文章编号:2096-1472(2024)02-0001-09
DOI:10.19644/j.cnki.issn2096-1472.2024.002.001

面向查询的文本摘要关键技术研究综述

徐 睿[1]，刘 金[1]，亚 静[2,3]

(1.华北计算机系统工程研究所，北京 102209；
2.京北方信息技术股份有限公司，北京 100081；
3.北京大学智能学院，北京 100871)
✉ xurui@ncse.com.cn; liujin972181761@126.com; jing.ya@pku.edu.cn

摘 要:面向查询的文本摘要是自动文摘中的一个特殊领域,可以根据用户个性化的查询需求,从原始文档或文档集中提取有价值的摘要信息。目前,该技术已经在面向查询的搜索引擎、智能化信息检索、问答系统等领域得到广泛应用,并受到越来越多的关注。文章基于面向查询的文本摘要任务的典型技术框架,从查询理解、文档处理和信息组织三个方面对其国内外研究方法的现状进行对比和分析,对不同业务场景的应用进行了举例,归纳了面向查询的文本摘要面临的挑战及发展趋势。

关键词:面向查询;文本摘要;自然语言处理
中图分类号:TP391 **文献标志码**:A

Research Overview of Key Techniques for Query-Focused Summarization

XU Rui[1], LIU Jin[1], YA Jing[2,3]

(1.*National Computer System Engineering Research Institute of China, Beijing* 102209, *China*;
2.*Northking Information Technology Co., Ltd., Beijing* 100081, *China*;
3.*School of Intelligence Science and Technology, Peking University, Beijing* 100871, *China*)
✉ xurui@ncse.com.cn; liujin972181761@126.com; jing.ya@pku.edu.cn

Abstract: Query-Focused Summarization (QFS) is a special area of automatic summarization, which can be used to extract valuable summary information from original document or multi-documents based on the user's personalized query requirements. The technology has been widely used in the search engine, intelligent information retrieval, Q&A system and other areas, and has received more and more attention. In this paper, we start from the typical technical framework for achieving this task, compare and analyze the existing research methods in three aspects: query comprehension, document processing and information organization, and summarize the challenges and development trends QFS.

Key words: query-focused; summarization; natural language processing

0 引言(Introduction)

随着信息技术的不断进步,各类信息数量快速增长,促进了信息的交流与共享。在信息获取过程中,如何利用先进技术从海量的复杂数据中更快速、准确地筛选出有价值的信息,成为各机构及专家、学者的研究热点。面向查询的文本摘要(Query-Focused Summarization, QFS)是自动文摘的一个特殊领域,旨在依据用户的查询需求,从源文档中自动提取重要信息,将其组织成与查询相关的简短摘要进行呈现。与通用文本

收稿日期:2023-07-01

摘要不同，面向查询的文本摘要主要面向特定用户，文本摘要内容不仅是对原始静态文本的反映，更带有主观倾向及侧重，满足个性化查询的需求，通常又被称为针对式文本摘要、面向用户的文本摘要或面向主题的文本摘要[1-3]。面向查询的文本摘要在面向查询的搜索引擎、智能化信息检索、问答系统中均有着重要的应用。

本文基于面向查询的文本摘要典型技术框架，从查询理解、文档处理和信息组织三个方面对其国内外研究现状进行梳理和分析，总结当前技术应用的现状、存在的问题及面临的挑战，分析未来发展趋势。

1 研究背景(Research background)

1.1 典型技术框架

面向查询的文本摘要任务的典型技术框架如图1所示，输入的文档类型包含单文档与多文档两类。与面向查询的单文档文本摘要相比，面向查询的多文档文本摘要(Query-Focused Multi-Document Summarization，QMDS)对具有相同话题的文档集进行了统一处理，可以满足用户全方位查询的需要，具备更高的应用价值。但是，鉴于不同文档可能会包含相同信息，多文档文摘需要充分考虑信息冗余(Redundancy)因素，消除冗余影响。1997年，CARBONELL[4]首次提出面向查询的文摘任务，并提出最大边界相关(Maximal Marginal Relevance，MMR)算法，考察查询相关性，作为语句之间的冗余消除策略[5]，提高摘要信息提取的准确性。针对输入的查询和文档信息处理，包括查询理解、文档处理和信息组织三个步骤。查询理解主要实现对用户查询意图的细化分析和理解；文档处理主要完成对文档或文档集内语句信息的处理，摘录候选语句或生成新的摘要语句；信息组织建模主要实现摘要语句的连贯性处理，保证输出信息可读。

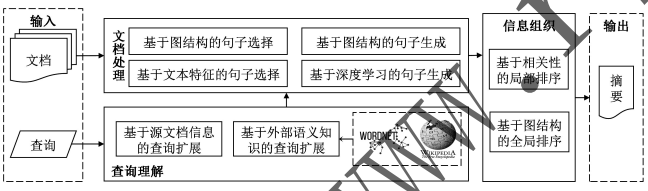

图1 面向查询的文本摘要任务的典型技术框架

Fig. 1 Typical technology framework of query-focused summarization tasks

针对上述框架，现有研究主要为不同的应用场景提供不同的解决方案。针对多个技术点提供统一的解决方案仍需要技术突破，其技术挑战主要来自三个方面：(1)用户输入的查询信息概括性强，包含内容有限，文档间存在语义描述鸿沟，因此生成与查询相关性强的概括性摘要，准确反映原文档信息成为难点；(2)信息量的快速增长导致过载问题严重，因此在限定空间对文本信息进行合理压缩，进而容纳更多有价值的内容成为难点；(3)抽取或生成的文摘语句的排列顺序会直接影响文摘可读性，因此确定生成的文本摘要语句的排列顺序成为难点。

1.2 文摘评测

对文摘信息的合理程度进行评测，是保证文摘质量的重要方法之一，通常可分为内部评测和外部评测。文摘评测方法如图2所示。

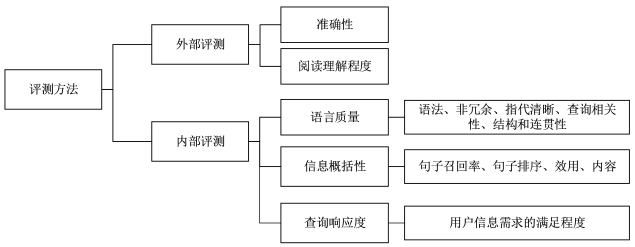

图2 文摘评测方法[6]

Fig. 2 Summarization evaluation methods

外部评测主要是将生成的摘要信息应用到实际的信息检索、问答系统任务中，根据对任务完成的贡献进行摘要的性能评测，通过与参考摘要进行对比，评测当前摘要内容的准确性和对原始内容的阅读理解程度。外部评测受相关任务的影响较大，因此对文摘的评测多采用内部评测，即根据独立的自动文摘系统的语言质量、信息概括性和查询响应度进行评测。

早期的内部评测多采用语言质量和查询响应度等指标进行评分。为节约时间和降低成本，在信息的概括性方面，研究人员提出了多种自动化评测方法。最常用的评测指标是由LIN[7]提出的基于内容的文本摘要自动评价方法ROUGE(Recall-Oriented Understudy for Gisting Evaluation)，其主要思想是将机器生成的文本摘要信息与人工总结的参考文摘进行对比，通过重叠的单词序列、N-Gram模型对摘要进行评价，具体准则包括基于N-Gram召回率的ROUGE-N，通过计算公共子系列匹配率获取最长公共子序列的ROUGE-L，基于权重的最长公共子序列的ROUGE-W，基于间隔二元组(Skip-Bigram)重叠度的ROUGE-S等。

对文本摘要生成模型进行训练和自动评测，需要依靠各种数据集，目前已公开的经典数据集概览如表1所示。

表1 数据集概览

Tab.1 Datasets overview

数据集名称	任务	说明	查询数/次	总文档数/个
DUC(Document Understanding Conference)2005	多文档摘要	DUC/TAC提供的数据集都是小型数据集，用来进行模型评测。DUC从2005年开始设置面向查询的多文档摘要任务；2008年，DUC会议停办，由TAC取代	50	1 593
DUC 2006	多文档摘要		50	1 250
DUC 2007	多文档摘要		45	1 125
TAC(Text Analysis Conference) 2008	多文档摘要		48	480
TAC 2009	多文档摘要		44	440
Wikipedia(维基百科)	单文档摘要	MISHRA等[8]基于维基百科数据构造了面向查询的抽取式文本摘要数据集。从维基百科1987－2007年发生的历史事件中随机选取100个事件，事件的简短描述作为查询，页面作为摘要，互联网上发布的新闻作为原始文档	100	100

续表

数据集名称	任务	说明	查询数/次	总文档数/个
WikiQA(Wikipedia Open-domain Question Answering)	多文档摘要	WikiQA是微软发布的一个开放域问答数据[9]。使用Bing查询日志作为问题源,维基百科页面的摘要部分作为问题候选答案	3 047	29 258
Debatepedia	单文档摘要	Debatepedia是一个关于辩论的"维基百科"网站,NEMA等[10]基于网站数据构造数据集,辩题作为查询,论点作为摘要,论述作为文档	12 695	12 695
CNN(Cable News Network)/Dailymail	单文档摘要	CNN/Dailymail[11]是针对通用性摘要的数据集。KRISHNA等[12]对CNN/Dailymail进行改造,利用2017年KDD(Knowledge Discovery and Data Mining)VoxMedia新闻数据集[13]对CNN/Dailymail数据添加主题信息,得到一个面向主题的文本摘要任务数据集	112 360	112 360

针对尚缺乏大量多文档摘要数据集的问题,PASUNURU等[14]通过汇总有线电视新闻网(CNN)和每日邮件信息以及挖掘搜索日志的方式,形成文档集,并进行查询模拟。

2　查询理解(Query understanding)

根据用户有限的输入理解其查询意图,是面向查询的文本摘要需要解决的重要问题之一。早期研究大多采用在通用文摘中加入查询相关特征的方法,对查询文本做简单的处理,如关键词抽取、词权重计算,缺少对查询的深入理解[15]。查询通常具有概括性,在多文档中的描述有所不同,存在语义描述鸿沟。通过查询扩展的方式可以有效解决信息的限制问题,弥补语义的缺失。目前,查询理解技术多被应用于抽取式摘要中,本文主要介绍基于外部语义知识的查询扩展和利用源文档信息的查询扩展。

2.1　基于外部语义知识的查询扩展

基于外部语义知识的查询扩展技术利用外部知识,学习查询词与文档词的相似性,实现对查询词的同义词进行扩展,达到提高查询能力的目的。常用的外部语义知识扩展包括基于WordNet的查询扩展和基于维基百科的查询扩展两类。

WordNet是一种基于认知语言学的英文语义词典,由普林斯顿大学设计开发。与传统词典按照字母顺序进行组织不同,WordNet将词汇划分为名词、动词、形容词、副词和虚词5类,每类词汇各自被组织成同义词网络,代表基本的语义概念,依据语义关系进行连接。WordNet常用的语义关系包括同义关系、反义关系、上位关系、下位关系、整体关系、部分关系、蕴含关系、因果关系和等级关系等。ZHOU等[16]利用TF-IDF(Term Frequency-Inverse Document Frequency)算法计算查询词的重要性程度,基于WordNet对超过重要性阈值的名词和动词进行同义词扩展,再根据文档句中的基本要素(Basic Elements)对句子进行排序和选择,使用简化的MMR(Maximal Marginal Relevance)技术消除冗余,首次将WordNet应用于面向查询的文摘任务中。为解决同义词扩展引入不相关信息的问题,ABDI等[17]利用WordNet计算了查询词和文档词的语义相似度,发现WordNet局限于有限的词覆盖范围,可以利用其他知识资源(如Wikipedia)及大型语料库解决此问题。

Wikipedia是一个基于超文本系统的网络百科全书[18],其中的概念多使用重定向关系、歧义关系、类关系和内部维基链接,构成层次化的网络结构。在概念的表示上,Wikipedia为每个概念提供了细致且丰富的描述形式。NASTASE[19]将查询文本中的命名实体等词汇与Wikipedia词条页面进行匹配,获取了查询的概念集合,利用维基词条的首段文本中相关概念对查询进行扩充,以提取面向查询的文本摘要。不同于词扩展,MIAO等[20]通过研究句子概念含义,用概念相关度的向量表示句子,向量值为句子内所有概念与Wikipedia内某个概念的相关度。MOHAMED等[21]提出了一种基于增强知识资源的方法用于解决单一知识源覆盖不全的问题。依靠度量短文本的语义相似度,将WordNet与分类变体数据库(CatVar)以及词法链接(Morphosemantic Links)结合,利用Wikipedia丰富WordNet,确定查询词与句子相似性及句子之间的相似程度。陈维政等[22]把图排序引入查询扩展中,抽取文档集合中频繁出现的实体对应的维基词条内容,形成文档集合知识库。利用页面排序算法(PageRank)对文档中的句子进行排序,利用改进的DivRank算法对文档和知识库词条句子进行再次排序。通过线性组合,综合两次排序的结果,最终确定句子的排序,从而选择适当的句子形成摘要。

2.2　利用源文档信息的查询扩展

基于外部语义知识的查询扩展方法存在以下弊端:外部语义知识无法提供与原始文档相关的上下文信息;扩展词有限,受词覆盖范围的限制,不存在于外部语义知识中的单词无法扩展;引入不相关或歧义信息,需要词义消歧,而词义消歧本身就是一个很难完成的任务,会影响最终摘要的性能。为了避免上述局限性,研究机构开始利用源文档信息进行查询扩展。

AMINI等[23]基于EM(Expectation Maximization)算法的变形实现对文档和查询中词项的聚类,依靠此方法实现查询词扩展,再通过分类模型选取摘要句子。ZHAO等[24]运用PageRank算法从原始文档中选择扩展词,综合句子自身的重要性以及句子和词汇间的关系。利用句子之间的关系及句子和词之间的关系寻找信息量大且与查询相关的词进行扩充查询,在引入较少干扰的同时,捕捉到更多有价值的信息。叶娜等[25]采用主题分析技术,识别出当前主题的各个子主题,计算子主题重要度及句子所在的子主题与查询的相关度,依靠计算结果选取摘要句,同时根据词语在子主题之间的共现信息,结合外部语义知识,实现查询扩展。

查询理解技术的优点和缺点如表2所示。

表2 查询理解技术的优点和缺点

Tab.2 The advantages and disadvantages of query understanding technology

方法	优点	缺点
基于外部语义知识的查询扩展	结构灵活;易于理解和使用	无法提供与原始文档相关的上下文信息,扩展词有限;易引入不相关歧义信息
利用源文档信息的查询扩展	更好地理解查询意图,明确查询目标;提高信息查询覆盖面和准确率	质量低的源文档会导致摘要结果不准确

3 文档处理(Document processing)

文档处理是指对原始文档中文本内容进行分析,依据分析情况对句子进行处理。目前对文档处理的方式主要有应用于抽取式文本摘要任务的句子选择技术和应用于生成式文本摘要任务的句子生成技术。其中,抽取式文本摘要主要是从原始文档中抽取单词或句子组成摘要;生成式文本摘要需要对原始文档进行理解,通过自然语言处理算法对其内容进行转述、压缩及同义替换,生成摘要信息。抽取式文本摘要与生成式文本摘要对比情况如表3所示。

表3 抽取式文本摘要与生成式文本摘要对比情况

Tab.3 The comparison between extractive summarization and abstractive summarization

形式	优点	缺点
抽取式文本摘要	连贯性高;可读性强;技术成熟	主题思想理解程度低;在长且复杂的文档中应用存在缺陷
生成式文本摘要	压缩性高;概括性强	技术难度大;可能生成与原文不一致的信息

3.1 句子选择技术

3.1.1 基于文本特征的方法

基于文本特征对句子进行选择,即利用人工智能技术提取句子特征实现句子选择,是最常见的方法之一。

(1)基于聚类的方法。当文档信息不带有标签时,通常采用无监督聚类的方法,根据句子的相似度和权重对其进行选择和排序。SCHILDER等[26]提出了一种面向查询的多文档文本摘要方法FastSum,依据文档集和主题的词频特征,应用最小角回归算法对特征进行详细分析,再利用支持向量机(SVM)输出摘要。YANG等[27]为了解决词向量余弦相似度不适用于短句子的问题,将单词视为独立的文本对象,提出一种噪声检测增强型共聚框架,同时对句子和单词进行聚类,输出摘要。YIN等[28]为减小目标聚类大小,利用高斯混合模型在特征空间上对句子进行聚类,对文档集中的句子进行排序。JAGADEESH等[29]把信息查询技术与摘要技术相结合,使用所有句子中的一组特征对句子进行评分,并以最大分数进行归一化,使用各个特征值的加权线性组合计算句子的最终分数。FEIGENBLAT等[30]通过提取相关性、多样性、长度、位置等特征迭代优化目标函数,计算句子权重,取得了较好的效果。

(2)基于分类的方法。基于分类的方法,通常先对文本对象进行打标签处理,再根据标签信息,将问题转换为二分类问题。应用此方法前需要对数据进行大量标记,不同的标记结果可能导致数据含有大量噪声。

LI等[31]基于贝叶斯主题模型将句子特征融合到主题模型中进行有监督训练,尝试通过在提取的特征中加入句子与查询语句的相似度提升句子的选择效果。AZAR等[32]在训练过程中通过添加噪声改进效果,将查询语句和文章句子一起使用TF-IDF向量编码后加入随机噪声,放入自编码器中进行训练。VALIZADEH等[33]采用多模型融合的方法进行分类效果改进,对于给出的多份人工摘要,考虑到每个人的认知和行为偏好各不相同,在使用标记语料时保留了每个摘要特点,给每个人单独摘要建立一个模型。OUYANG等[34]将回归模型应用到面向查询的多文档摘要任务中,使用SVM评估句子在多文档中的重要性。

3.1.2 基于图结构的方法

基于图结构的方法利用文档结构,将文档表示成一个图模型(节点为文本单元,边用来连接具有关联的节点),从全局角度确定词、句子等文本单元的重要程度,依据节点的连接方式有以下两种方法。

(1)基于传统图的方法。LexRank算法[35]首次将图排序算法引入抽取式文本摘要任务中,使用图结构的方法,综合全文信息,计算句子的权重。LexRank算法的变种Biased LexRank[36-37]是利用马尔科夫模型统计句子转移到查询的加权概率,计算句子的权重,将图排序算法应用到面向查询的文本摘要任务中。BADRINATH等[38]利用先行策略(Look Ahead)寻找与查询相关的句子,并对其相似性进行评分。MOHAMED等[39]通过计算句子和查询的相似性,从文档中选择最合适的句子,并按照句子在文档中出现的时间顺序构建句子和查询的关系图,形成摘要。WAN等[40]通过分析单文档内句子的关系和多文档之间句子的关系,构建跨多文档的句子关系图,采用线性形式、顺序形式和得分组合形式3种不同的融合方案,提出一种多模式图排序算法。WEI等[41]通过计算单文档之间句子相似度和多文档之间句子关联关系,对句子进行排序,充分考虑句子之间和文档之间的相似性,构造了句子层和文档-句子层的两层图。PANDIT等[42]利用离线模型将段落作为节点,依靠TF-IDF算法计算节点间的相似度和节点评分,依据计算结果段落分类,再利用在线模型构建包含查询关键词的子树,计算查询语句与类的相似度以及与节点的相似度,对类和类内节点排序。LI等[43]引入主题信息,基于主题建模技术构造包含句子层和主题层的两层图。随后,CANHASI等[44]通过计算查询词与句子的相似度,构造包含文档、句子、主题的3层图。SAKAMOTO等[45]通过对文档、句子、单词3种异质信息进行融合,用来表示不同语言单元间整体与部分的关系,构建3层图模型。CANHASI[46]构建了句子、查询、段落、文档、框架5层图模型,并通过PageRank算法计算每层信

息和图层间信息的相似度,改善了图模型效果。HU 等[47]通过引入亲和图估算句子之间的相似性,基于局部几何结构和句子内容实现对句子的排序。

(2)基于超图的方法。传统图的一条边只能连接两个节点,无法表示多个句子之间共享的复杂关系,导致大量文档信息损失。超图的一条边可以连接多个节点,应用超图可以简化句子间关系的复杂度,并且利于整合文档全部信息。WANG 等[48]应用基于密度的聚类(Density-Based Spatial Clustering of Applications with Noise, DBSCAN)算法进行聚类,基于聚类结果构建超图:若两个句子间的余弦相似度超过设定阈值或两个句子在同一个类中,添加一条边,并依据节点间的相似度计算句子权重。D'SILVA 等[49]使用改进的 K 均值(K-Means)聚类算法代替了 DBSCAN 算法:根据词 TF-IDF 值、句子间相似度以及与查询语句的相关性,选择距离得分最高的 K 个句子作为中心节点进行聚类,并依据句子之间的相似度和类之间的相似度构建句子关系图模型,并通过图排序算法对模型中的句子进行排序,获得文本摘要。XIONG 等[50]结合主题模型获得主题分布,使用超图获得词与主题、句子与句子的主题分布,应用节点增强和随机游走模型对句子进行排序。ZHENG 等[51]从句子中提取概念,构建概念与句子、概念与查询的二分图,并基于构建超图模型,对句子评分。VAN 等[52]引入超图解决信息冗余或主题覆盖不全的问题,首先,引入一种基于术语语义聚类的新主题模型,以发现语料库中的主题;其次,将这些主题建模为超图中的超边、句子为节点;最后,通过在超图中选择交叉覆盖所有主题的节点生成摘要。

3.1.3 基于神经网络的方法

针对长文档或文档集中句子间复杂的依赖关系,应用深度学习技术中的神经网络方法对其进行分析,成为当前的研究热点。

LIU 等[53]将深度学习各个隐藏层看作表示文本的复杂结构,将网络结构分成内容过滤、结构重组和摘要生成三个部分:通过内容过滤实现对非关键词的过滤,提取重要词汇,选择候选句;通过结构重组对网络结构进行剪枝优化;通过动态规划,选择可作为摘要的句子。CAO 等[54]利用注意力机制实现对相关性和显著性的联合训练:利用卷积神经网络(CNN)学习句子的向量表示,将句子向量加权求和作为文档的向量表示,映射句子和文档到同一向量空间,并在语义层计算其相似度。GAO 等[55]提出了一种协同表示框架,利用当前句子表示、单词内容和主题预测下一个句子的表示,使用句子表示法判断适合作为摘要的句子。JESSE 等[56]通过探索抽取-生成联合模型解决面向查询的文本摘要任务,并结合了迁移学习策略增强模型的性能。

3.2 句子生成技术

句子生成技术通过获取文档或文档集的核心思想,以不同的表达方法生成摘要信息,可以满足多样性文本摘要的需求。

3.2.1 基于图结构的方法

SHAFIEIBAVANI 等[57]提出一种基于图结构的生成式方法。首先,用消歧算法计算文档句子之间以及文档句子和输入查询之间的语义相似性,构建无向图;其次,使用聚类算法对与查询相关的句子进行聚类,在每个类中构建词级 MSC(Multi-Sentence Compression)网络;最后,利用语言模型,考虑词权重、词性和语法结构,生成文本摘要。

3.2.2 基于深度学习的方法

RUSH 等[58]提出了一种基于序列到序列模型(Seq2Seq)的生成式摘要方法,正式将深度学习应用于生成式摘要任务。NEMA 等[10]基于 Seq2Seq 模型,在查询中使用 Attention 的机制获取查询相关的上下文向量,并引入正交变换的方法,解决了生成式摘要当中重复词的问题。KIMURA 等[59]通过实验证明,当输入序列的长度超过 60 时,长短期记忆网络(Long Short-Term Memory, LSTM)实现的编码器模型的准确性会降低。为解决文摘中长文本编码失败的问题,可以引入句子向量表示和原始文档单词向量表示。

3.3 联合训练技术

为解决句子生成技术无法准确地复述原始文档中的事实细节的问题,SEE 等[60]对抽取式模型和生成式模型联合训练,根据选择概率,软性结合 Seq2Seq 模型生成的文本和指针网络抽取的关键信息,既可以生成新文本,又可以复制原文本。

对文档进行处理的技术特点及其局限性如表 4 所示。

表4 ●文档处理技术特点及其局限性

Tab.4 The characteristics and limitations of document processing technology

技术	方法	特点	局限性
句子选择技术	基于聚类的方法	根据句子的相似度和权重,对其进行选择和排序	忽略了文档整体结构对摘要生成的影响
	基于分类的方法	根据文本对象的标签信息,将摘要问题转换为二分类问题	需要对数据进行提前标记,易引入噪声忽略了文档整体结构对摘要生成的影响
	基于传统图的方法	根据全局信息确定词、句子等文本单元的重要程度	无法表示多个句子间的复杂关系,可能存在信息缺失
	基于超图的方法	可以表示多个句子之间共享的复杂关系	计算复杂度较高,概括性较低
	基于神经网络的方法	可以分析长文档或文档集中句子间复杂的依赖关系	计算复杂度高,模型调优难度大
句子生成技术	基于图结构的方法	能够生成具有多样化的文本摘要	过度依赖句子间的相似性
	基于深度学习的方法	生成的摘要更连贯,包含原文的主要信息的可能性更大	不可解释性,生成质量和一致性较低
联合训练技术	—	综合句子选择和句子生成技术准确地复述原始文档中的事实细节	难度大,调优困难

4 信息组织(Information organization)

句子顺序直接影响摘要的可读性。在单文档摘要中,依据句子在原文档中的顺序,即可确定句子顺序。在多文档摘要中,对不同文档中的句子进行排序,需要考虑句子所处上下文

的综合信息[61]。当下,由于对文本摘要中连贯性的研究工作相对缺乏,可以将其分为局部排序和全局排序两类。

4.1 基于相关性的局部排序

局部排序是一种贪婪算法,每次对两个句子进行组织排序。NAYEEM[62]认为实体相同是文本连贯性的标志之一,良好的句子顺序表示所有相邻句子之间具有相似性,基于该假设量化文档连贯性。

BOLLEGALA 等[63]依据时间、概率、主题相关性、前序和后序等 5 种相关性综合判断两个句子的相关性,对两个句子进行排序并得到顺序关系。

4.2 基于图结构的全局排序

与局部排序不同,全局排序方法要求输入所有的句子,基于整体篇章的语义逻辑关系,输出全局最优解。

HE 等[64]通过对句子之间的时间关系、位置关系、主题关系和从属关系进行抽取,构建句子关系图,应用 PageRank 算法确定多文档文本摘要的句子顺序。CHOWDARY 等[65]基于句子的余弦相似度对每个文档构造图,按照文档句子数量进行排序,构建增强集成图,依据句子在增强集成图中的位置进行摘要句子组织和排序。

5 业务场景(Business scenario)

近年来,针对不同的业务场景,研究人员做了很多适配工作。魏鑫炀等[66]根据民事裁判文书的文本结构与其信息分布的特点,对裁判文书重要模块信息文本进行粗粒度抽取,再利用 BERT 的序列标注方法构建细粒度的抽取模型,从句子级别对重要信息进一步抽取,形成最终摘要。ALROSHDI 等[67]利用深度学习算法对电子教育课程的思想、内容和培训目标等进行文本摘要的抽取,帮助学生选择合适的课程来提高成绩。XIAO 等[68]提出了一种称为卷积层次结构注意力网络(CHAN)的方法,以用户查询和长视频为输入,利用编码网络和查询相关性计算,生成满足用户查询偏好的文本视频摘要。

6 结论(Conclusion)

未来,查询文本摘要的发展趋势主要包括以下内容:(1)基于外部语义知识和源文档信息的查询理解技术可以联合使用。利用多个模型联合进行训练,充分发挥各自的优点,提高生成与原文的相关性;(2)基于深度学习的文本分析技术将成为未来研究的热门方向。通过深度学习的方法,可以更准确地表达文本的语义信息,为解决向查询的文本摘要中的技术难题提供新的解决思路;(3)面向查询的文本摘要技术在跨领域中的应用价值将凸显,如面向查询的搜索引擎、个性化信息检索、问答系统,将成为研究热点;(4)针对不同业务场景,面向查询的文本摘要的生成,具有不同的侧重性;(5)针对多文档的摘要生成,目前尚缺乏大规模高质量的训练及评测数据集,因此需要加强数据的构建、增强、共享与评测。生成符合特定需求的文本摘要是面向查询的文本摘要的目标,对于其他方面问题的验证,则是评估模型的重要指标。

参考文献(References)

[1] 曾昭霖. 面向查询的新闻多文档抽取式摘要方法研究[D]. 昆明:昆明理工大学,2023.

[2] 曾昭霖,严馨,徐广义,等. 基于层级 BiGRU+Attention 的面向查询的新闻多文档抽取式摘要方法[J]. 小型微型计算机系统,2023,44(1):185-192.

[3] 何东欢. 查询式文本摘要生成方法研究[D]. 太原:山西大学,2023.

[4] CARBONELL J. Automated query-relevant summarization and diversity-based reranking[J]. IJCAL-97,AI and digital libraries,1997:9-14.

[5] CARBONELL J,GOLDSTEIN J. The use of MMR,diversity-based reranking for reordering documents and producing summaries[C]//CROFT W B,MOFFAT A,VAN RIJSBERGEN C J,et al. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM,1998:335-336.

[6] GAMBHIR M,GUPTA V. Recent automatic text summarization techniques:a survey[J]. Artificial intelligence review,2017,47:1-66.

[7] LIN C Y. Rouge:a package for automatic evaluation of summaries[C]//ACL. Proceedings of the Workshop on Text Summarization Branches out. Stroudsburg:ACL,2004:74-81.

[8] MISHRA A,BERBERICH K. Event digest:a holistic view on past events[C]//PEREGO R,SEBASTIANI F. Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM,2016:493-502.

[9] YANG Y,YIH W T,MEEK C. WikiQA:a challenge dataset for open-domain question answering[C]//MÀRQUEZ L,CALLISON-BURCH C,SU J. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Stroudsburg:ACL,2015:2013-2018.

[10] NEMA P,KHAPRA M,LAHA A,et al. Diversity driven attention model for query-based abstractive summarization[DB/OL]. (2017-04-26)[2018-07-13]. arXiv preprint arXiv:1704.08300.

[11] Hermann. CNN/Dailymai[EB/OL]. (2015-06-10)[2015-11-19]. https://cs.nyu.edu/~kcho/DMQA/.

[12] KRISHNA K,SRINIVASAN B V. Generating topic-oriented summaries using neural attention[C]//WALKER M

A,JI H,STENT A. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies, Volume 1 (Long Papers). Stroudsburg: ACL, 2018: 1697-1705.

[13] Vox Mediaz. KDD VoxMedia articles[EB/OL]. (2017-03-21)[2020-06-16]. https://data.world/elenadata/vox-articles.

[14] PASUNURU R,CELIKYILMAZ A,GALLEY M,et al. Data augmentation for abstractive query-focused multi-document summarization[C]//AAAI. Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI,2021,35(15):13666-13674.

[15] ZHU H C,DONG L,WEI F R,et al. Transforming wikipedia into augmented data for query-focused summarization[J]. IEEE/ACM transactions on audio, speech, and language processing,2022,30:2357-2367.

[16] ZHOU L,LIN C Y,HOVY E. A be-based Multi-document summarizer with query interpretation[EB/OL]. (2005-10-10)[2005-10-10]. https://duc.nist.gov/pubs/2005papers/usc-isi-zhou1.pdf.

[17] ABDI A, IDRIS N, ALGULIYEV R M, et al. Query-based multi-documents summarization using linguistic knowledge and content word expansion[J]. Soft computing, 2017,21:1785-1801.

[18] 康扬,李梦琳,王晓光. 维基百科词条语义结构研究[J]. 信息资源管理学报,2017,7(3):88-96.

[19] NASTASE V. Topic-driven multi-document summarization with encyclopedic knowledge and spreading activation [C]//LAPATA M, NG H T. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Stroudsburg:ACL,2008:763-772.

[20] MIAO Y J,LI C P. Enhancing query-oriented summarization based on sentence wikification[C]//ACM. Workshop of the 33rd Annual International. New York:ACM,2010: 32-35.

[21] MOHAMED M A,OUSSALAH M. Similarity-based query-focused multi-document summarization using crowdsourced and manually-built lexical-semantic resources[C]//IEEE. Proceedings of the IEEE:2015 IEEE Trustcom/BigDataSE/ISPA. Piscataway:IEEE,2015,2:80-87.

[22] 陈维政,严睿,闫宏飞,等. 利用维基百科实体增强基于图的多文档摘要[J]. 中文信息学报,2016,30(2):153-159.

[23] AMINI M R,USUNIER N. A contextual query expansion approach by term clustering for robust text summarization[EB/OL]. (2007-04-27)[2007-04-27]. https://duc.nist.gov/pubs/2007papers/lip6.pdf.

[24] ZHAO L,WU L D,HUANG X J. Using query expansion in graph-based approach for query-focused multi-document summarization[J]. Information processing & management,2009,45(1):35-41.

[25] 叶娜,蔡东风. 一种面向查询的多文档摘要方法[J]. 中文信息学报,2010,24(6):69-74.

[26] SCHILDER F,KONDADADI R. FastSum:fast and accurate query-based multi-document summarization[C]// ACL. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. Stroudsburg: ACL, 2008:205-208.

[27] YANG L B,CAI X Y. Semi-supervised co-clustering for query-oriented theme-based summarization[J]. Research journal of applied sciences, engineering and technology, 2012,4(18):3410-3414.

[28] YIN W P, PEI Y L, ZHANG F, et al. Query-focused multi-document summarization based on query-sensitive feature space[C]//ACM. Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York:ACM,2012:1652-1656.

[29] JAGADEESH J, PINGALI P, VARMA V. Capturing sentence prior for query-based multi-document summarization[C]//ACM. Large Scale Semantic Access to Content (Text,Image,Video,and Sound). New York:ACM, 2007:798-809.

[30] FEIGENBLAT G,ROITMAN H,BONI O,et al. Unsupervised query-focused multi-document summarization using the cross entropy method[C]//ACM. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM,2017:961-964.

[31] LI J W,LI S J. A novel feature-based Bayesian model for query focused multi-document summarization[J]. Transactions of the association for computational linguistics, 2013,1:89-98.

[32] AZAR M Y, SIRTS K, MOLLA D, et al. Query-based single document summarization using an ensemble noisy auto-encoder[C]//ACL. Proceedings of the Australasian Language Technology Association Workshop 2015. Stroudsburg:ACL,2015:2-10.

[33] VALIZADEH M, BRAZDIL P. Exploring actor-object relationships for query-focused multi-document summarization[J]. Soft computing, 2015, 19: 3109-3121.

[34] OUYANG Y, LI W J, LI S J, et al. Applying regression models to query-focused multi-document summarization[J]. Information processing & management, 2011, 47(2): 227-237.

[35] ERKAN G, RADEV D R. LexRank: graph-based lexical centrality as salience in text summarization[J]. Journal of artificial intelligence research, 2004, 22: 457-479.

[36] OTTERBACHER J, ERKAN G, RADEV D R. Using random walks for question-focused sentence retrieval[C]//ACL. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2005: 915-922.

[37] OTTERBACHER J, ERKAN G, RADEV D R. Biased LexRank: passage retrieval using random walks with question-based priors[J]. Information processing & management, 2009, 45(1): 42-54.

[38] BADRINATH R, VENKATASUBRAMANIYAN S, VENI MADHAVAN C E. Improving query focused summarization using look-ahead strategy[C]//CLOUGH P, FOLEY C, GURRIN C, et al. Advances in Information Retrieval: 33rd European Conference on IR Research, ECIR 2011. Berlin: Springer, 2011: 641-652.

[39] MOHAMED A A, RAJASEKARAN S. Improving query-based summarization using document graphs[C]//IEEE. Proceedings of the IEEE: 2006 IEEE International Symposium on Signal Processing and Information Technology. Piscataway: IEEE, 2006: 408-410.

[40] WAN X J, XIAO J G. Graph-based multi-modality learning for topic-focused multi-document summarization[C]//KITANO H. Twenty-First International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann, 2009: 1587-1591.

[41] WEI F R, LI W J, HE Y X. Document-aware graph models for query-oriented multi-document summarization[M]// LIN W S, TAO D C, KACPRZYK J, et al. Multimedia analysis, processing and communications. Berlin: Springer, 2011: 655-678.

[42] PANDIT S R, POTEY M A. A query specific graph based approach to multi-document text summarization: simultaneous cluster and sentence ranking[C]//IEEE. Proceedings of the IEEE: 2013 International Conference on Machine Intelligence and Research Advancement. Piscataway: IEEE, 2013: 213-217.

[43] LI Y R, LI S J. Query-focused multi-document summarization: combining a topic model with graph-based semi-supervised learning[C]//TSUJII J, HAJIC J. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Stroudsburg: ACL, 2014: 1197-1207.

[44] CANHASI E, KONONENKO I. Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization[J]. Expert systems with applications, 2014, 41(2): 535-543.

[45] SAKAMOTO K, SHIBUKI H, MORI T, et al. Fusion of heterogeneous information in graph-based ranking for query-biased summarization[C]// ACM. International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2015: 19-22.

[46] CANHASI E. Query focused multi-document summarization based on five-layered graph and universal paraphrastic embeddings[C]// SILHAVY R, SENKERIK R, KOMINKOVA OPLATKOVA Z, et al. Artificial Intelligence Trends in Intelligent Systems: Proceedings of the 6th Computer Science On-line Conference 2017. Cham: Springer, 2017: 220-228.

[47] HU P, HE J C, ZHANG Y. Graph-based query-focused multi-document summarization using improved affinity graph[C]//ZHANG S M, WIRSING M, ZHANG Z L. Knowledge Science, Engineering and Management: 8th International Conference, KSEM 2015. Cham: Springer, 2015: 336-347.

[48] WANG W, WEI F R, LI W J, et al. HyperSum: hypergraph based semi-supervised sentence ranking for query-oriented summarization[C]// ACM. Proceedings of the 18th ACM conference on information and knowledge management. New York: ACM, 2009: 1855-1858.

[49] D'SILVA S, JOSHI N, RAO S, et al. Improved algorithms for document classification & query-based multi-document summarization[J]. International journal of engineering and technology, 2011, 3(4): 404-409.

[50] XIONG S F, JI D H. Query-focused multi-document summarization using hypergraph-based ranking[J]. Information processing & management, 2016, 52(4): 670-681.

[51] ZHENG H T, GUO J M, JIANG Y, et al. Query-focused multi-document summarization based on concept impor-

tance[C]//BAILEY J,KHAN L,WASHIO T,et al. Advances in Knowledge Discovery and Data Mining:20th Pacific-Asia Conference, PAKDD 2016. Cham: Springer, 2016:443-453.

[52] VAN LIERDE H,CHOW T W S. Query-oriented text summarization based on hypergraph transversals[J]. Information processing & management, 2019, 56(4): 1317-1338.

[53] LIU Y,ZHONG S H,LI W J. Query-oriented multi-document summarization via unsupervised deep learning[C]// AAAI. Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park:AAAI,2012,26(1):1699-1705.

[54] CAO Z,LI W,LI S,et al. Attsum:Joint learning of focusing and summarization with neural attention[DB/OL]. (2016-04-01)[2016-09-27]. arXiv preprint arXiv:1604.00125.

[55] GAO Y,WEI L J,HUANG H Y,et al. Topical sentence embedding for query focused document summarization[C]// KANAGASABAI R, MORSHED A, PUROHIT H J. Proceedings of the IJCAI Workshop on Semantic Machine Learning(SML 2017). San Francisco: Morgan Kaufmann,2017:21-26.

[56] JESSE V,FABBRI A R,KRYSCINSKI W,et al. Exploring neural models for query-focused summarization[DB/OL]. (2021-12-14)[2022-04-26]. arXiv preprint arXiv:2112.07637.

[57] SHAFIEIBAVANI E,EBRAHIMI M,WONG R,et al. A query-based summarization service from multiple news sources[C]//IEEE. Proceedings of the IEEE:2016 IEEE International Conference on Services Computing. Piscataway:IEEE,2016:42-49.

[58] RUSH A M,CHOPRA S,WESTON J. A neural attention model for abstractive sentence summarization[DB/OL]. (2015-09-02)[2015-09-03]. arXiv preprint arXiv:1509.00685.

[59] KIMURA T, TAGAMI R, MIYAMORI H. Query-focused summarization enhanced with sentence attention mechanism[C]//IEEE. Proceedings of the IEEE: 2019 IEEE International Conference on Big Data and Smart Computing (BigComp). Piscataway:IEEE,2019:1-8.

[60] SEE A,LIU P J,MANNING C. Get to the point:Summarization with pointer-generator networks[DB/OL]. (2017-04-14)[2017-04-25]. arXiv preprint arXiv:1704.04368.

[61] BARZILAY R, ELHADAD N. Inferring strategies for sentence ordering in multidocument news summarization[J]. Journal of artificial intelligence research, 2002, 17: 35-55.

[62] NAYEEM M T. Methods of sentence extraction,abstraction and ordering for automatic text summarization[D]. Lethbridge:University of Lethbridge (Canada),2017.

[63] BOLLEGALA D,OKAZAKI N,ISHIZUKA M. A machine learning approach to sentence ordering for multidocument summarization and its evaluation[C]//DALE R,WONG K F,SU J,et al. Natural Language Processing-IJCNLP 2005: Second International Joint Conference. Berlin:Springer, 2005:624-635.

[64] HE Y,LIU D X,YANG H,et al. A hybrid sentence ordering strategy in multi-document summarization[C]// ABERER K,PENG Z Y,RUNDENSTEINER E A,et al. Web Information Systems - WISE 2006:7th International Conference on Web Information Systems Engineering. Berlin:Springer,2006:339-349.

[65] CHOWDARY C R,SREENIVASA KUMAR P. Sentence ordering for coherent multi-document summary generation[C]//GRAY A,JEFFERY K,SHAO J H. Sharing Data,Information and Knowledge:25th British National Conference on Databases,BNCOD 25. Berlin: Springer, 2008:40-50.

[66] 魏鑫炀,唐向红. 基于 BERT 的抽取式裁判文书摘要生成方法研究[J]. 软件工程,2022,25(5):1-4.

[67] ALROSHDI Y,ALBADAWI M,ALHAMADANI A,et al. An extractive summarization for utilizing learning content using deep learning algorithm:proposed framework and implementation[J]. International journal of computing and digital systems,2023,13(1):461-474.

[68] XIAO S W,ZHAO Z,ZHANG Z J,et al. Convolutional hierarchical attention network for query-focused video summarization[C]//AAAI. Proceedings of the AAAI conference on artificial intelligence. Palo Alto:AAAI,2020, 34(7):12426-12433.

作者简介:

徐　睿(1990-),男,博士,高级工程师。研究领域:信息内容安全,网络安全。

刘　金(1990-),女,硕士,工程师。研究领域:工控安全,网络安全。

亚　静(1990-),女,博士,AI 架构师。研究领域:自然语言处理,信息内容安全。本文通信作者。