/DailyMail 上的实验结果证明,两种训练方法都取得了显著的效果。通过实验还发现,基于引导训练的摘要模型可以有效提高摘要内容的显著性,同时基于蒸馏训练的模型在提高摘要相关性和显著性方面达到了最好的效果。

关键词:查询导向抽取式摘要;扩展数据集;BERT;知识蒸馏
中图法分类号　TP391

Extractive Automatic Summarization Model Based on Knowledge Distillation

ZHAO Jiangjiang[1],WANG Yang[2],XU Yingying[1] and GAO Yang[2]
1 China Mobile Online Services Company Limited,Beijing 100033,China
2 School of Computer Science and Technology,Beijing Institute of Technology,Beijing 100081,China

Abstract　The objective of the extractive summarization is to extract the important sentences from the original text to form a short summary while retaining the important content of the original text. Query-focused extractive summary model can further satisfy users' different needs for summary content. Extractive summary model has the natural advantage of ensuring the correctness of summary and the readability of sentences. On this basis,ensuring the relevance and importance of the summary content is the key to the goal of the model. In order to satisfy the relevance of query and ensure the importance of summary content,this paper uses query information as a model study target,creates an extended summary data set based on the title and picture information,an extractive summary model based on knowledge distillation is proposed. In experiments,the pre-training language model BERT is adopted as the encoder,two model training strategies based on knowledge distillation theory are proposed:guided training and distillation training. Experimental results on CNN/DailyMail,a publicly available data set of news summaries,show that both training methods have achieved significant effects. It is also found that the model based on guiding training could effectively improve the significance of the summary content,while the model based on distillation training achieves the best effect in improving the relevance and significance of the summary.

Keywords　Query-focused extractive summarization,Extended data set,BERT,Knowledge distillation

1　引言

自动摘要是自然语言处理领域中一个重要的研究方向,而基于查询信息的自动摘要具有依据特定要求进行摘要生成的特点,可以满足不同用户对摘要内容的个性化要求。单文档自动摘要一般分为抽取式摘要和生成式摘要两种,抽取式摘要通过抽取原文档中的句子来构成摘要,而生成式摘要则由模型通过逐个词的生成来产生摘要,因此摘要会包含原文没有的词汇和句子。尽管生成式摘要的方法通过更深层次的架构能够提供更好的表意能力,但是抽取式摘要能够从原文中获取语义、语法正确且可读性良好的句子。相比之下,抽取式摘要更为保险且效率更高,能够满足在线服务所需的时效性。因此,本文针对单文档进行以查询为导向的抽取式摘要任务。然而,高质量的基于查询的摘要需要满足两个主要条件,即内容显著性与查询相关性。显著性确保摘要能够捕捉文档大部分的重要信息,实现摘要显著性主要有两种方法,第一种是利用特征、表征学习的方式对候选句子进行评分或者排序,并且选择得分高的句子作为摘要,第二种方法则是将模型抽取句子构成摘要的过程看作 0/1 分类的问题或者 0/1 序列标注的问题。相关性是确保摘要内容满足查询的需要,本文提出将相关性也作为摘要模型学习的目标,具体来说,就是将抽取式摘要任务拆分成两个具有具体目标的任务,

通信作者:高扬(gyang@bit.edu.cn)

第一个任务目标是预测摘要内容的显著性，第二个任务目标是预测摘要内容和查询信息的相关性。本文为这两个任务分别构建基于原始文档的标注数据集，一组是标准摘要数据集，即根据人类参考摘要构建，另一组是基于查询的摘要数据集，即根据查询信息(标题或图像)进行标注。因此，后者基于查询标注的数据集能够扩展出额外的与查询相关的知识，从而影响自动摘要模型。

本文提出了两种新的训练策略，即查询导向的引导训练方法和基于知识蒸馏的摘要模型训练方法，为了实现这样的训练方法，本文利用新的标注数据集进行训练。本文研究的基础抽取式摘要模型框架为多任务学习，并且利用了基于双向 Transformer[1]编码器的 BERT[2]编码表示。在不从根本上改变 BERT 结构的前提下，通过两个独立的[CLS]标记获得句子的表示，对句子间的语义信息进行协同训练，突出句子的相关性和显著性。具体来说，引导训练是典型的两阶段训练，第一阶段的训练目标是产生以查询信息为中心的摘要，即相关性，第二阶段的训练目标是产生以标准摘要为目标的摘要，即显著性。同时，本文借鉴了学生网络与教师网络具有相同架构的知识蒸馏网络[3]的成功经验，提出了一种新的蒸馏学习方法。在该网络中，通过同时匹配引导模型的预测输出和标准摘要，将查询相关知识从引导模型转移到面向显著性的学生模型，以进行蒸馏训练。在实验中验证中，引导训练和蒸馏训练可以相互加强，从而提高模型的整体效果。

为了证明所提模型的有效性，本文在两个基准数据集 CNN 和 DailyMail 上进行了实验。实验结果表明，该方法优于现有的方法。总的来说，本文的主要贡献如下：

(1)创造性地将以查询为中心的抽取式摘要定义为摘要显著性预测和相关性预测的多任务学习。此外，还创建了一个扩展摘要集，该集合可以利用与查询相关的知识对模型进行训练。

(2)通过插入两个相邻的[CLS]修改 BERT 的微调结构来获得句子的两个表示，从而分别进行摘要相关性和显著性的训练。因此，在提出的训练策略中，这些句子表示可以被充分地学习和被特定地使用。

(3)针对抽取式模型提出了两种有效的训练策略，即引导训练和蒸馏训练。引导训练在摘要的显著性方面表现最好，而蒸馏训练则利用引导训练模型成功地将查询相关知识融入到学生模型产生的摘要中，使模型整体兼顾了查询导向摘要的显著性和相关性。

2 相关工作

2.1 预训练语言模型

预训练语言模型(如 ELMo[4]，OpenAI GPT[5]，BERT[2]，RoBERTa[6]，AlBERT[7])在 NLP 任务表示学习方面取得了一系列突破，在 NLP 的各个领域取得了令人瞩目的成绩。典型的预训练操作是通过使用大量文本数据，根据上下文预测单词和进行句对预测任务来学习包含上下文内容的文本表示，在具体下游任务中进行微调训练即可。基于 BERT 的 BERTSUM[8]和 HIBERT[9]在抽取式摘要和生成式摘要任务中取得了成功，UniLM[10]统一了不同类型的语言模型，如单向、双向和序列到序列的预测，然后对模型进行微调，以帮助生成任务：摘要和问题生成。

2.2 自动摘要

抽取式摘要和生成式摘要是摘要系统的两种主要形式。本文主要研究基于查询的抽取式摘要，重点评价摘要与给定查询的相关性和摘要句子的显著性，在这两个方面的改进可以提高摘要的整体质量。在文档摘要任务中定义了一系列的特征来衡量相关性，包括 TF-IDF 特征、关键词、主题和概念相似性等。然而，这些特性通常缺乏文本语义的表达能力。更先进的神经网络模型最近将抽取式摘要任务作为分类任务处理，相关性和显著性总是通过端到端的方式一起训练。对于通用文档摘要，大多数模型都将抽取式摘要视为一个二分类问题，即预测一个句子是否作为摘要句。SummARuNNer[11]是采用递归神经网络作为句子编码器的最早的神经模型之一。REFRESH[12]采用了一个基于强化学习的子模块，该子模块是通过优化全局 ROUGE 评价分数作为目标来训练的。此外，抽取式摘要也可以建模为一个排序问题。例如，NEUSUM[13]联合评分和句子选择，而不是两阶段过程。最近，预训练语言模型的发展极大地促进了 NLP 领域的发展。对于摘要任务，HIBERT[9]框架将抽取式摘要作为序列标记问题，并应用分层双向 Transformer 编码器表示文档。BERTSUM[8]对改进的 BERT 模型进行了微调，用于抽取式和生成式摘要，并在抽取式摘要方面达到了最先进的水平。

对于以查询为中心的摘要，Cao 等[14]提出了一种基于注意力的模型，该模型联合计算句子显著性排名和查询相关性。为了利用上下文关系进行句子建模，Ren 等[15]提出了联合学习句子表示和上下文表示并注意基于查询特征的 CRSum 模型。目前来说，除了一些与问答系统相关的工作，很少有相关的工作是针对以查询为中心的单文档摘要。然而，这些问答系统相关工作很少用于摘要，因为他们只考虑了查询相关性，而没有考虑内容的显著性。

2.3 知识蒸馏

知识蒸馏通过训练学生模型来模仿教师模型的输出，把知识从教师模型转移到学生模型。Hinton 等[16]认为教师模型的输出中包含了训练数据中明确标签之外的额外信息。Zhang 等[17]进一步证明了“软目标”为计算熵损失带来了鲁棒性，并在正则化方面优化了模型。Phuong 等[18]则从理论上证明了蒸馏的成功源于数据类分离和优化偏差。知识蒸馏模型在分类、序列生成、自然语言理解等领域已经得到广泛应用。

压缩蒸馏是一种被广泛接受的方法，它将一个大的模型或集成网络压缩成一个小的模型[19]或高效模型[20]。这样做的目的是让学生模型更接近老师模型的预测能力。令人惊讶的是，再生模型[21]引起了极大的关注，因为在与教师模型相同的尺寸和结构设置下，学生模型的表现超过了教师模型。Clark 等[21]采用再生网络进行多任务学习，提供了更好的正则化和相关任务间的迁移。

3 基于预训练语言模型的抽取式摘要

3.1 BERT 模型

最近关于预训练语言模型的研究在文本表示的语境学习中表现出了显著的效果，并使各种 NLP 任务受益。这些

模型，如主要使用的 BERT[5]，从大规模语料库中学习了由 Transformer 编码的上下文表示，学习策略的本质是在大型自然语言语料库上进行预测被遮蔽词汇任务和下一个句子任务。

输入文本一般表示为一个单词序列 $D=[w_1,w_2,\cdots,w_n]$。然后，通过在基于 BERT 的结构中插入两个特殊标记，对所有的句子进行预处理。[CLS]标记被添加到文本的开头，[SEP]标记被插入到每个句子的后面，作为句子切分的指示。然后，为每个 w_i 分配 3 种嵌入，分别是表示每个单词的语义的词嵌入、表示两个句子之间的分割的分句嵌入以及表示每个单词在文本序列中的位置嵌入。将 3 个嵌入结合为单个向量 $\boldsymbol{x}_i$，并馈入由双向 Transformer 单元构成的 N 层结构中。每一层由两个子层组成：

$$\tilde{h}^l=LN(h^{l-1}+MHAtt(h^{l-1})) \tag{1}$$

$$h^l=LN(\tilde{h}^l+FFN(\tilde{h}^l)) \tag{2}$$

其中，h^0 是 T 经过位置编码后的表示，$\boldsymbol{T}$ 是 BERT 输出的句子向量表示，LN 是层标准化操作，$MHAtt$ 是多头注意力机制，FFN 是全连接层操作，l 表示多层的深度。

对于将抽取式摘要任务视为分类任务的模型只需要增加一个简单的分类层就可以完成摘要抽取任务，利用 sigmoid 激活函数来预测句子的得分：

$$\hat{Y}_i=\sigma(W\,T_i+b) \tag{3}$$

其中，σ 是 Sigmund 函数，T_i 是 BERT 输出的某个句子的向量表示。对于模型预测的每个句子分数 $\hat{Y}_i$，只需要与真实的标签计算损失并更新模型参数即可。

3.2 基于微调 BERT 的自动摘要

接下来，针对抽取式摘要任务对 BERT 参数进行微调。本文遵循 BERTSUM[8] 中使用的微调 BERT 的设置，它与原来的 BERT 略有不同。具体来说，为了表示单个句子，在每个句子的开头插入[CLS]标记，因此每个[CLS]都可以用来表示对应的句子。对于句子的词交替使用两种不同的分句嵌入，来区分词所属的句子。例如，对于一个文档$[s_1,s_2,s_3,s_4]$，分配的分段嵌入是$[E_A,E_B,E_A,E_B]$。最后，通过交叉熵函数将最顶层的句子表示与分类网络连接，并与特定的标签进行微调训练。大多数传统的抽取式摘要模型框架将摘要任务作为一个二分类问题处理。该方法建立了句子的表示，并在这些表示上应用二值分类器来预测是否应该将这些句子包含在摘要中。为了适应监督系统，给定预测分数 $P(c|D)$ 和标签 Y，损失函数可以通过交叉熵函数计算为：

$$L=-\sum_{i=1}^{m}(Y_i\log P(c|D_i)+(1-Y_i)\log(1-P(c|D_i))) \tag{4}$$

其中，i 是文档中句子的索引，m 是文档句子个数。

在实际情况中，人类书写的摘要通常是抽象的形式。当以监督的方式抽取摘要时，本文通常将具有最大重叠短语的文档原始句子标记为目标摘要句。本文关注以查询为中心的摘要，重点评估句子与给定查询的相关性以及句子的显著性。为此，将以查询为中心的抽取式摘要定义为显著性预测和相关性预测的多任务。

4 扩展数据集

很少有数据集是专门为基于查询的单文档摘要而构建的，为了完成该任务，本文将 CNN/DailyMail[22] 新闻数据集重构为文档-查询-摘要三元组。具体来说，CNN/DailyMail 新闻包含文章和它们相应的人工创建的摘要(highlight)，这些人工摘要还不能直接用于有监督的抽取式摘要训练。

本文使用贪婪搜索从原始文档中选择最佳的句子组合，以最大化与人类摘要评分的 ROUGE-2 Recall 和 ROUGE-2 F1 为目标创建目标摘要。当句子被添加到当前摘要集后，在剩余的候选语句加入不会再带来 ROUGE 分数的提高时，算法停止。本文将这个句子子集作为抽取的目标摘要返回，命名为标准摘要，标准摘要中包含的句子标记为 1，否则为 0。

查询信息是通过文档的标题和图像获得的。本文使用 Narayan 等[23] 提出的方式来提取数据集中每个文档对应的标题和图像描述，并将它们与相应的文档对应起来。不同文档的图片数量不同，只有 40% 的文档至少有一个图片。对于有多个图片的文档，本文选择第一个作为查询信息；而对于那些没有图像标题的文档，图像查询信息为空。

除了用于模型显著性训练的标准摘要外，本文还在原数据集的基础上利用标题和图片信息构建了两个用于模型相关性训练的以查询为中心的扩展摘要数据集：标题内容相关摘要和图片内容相关摘要。标签的执行方式与标准摘要集合类似，对原始文档进行查询信息的贪婪搜索，并将其标记为 1(值得作为查询摘要句)和 0(不值得作为查询摘要句)。

直觉上，位于文档前面的句子对于新闻类的文章很重要。Liu 等[8] 以及 Zhou 等[13] 也发现了 CNN/DailyMail 数据集更倾向于使用位置靠前作为标准摘要。为了显示查询为中心的信息的影响，图 1 分别给出了在标准摘要、标题相关摘要和图像相关摘要中所选摘要句子的不同位置分布。如图 1 所示，标题相关的摘要相对而言更倾向于首句，因为一篇组织良好的新闻通常会在文章的第一句话显示吸引眼球的信息，这与标题的效果类似。另外，新闻通常会在文章中插入图片来解释背景或状态，因此图像相关的摘要主要分布在中间。这两个独立的以查询为中心的摘要从不同的方面补充了标准摘要。

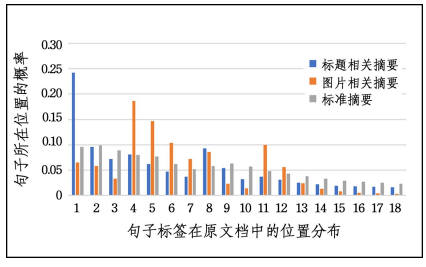


图 1 标题相关摘要、图片相关摘要和标准摘要训练集上的选择语句的位置分布

Fig. 1 Position distribution of selected sentences on the title-related summary, image-related summary and standard summary training

5 引导和蒸馏

本节介绍了一个基于微调 BERT 的抽取式摘要编码器，以及针对基于查询的单文档抽取式摘要提出的两种训练策略，即引导训练模型(Guidance Training Model，GTM)和蒸馏训练模型(Distillation Training Model，DTM)。摘要模型的框架如图 2 所示。

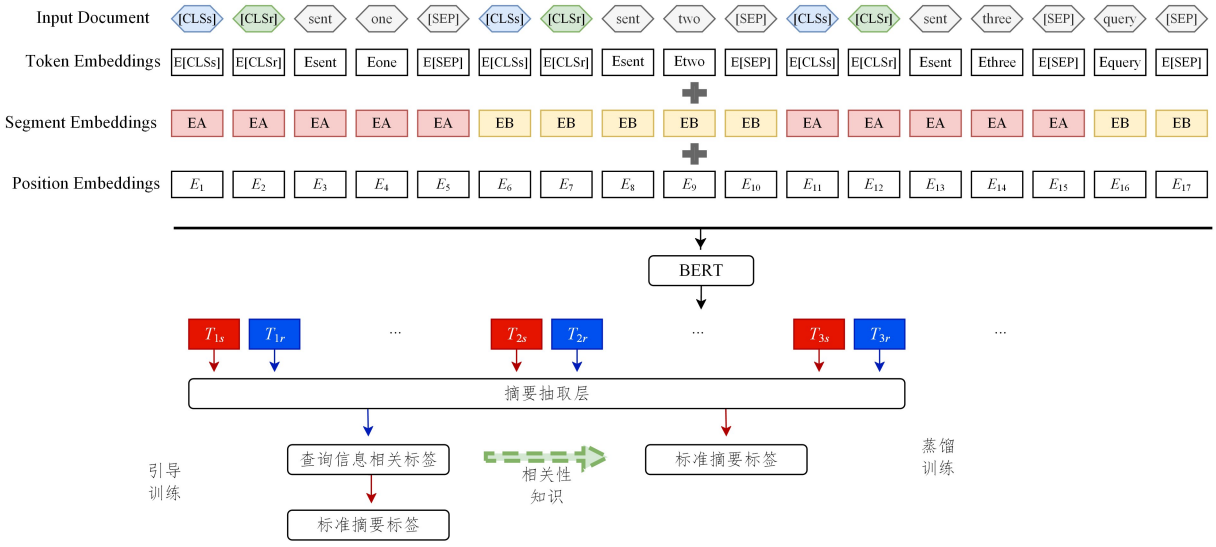


图 2 本文用于摘要抽取的摘要模型的总体框架

Fig. 2 General framework of the summary model used in this paper for summary extraction

5.1 抽取式摘要编码器

本节将基于第 3 节中的微调 BERT 和第 4 节中的标记摘要数据集，介绍一种用于基于查询的抽取式摘要的编码器。

输入文本 $D=[w_1,w_2,\cdots,w_n]$ 的序列长度为 n，与输入的查询项序列对齐，$Q=(q_1,q_2,\cdots,q_t)$ 的长度为 t。第 3 节描述的微调 BERT 结构的差异中，所有的句子都被插入了 3 个特殊标记。具体来说，其中两个特殊标记[CLSs]和[CLSr]被插入到每个句子的前面，分别代表用于相关性计算的表示和用于显著性计算的表示，如图 2 所示。在面向任务的微调过程中，附加在句子中的两个相邻的[CLS]在句子间转换层中交互地接受训练。这样，通过不同类型的摘要标签，在丰富的上下文信息中充分捕捉到具体句子的相关性和显著性。另一种特殊标记是在每个句子后面插入[SEP]，将其作为句子分割的标记。词嵌入、分句嵌入(即 E_A 和 E_B)和位置嵌入的设置与第 3 节中的设置相似。

微调编码器层指用于微调的编码器层，是一个两层的句子级 Transformer 编码器，位于 BERT 的最后一层之上，这一层很好地捕获了文档级语义信息。然后，将结果输入到一个 sigmoid 分类器中，它预测一个句子 i 是否应该包含在摘要中，表达式如下：

$$P(c|D_i)=\sigma(W T_i+b) \tag{5}$$

其中，$P(c|D_i)$ 为句子 i 的得分，T_i 为句子 i 对应的[CLS]的表示，其中使用[CLSr]来匹配查询导向的摘要标签，标准摘要标签使用[CLSs]来匹配。模型的损失如式(4)所示。

5.2 引导训练

为了在训练抽取式摘要编码器时同时考虑摘要的相关性和显著性，本文提出了两阶段微调的方法。第一个阶段是将基于查询导向的摘要作为目标(第 4 节中构造的查询相关的摘要)进行训练，其中[CLSr]被提取为句子的表示，损失函数如式(4)所示。第二阶段紧接着第一阶段的微调模型，继续用标准摘要的标签进行训练，其中[CLSs]被提取为句子表示，这个策略简单而直接。它对以查询为导向的任务有效的原因为：1)以查询为导向的摘要和标准摘要共享了信息，加强了抽取式摘要编码器的有用上下文表示；2)引导训练策略训练的句子表征保持了摘要的相关性和显著性。

需要强调的是，虽然两阶段的微调目标明确，但是实验中第二阶段的微调效果过于明显，导致第一阶段微调的效果下降太多。因此本文将第二阶段微调的目标标签升级为在标准摘要标签和第一阶段预测的查询的相关摘要之间进行权衡。损失函数如下：

$$L_G=(1-\lambda_g)F_c(Y_i,P(c|D_i))+\lambda_g F_c(\ell_{ri},P(c|D_i)) \tag{6}$$

其中，ℓ_{ri} 为第一阶段输出预测中句子 i 的相关性得分，$P(c|D_i)$ 为当前训练模型的预测得分，F_c 为交叉熵函数，Y_i 为标准摘要的标签，λ_g 是权衡参数。

5.3 蒸馏训练

知识蒸馏的原理是让学生模型不仅通过真实标签提供的信息进行训练，还通过教师模型携带的知识进行训练。然而，在本文的案例中，以显著性为导向的摘要与以相关性为导向的摘要略有不同，本文认为不能过分依赖教师模型的表现。为了缓解这一问题，从以查询为导向的摘要中提炼出的知识可以被视为一种引导，而不是“教师”，对学习过程产生轻微的影响和引导。引导知识是查询导向摘要系统中所需要的，因此本文提出了一种新的蒸馏学习方法，通过将相关知识从引导模型转移到以显著性为导向的学生模型来进行蒸馏训练。

继承重生网络的思想，学生模型可以与引导模型具有相同的架构。训练摘要网络匹配标准摘要的标签来学习显著性，并匹配引导训练后模型的预测输出来获得查询相关知识，将标准摘要标签视为“硬目标”，将引导训练第一阶段的输出视为“软目标”，构造基于知识蒸馏的抽取式摘要模型训练方法。损失函数为：

$$L_D=(1-\lambda_d)F_c(Y_i,P(c|D_i,\theta'))+\lambda_d F_{L_2}(\ell_{ri},P(c|D_i,\theta')) \tag{7}$$

其中，ℓ_{ri} 是引导训练第一阶段的输出，即每个句子的相关性得分；$P(c|D_i,\theta')$ 是学生的分数预测模型；F_{L_2} 是基于 $L2$ 距离的损失函数，本文训练了学生模型的 λ_d 权衡参数，使其预测和引导之间的 $L2$ 距离最小化，而不是使用交叉熵损失，知识蒸馏从本质上提高了模型的性能，这是由于输出中经过良好训练的丰富知识完全分布在标签上，用 L_2 规范化可以更好地计算回归分类任务；λ_d 是权衡参数。

引导训练是一个两阶段的训练，模型的参数经历了两次学习。蒸馏训练基于知识蒸馏的理论将引导训练的第一阶段输出作为一种“软目标”和标准摘要(硬目标)相结合进行模型

训练，模型的参数只经历了一次学习。

6 实验

本节在两个数据集 CNN 和 DailyMail[22] 上进行了全面的实验，以分析和研究本文所提模型的有效性和整体性能。本文尤其关注了模型抽取的摘要在显著性和相关性方面的表现，并对结果进行了进一步分析。

6.1 数据集和评价指标

本文在第 4 节介绍的 CNN 和 DailyMail 两个基准数据集上评估了所提模型，这些数据集的统计信息如表 1 所列，其中文档和摘要长度以句子为单位，标题和图片描述以单词为单位。数据集在未匿名实体的情况下进行了预处理，所有的句子都是用斯坦福 CoreNLP 工具包[24] 分割的，输入文档长度被截断为 512，对于 BERT 输入的查询信息长度被截断为最多 12。由于加入了查询信息和额外的[CLS]，有效的输入内容被缩短。

本文使用 ROUGE[25] 作为评价指标，通过计算候选摘要和参考摘要之间的重叠词汇元素来度量摘要的质量。按照之前的做法，本文评估了将 ROUGE-1，ROUGE-2 作为评估摘要内容信息量的方法，以及将 ROUGE-L 作为评估摘要流畅度的方法，同时给出了它们的 F1 和 Recall 值，如表 2－表 4 所列。

表 1 实验数据集统计

Table 1 Experimental dataset statistics

Datasets	# docs(train/val/test)	avg. doc length	avg. summary length	avg. Title	avg. Image Captions
CNN	90 266/1 220/1 093	33.98	3.59	9.8	19.5
Daily Mail	196 961/12 148/10 397	29.33	3.86	18.1	21.2

注：表中的数据分别为训练集、验证集和测试集的大小，平均文档和摘要长度以及标题/图像单词的平均长度

表 2 模型在 CNN/DailyMail 数据集上实验的显著性得分 ROUGE-F1

Table 2 Significance scores ROUGE-F1 for model experiments on the CNN/DailyMail dataset

Models	ROUGE-1	ROUGE-2	ROUGE-L
标准摘要	52.59	31.24	48.87
LEAD-3	40.37	17.45	36.61
TextRank	39.53	17.28	36.38
SummARuNNer	39.60	16.20	35.30
REFRESH	40.98	17.85	37.10
LATENT	41.05	18.77	37.54
BANDITSUM	41.50	18.70	37.60
NEUSUM	41.59	19.01	37.98
HiBERT	42.31	19.87	38.78
BERTSUM	43.25	20.24	39.63
Basic model	43.14	20.10	39.51
GTM(Title)	43.34	20.28	39.72
DTM(Title)	**43.36**	**20.30**	**39.78**

表 3 相关性：模型在 CNN/DailyMail 数据集上的 ROUGE-Recall 得分

Table 3 Correlation: ROUGE-Recall scores for model experiments on the CNN/DailyMail dataset

Models	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	52.38	22.75	47.47
AttSum	52.81	22.95	47.74
CRSUM	53.10	23.15	48.07
XNET	52.72	22.95	48.10
Basic model	53.19	24.78	48.66
DTM(Title)	**53.92**	**25.13**	**49.29**
DTM(Image)	53.30	24.97	48.77

表 4 模型在基于标题信息的扩展摘要数据集上的 ROUGE-F1 得分

Table 4 ROUGE-F1 scores for the model's experiments on the extended summary dataset based on title information

Models	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	16.59	4.95	13.04
AttSum	17.17	5.24	13.84
CRSUM	17.47	5.31	13.91
XNET	17.01	5.17	13.74
Basic model	17.73	5.68	14.36
GTM First Stage(Title)	**18.58**	**6.43**	**15.62**
GTM Second Stage(Title)	18.41	6.36	15.48
DTM(Title)	18.53	6.4	15.53

6.2 基线模型

本文将基线模型分为 3 类，分别包括无监督方法、无查询的监督方法和有查询的监督方法。

(1)无监督模型。LEAD-3 基线选择每个文档中的前三句话作为摘要。虽然很直接简单，但在 CNN/DailyMail 上却是不容易达到的基线，因为突出的句子大多集中在文档的开头。TextRank[26] 是基于 PageRank 开发的基于加权图的无监督摘要算法。

(2)无查询的监督模型。SummARuNNer[11] 考虑新奇性、突出性和位置作为特征，用 RNN 网络进行句子表示；REFRESH[12] 利用强化学习算法对 ROUGE 评价结果进行全局优化来训练模型，以提高自动评价的得分；LATENT[27] 把句子看成是潜在变量；BANDITSUM[28] 把它当作一个上下文“强盗”来预测选择行为；NEUSUM[13] 将选择策略集成到评分模型中；HiBERT[9] 在 BERT 之上使用了层次表示；BERTSUM[8] 是一个改良的微调伯特模型。

(3)查询导向的监督模型。AttSum[14] 联合学习了查询相关性排序和句子显著性排序。CRSUM[15] 考虑了句子的上下文特征。本文使用相同的训练数据集实现了这个基线模型，并计算了查询相关性。XNET[23] 基于 REFRESH 模型的基本结构并把查询信息作为外部信息融入到模型学习中。

6.3 实验细节

本文使用的 BERT 是“bert-base-uncased”版本。源词汇和目标词汇都用 BERT 的词集进行分词处理。Transformer 层隐藏单元尺寸为 768，前馈层共有 2 048 个隐藏单元，使用两个 Transformer 层来获得句子表示，head 数目为 8，dropout 为 0.1。引导训练和蒸馏训练的每一阶段训练 5 万步，每 2 步进行梯度积累在 1 块 GPU(GTX 1080 Ti)上进行训练。在预测过程中，使用 n 元重合方法来减少句子冗余，n 取值 3。模型每 5 000 步被验证并保存。通过验证过程选择了最好的 3 个模型，并在测试集上进行测试获得结果取平均，使用了 Adam 优化器，$\beta_1=0.9$，$\beta_2=0.999$。模型训练前 1 万步是预热阶段，学习率变化遵循文献[8]中的设置。抽取式摘要模型训练的时间如表 5 所列。

表 5 抽取式摘要模型的训练时长

Table 5 Training time of the extractive summary model

(单位:h)

Models	Training duration
Basic model	5～6
GTM(Title)	10～12
DTM(Title)	5～6

6.4 ROUGE 结果分析

6.4.1 显著性分析

本文使所有方法选择得分最高的前 3 个句子作为摘要进行公平比较。作为上界,本文提供了一个有效的标准系统。统计结果表明,抽取式摘要内容的显著性还有很大的提高空间。表 2 列出了所有模型的 ROUGE-F1 指标的测试结果,包括 2 个无监督的基线模型 LEAD-3 和 TextRank、5 个基于神经网络的抽取式摘要模型以及 2 个基于 BERT 的抽取式摘要模型,其中 BERTSUM 在所有基线中获得了最好的结果。最后是所提模型的表现,引导训练模型(GTM)使用本文提出的引导训练方式进行训练,λ_g 取值为 0.1。蒸馏训练模型(DTM)是基于引导训练得到的输出分布的一次蒸馏,引导训练使用以标题(Title)相关的扩展摘要数据集,其中损失按 L2 距离计算,如 5.3 节所述。

为了验证所提模型所产生摘要内容的显著性,表 2 中的所有基线都是无查询的摘要,通过与推荐摘要进行比较来评估(本文认为推荐摘要是 CNN/DailyMail 人工编辑提供的标准重要摘要)。本文模型的性能比 BERTSUM 略差,这是因为本文模型的输入加入了查询和两倍数量的[CLS],因此与 BERTSUM 相比,在 BERT 结构中编码的有效内容较少,但提出的引导训练在很大程度上提高了基本模型的性能,从而超越了 BERTSUM,证明了引导训练策略有效地提高了摘要的显著性。此外,它还表明,以查询为中心的扩展摘要数据集也有助于共享表征学习的重要信息。采用蒸馏训练的方法效果最好,证明了知识蒸馏策略在抽取式摘要任务中的有效性。

6.4.2 相关性分析

为了研究相关性,本文计算了 ROUGE 的召回分数,它更适合评价以查询为中心的抽取式摘要。因为与推荐摘要相比,摘要需要收集更多与查询内容相关的句子。在表 3 中,本文实验了两种类型的查询,即标题(title)和图片(image),基本模型的性能超过了所有的基线。此外,两种蒸馏系统,标题查询和图像查询如预期得到了更相关摘要,其中标题蒸馏达到最佳($\lambda_d=0.3$)。蒸馏训练的显著效果表明,以查询为中心的关联知识已成功地从引导模型转移到学生模型。参数λ_d对 DTM(Title)模型摘要效果的影响,使用 ROUGE 评估的 F1 值增益作为指标,如图 3 所示。

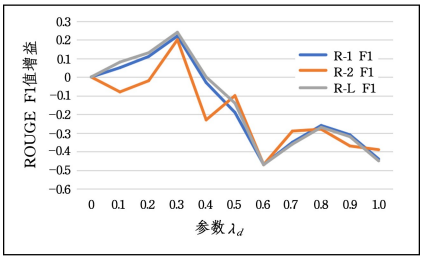


图 3 蒸馏训练中参数λ_d对模型效果的影响

Fig. 3 Effect of the paramete λ_d on the model effect in distillation training

知识蒸馏方法所具有的知识迁移的能力,可以将查询信息中的关键知识迁移到摘要模型中,让模型可以在进行摘要抽取时关注到重要的信息,提高模型的摘要效果。

无论是从摘要的显著性角度还是相关性角度,本文模型都优于所有的基线,模型对于查询信息的有效处理不仅提升了摘要的相关性表现,还进一步提高了模型产生摘要的显著性表现。

6.5 消融实验

为了验证引导训练的两阶段微调方法的有效性和效果以及蒸馏训练中加入"软目标"对模型的影响,本文进行了消融实验。通过将模型产生的摘要在基于标题信息的扩展数据集(Title)上进行 ROUGE 评估的方式,采用 ROUGE-F1 的得分结果来验证引导训练和蒸馏训练对摘要模型查询相关性的直接影响,结果如表 4 所列。

从实验结果中可以发现,引导训练的第一阶段(GTM First Stage)训练后的测试结果与其他基线模型相比给摘要相关性带来了显著提升。在进行第二阶段(GTM Second Stage)训练后模型所产生的摘要在相关性方面只出现了略微的下降,同时在正式的评价结果(见表 2)中有着很好的表现,这与本文模型使用两个独立的[CLSr]和[CLSs]有着密不可分的关系,在实验中发现采用同一个[CLS]会导致第二阶段训练后摘要相关性的显著下降,这可能是由于之前训练好的参数被新的训练覆盖导致的结果。

蒸馏训练模型(DTM)方法中相关性被作为"软目标"加入到模型的训练目标中,同样给摘要的相关性带来了较大提升,虽然效果没有引导训练第一阶段的效果明显,但是从正式的测试结果(见表 2 和表 3)可以得出蒸馏训练给模型摘要的相关性和显著性带来了综合提升,在所有模型中取得了最好的效果。

结束语 本文将以查询为中心的抽取式摘要分为面向显著性的摘要和面向相关性的摘要,并提出了两种训练策略:引导训练和蒸馏训练。相比在传统方法中通过计算句子与查询信息相似度来进行句子筛选,本文还同时考虑了摘要内容的显著性;相比将查询信息视为一种额外的注意力的方法,本文更倾向于将其视为目标的一部分,通过模型参数的学习来生成与查询更相关的摘要内容。

实验充分展示和分析了所提出的引导和蒸馏训练策略如何应用于改进 BERT 微调以获得摘要。实验结果证明,本文模型在整体结果上优于目前最先进的基线模型,同时摘要相关性的提升并没有导致显著性的下降,反而提高了摘要的 ROUGE 评价结果,这也证明了本文对于查询信息的处理是合理且有效的。同时,本文扩展的查询导向摘要数据集对模型的训练也起到了重要作用。未来工作可以进一步研究和探索多任务学习在摘要领域的应用。同时在数据集方面为面向查询的摘要任务创造更多有价值的训练和测试数据,例如面向用户评价内容的摘要等。

参 考 文 献

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[2] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of

Deep Bidirectional Transformers for Language Understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 4171-4186.

[3] FURLANELLO T, LIPTON Z, TSCHANNEN M, et al. Born-Again Neural Networks[C]// International Conference on Machine Learning. 2018: 1602-1611.

[4] EPETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv: 1802. 05365, 2018.

[5] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[J/OL]. 2018. https://scholar. google. com/scholar?q = Improving + Language + Understanding + by + generative + pre-training&hl = zh-CN&as_sdt = 0&as_vis = 1&oi = scholart.

[6] LIU Y H, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv: 1907. 11692, 2019.

[7] LAN Z Z, CHEN M D, GOODMAN S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv: 1909. 11942, 2019.

[8] LIU Y, LAPATA M. Text Summarization with Pretrained Encoders[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 3721-3731.

[9] ZHANG X X, WEI F R, ZHOU M. HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization[C]// Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019). Florence, Italy, 2019: 5059-5069.

[10] DONG L, YANG N, WANG W H, et al. Unified Language Model Pretraining for Natural Language Understanding and Generation[J]. arXiv: 1905. 03197, 2019.

[11] NALLAPATI R, ZHAI F F, ZHOU W. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents[C]// Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[12] NARAYAN S, COHEN S B, LAPATA M. Ranking Sentences for Extractive Summarization with Reinforcement Learning [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long Papers). 2018: 1747-1759.

[13] ZHOU Q Y, YANG N, WEI F R, et al. Neural Document Summarization by Jointly Learning to Score and Select Sentences [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers). 2018: 654-663.

[14] CAO Z Q, LI W J, LI S J, et al. AttSum: Joint Learning of Focusing and Summarization with Neural Attention[C]// The 26th International Conference on Computational Linguistics (COLING 2016). 2016: 547-556.

[15] REN P J, CHEN Z M, REN Z C, et al. Leveraging contextual sentence relations for extractive summarization using a neural attention model [C] // Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2017: 95-104.

[16] HINTON G, VINYALS O, DEAN J. Distilling theknowledge in a neural network[J]. arXiv: 1503. 02531, 2015.

[17] ZHANG Y, XIANG T, HOSPEDALES T M, et al. Deep mutual learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4320-4328.

[18] PHUONG M, LAMPERT C. Towards Understanding Knowledge Distillation [C] // International Conference on Machine Learning. 2019: 5142-5151.

[19] ROMERO A, BALLAS N, KAHOU S E, et al. Fitnets: Hints for thin deep nets[J]. arXiv: 1412. 6550, 2014.

[20] POLINO A, PASCANU R, ALISTARH D. Model compression via distillation and quantization[J]. arXiv: 1802. 05668, 2018.

[21] CLARK K, LUONG M T, KHANDELWAL U, et al. BAM! Born-Again Multi-Task Networks for Natural Language Understanding[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 5931-5937.

[22] HERMANN K M, KOCISKY T, GREFENSTETTE E, et al. Teaching machines to read and comprehend[C]// Advances in Neural Information Processing Systems. 2015: 1693-1701.

[23] NARAYAN S, CARDENAS R, PAPASARANTOPOULOS N, et al. Document modeling with external attention for sentence extraction[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers). 2018: 2020-2030.

[24] MANNING C D, SURDEANU M, BAUER J, et al. The Stanford CoreNLP Natural Language Processing Toolkit[C]// Association for Computational Linguistics (ACL) System Demonstrations. 2014: 55-60. http://www. aclweb. org/anthology/P/P14/P14-5010.

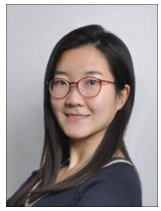
[25] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]// Text Summarization Branches Out. 2004: 74-81.

[26] MIHALCEA R, TARAU P. TextRank: Bringing Order into Text[C] // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2004: 404-411.

[27] ZHANG X X, LAPATA M, WEI F R, et al. Neural Latent Extractive Document Summarization[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 779-784.

[28] DONG Y, SHEN Y K, CRAWFORD E, et al. BanditSum: Extractive Summarization as a Contextual Bandit[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 3739-3748.

ZHAO Jiangjiang, born in 1987, Ph. D candidate, is a member of China Computer Federation. His main research interests include open domian dialog system, information extraction and network representation learning and natural language processing.



GAO Yang, born in 1987, Ph. D, associate professor, is a member of China Computer Federation. Her main research interests include information extraction, network representation learning and natural language processing.