计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2021.20190785
58(1): 1-21, 2021

自动文本摘要研究综述

李金鹏[1,2] 张 闯[1] 陈小军[1] 胡 玥[1,2] 廖鹏程[1,2]
[1](中国科学院信息工程研究所 北京 100093)
[2](中国科学院大学网络空间安全学院 北京 100040)
(lijinpeng@iie.ac.cn)

Survey on Automatic Text Summarization

Li Jinpeng[1,2], Zhang Chuang[1], Chen Xiaojun[1], Hu Yue[1,2], and Liao Pengcheng[1,2]
[1](*Institute of Information Engineering*, *Chinese Academy of Sciences*, *Beijing* 100093)
[2](*School of Cyber Security*, *University of Chinese Academy of Sciences*, *Beijing* 100040)

Abstract In recent years, the rapid development of Internet technology has greatly facilitated the daily life of human, and it is inevitable that massive information erupts in a blowout. How to quickly and effectively obtain the required information on the Internet is an urgent problem. The automatic text summarization technology can effectively alleviate this problem. As one of the most important fields in natural language processing and artificial intelligence, it can automatically produce a concise and coherent summary from a long text or text set through computer, in which the summary should accurately reflect the central themes of source text. In this paper, we expound the connotation of automatic summarization, review the development of automatic text summarization technique and introduce two main techniques in detail: extractive and abstractive summarization, including feature scoring, classification method, linear programming, submodular function, graph ranking, sequence labeling, heuristic algorithm, deep learning, etc. We also analyze the datasets and evaluation metrics that are commonly used in automatic summarization. Finally, the challenges ahead and the future trends of research and application have been predicted.

Key words automatic text summarization; extractive; abstractive; deep learning; ROUGE metric

摘 要 近年来,互联网技术的蓬勃发展极大地便利了人类的日常生活,不可避免的是互联网中的信息呈井喷式爆发,如何从中快速有效地获取所需信息显得极为重要.自动文本摘要技术的出现可以有效缓解该问题,其作为自然语言处理和人工智能领域的重要研究内容之一,利用计算机自动地从长文本或文本集合中提炼出一段能准确反映源文中心内容的简洁连贯的短文.探讨自动文本摘要任务的内涵,回顾和分析了自动文本摘要技术的发展,针对目前主要的2种摘要产生形式(抽取式和生成式)的具体工作进行了详细介绍,包括特征评分、分类算法、线性规划、次模函数、图排序、序列标注、启发式算法、深度学习等算法.并对自动文本摘要常用的数据集以及评价指标进行了分析,最后对其面临的挑战和未来的研究趋势、应用等进行了预测.

关键词 自动文本摘要;抽取式方法;生成式方法;深度学习;ROUGE指标

中图法分类号 TP391

收稿日期:2019-11-08;**修回日期**:2020-04-16

基金项目:国家自然科学基金项目(61602474)

This work was supported by the National Natural Science Foundation of China (61602474).

通信作者:张闯(zhangchuang@iie.ac.cn)

21世纪互联网快速发展,文本数据呈指数级增长,用户如何快速有效地从海量信息中提炼出所需的有用资料,已经成为一个亟待解决的问题.自动文本摘要(automatic summarization)技术,又被称为自动文摘,它的出现恰逢其时,为用户提供简洁而不丢失原意的信息,可以有效地降低用户的信息负担、提高用户的信息获取速度,将用户从繁琐、冗余的信息中解脱出来,节省了大量的人力物力,在信息检索、舆情分析、内容审查等领域具有较高的研究价值.

早期的文本摘要普遍是通过人工来完成的,文本数据量的激增使得这项工作日渐繁重且效率低下,逐渐不能满足用户的需求.近年来,随着对非结构化文本数据研究的进展,自动文摘任务得到了广泛的关注和研究,其已成为自然语言处理领域的研究热点之一.学术界涌现出大量围绕算法技术、数据集、评价指标和系统的相关工作,这些工作在一定程度上取得了较好的效果,快速应用到金融、新闻、医学、媒体等各个领域,如社交媒体摘要[1]、新闻摘要[2]、专利摘要[3]、观点摘要[4]以及学术文献摘要[5].尽管如此,目前计算机自动产生的摘要还远不能达到人工摘要的质量,在该任务上还有很大的提升空间,仍需要相关研究者进一步探索有效的自动文摘技术.

目前已有一些文献对自动文摘任务进行了调研和评估.在早期的工作中,万小军等人[6]首次将自动文摘的研究工作从内容表示、权重计算、内容选择、内容组织4个角度进行了深度剖析,并对发展趋势进行了展望,为之后的研究工作打下了良好的基础.王俊丽等人[7]则主要针对抽取式自动文摘的图排序算法进行了介绍.曹洋等人[8]重点分析了3种主要的机器学习算法在自动文摘中的应用.此外,还有一些相关的研究工作,但他们基本仅针对自动文摘中的单个技术方向进行详细综述,经过调研发现目前尚缺乏对自动文摘任务进行全面的研究综述.

基于此,为了便于研究者在现有研究工作的基础上取得更好的进展,非常有必要对目前自动文摘的研究成果进行全面的分析和总结.因此,我们查阅整理了近年来学术界相关的研究工作,包括自然语言处理、人工智能等相关领域的国际会议和学术期刊,对这些研究成果按照摘要产生的技术算法进行了详细的分类以及优缺点的对比与总结.除此之外,本文对自动文本摘要研究常用的数据集、评价方法进行归纳总结,最后对自动文摘任务未来的研究趋势进行展望与总结.

1 自动文本摘要问题定义

大量的文本数据涌现导致用户很难快速获取文本中的主题信息,所以需要通过技术手段将文本提炼成不丢失原意的摘要.摘要,维基百科对其的定义是指简洁准确记述重要内容,正确无误地摘录出来,不作主观解释和评论,使读者于最短的时间内以掌握内容,得知原著的大意;剑桥英文词典解释摘要是"A short, clear description that gives the main facts or ideas about something".自动文摘是利用计算机通过算法自动地将文本或文本集合转换成简短摘要,帮助用户通过摘要全面准确了解原始文献的中心内容.

美国IBM公司的Luhn[9]于1958年首次设计了一个自动文摘系统,拉开了该课题研究的序幕.自动文摘形式化定义为:设$D=\{w_1,w_2,\cdots,w_n\}$为包含n个单词的原始文档,自动文摘的目标是得到一个由单词y_i组成的包含原文中心内容的摘要$Y=\{y_1,y_2,\cdots,y_m\}$,需满足$n\gg m$.自动文摘根据不同的标准有不同的分类划分,按照是否提供上下文环境,可以分为面向查询的自动文摘和普通自动文摘;按照不同的用途,可以分为指示性文摘和报道性文摘等;按照文档数量,可以分为单文档自动文摘和多文档自动文摘;按照产生方法可以分为抽取式自动文摘和生成式自动文摘.单文档自动文摘和多文档自动文摘主要区别在于处理的文档数量,多文档存在的冗余信息较多[10].但这2种任务都需要对原文内容进行权重计算、排序组织、整理等,因此多文档自动文摘技术可看作单文档自动文摘技术的扩展.本文主要依据自动文摘产生方法(抽取式和生成式)的算法技术、数据集、评价指标对相关的研究工作进行综述.

2 自动文本摘要技术和方法

20世纪90年代以来,随着互联网的快速发展,自动文摘的应用价值越来越广,引起了越来越多的学者关注,深度学习的热潮更是为自动文摘的研究带来了新的机遇.目前,自动文摘实现方法主要分为抽取式方法和生成式方法.前者是从原始文档中提取关键文本单元来组成摘要,文本单元包括但不限于单词、短语、句子等.这种方法产生的摘要通常会保留源文章的显著信息,有着正确的语法,但不可避免

的是容易产生大量的冗余信息,且对于短文本摘要不太友好.后者是根据对输入原始文本的理解来形成摘要,模型试图去理解文本的内容,可以生成原文中没有的单词,更加接近摘要的本质,具有生成高质量摘要的潜力.自动文摘的研究工作的技术框架为:内容表示→权重计算→内容选择→内容组织[6].

内容表示是将原始文本划分为文本单元的过程,主要是分字、词、句等预处理工作.另有一些研究工作使用主题模型、图、语义表示的方法对原文进行深层次的表示,针对深度学习方法而言,需要将文本单元映射成由实数构成的向量,即词嵌入(word embedding)工作.权重计算则是要对文本单元计算相应的权重评分,权重的计算方式多样,如基于特征评分、序列标注、分类模型等提取内容特征计算权重.内容选择是对经过计算权重后的文本单元选择相应的文本单元子集组成摘要候选集,可根据要求的摘要长度、线性规划、次模函数、启发式算法等选择文本单元.内容组织是指对候选集的内容进行整理形成最终摘要,可根据字数要求按顺序输出,也有研究者提出使用基于语义信息、模板和深度学习的方法来产生符合要求的摘要.

目前主流的自动文摘技术方法的对比见表1.该技术方法也可根据是否有监督分为无监督学习方法(特征评分、图排序、主题模型等)和监督学习方法(分类算法、序列标注、深度学习等).前者不需要训练数据和人工参与,速度较快、效率较高,在缺乏高质量数据集的情况下取得了不错的效果,但无法避免的是应用场景简单,不能满足用户对高质量摘要的需求;而后者在自动文摘任务上得到了较快的发展并取得了突破性的进展.广义来看,抽取式方法将自动文摘简单地看作是二元分类问题,判断文档中的文本单元是否属于摘要内容,该类方法产生的摘要往往不够简洁,存在冗余文本,连贯性上也无法得到很好的保证;生成式方法则是对训练数据的文本-摘要数据对的学习,包括语言结构、词法、语法等,根据不同的算法生成摘要.不足之处是需要利用大量

Tabel 1 The Classification of Automatic Text Summarization Technology

表1 自动文摘技术分类

方法	技术	描述	缺点	类型
抽取式方法	主题模型	使用语义信息、主题模型,挖掘词句隐藏信息抽取重要句子	效果依赖于数据集质量和领域等情况	无监督学习
	基于图	句子作为顶点,2个句子的相似度作为边的权重,根据顶点的权重分数来确定关键词句	只依赖于句子相似度,计算量大、运算相对较慢	无监督学习
	特征评分	根据词频、句子位置或句子与首句相似度等来选择关键词句构成摘要	需要手工设置权重,质量低下	无监督学习
	序列标注	以句子为单位,利用HMM, CRF等进行序列标注,抽取句子组成摘要	特征复杂,执行速度慢	监督学习
	分类算法	利用SVM、贝叶斯、CNN、LSTM等分类模型来判别句子是否隶属于摘要	分类算法将句子看作独立的,忽略了句子之间的联系	监督学习
	启发式算法	利用遗传算法、蚁群算法等提取最优句子组成摘要	运算复杂,参数设置和迭代停止条件等依赖经验,但是却相当重要	强化学习
	线性规划	把自动摘要问题看成带约束的优化问题基于线性规划进行求解,文本单元以句子为主	求解时存在维数灾,计算复杂性上一般为NP-难问题	其他
	次模函数	把自动文摘问题当做一个预算约束下的次模函数最大化问题	求解过程在实际中较慢,如何设计最适合任务模型的次模函数较难统一	其他
	深度学习	利用CNN, RNN, LSTM等神经网络模型进行句子抽取	对数据要求较高,参数量较多,易出现梯度消失或爆炸等影响效果	监督学习
生成式方法	基于图	词作为顶点,2个词的相似度作为边的权重,根据顶点的权重分数来选择最优路径	只依赖于单词的相似度,计算量大、运算速度慢	无监督学习
	线性规划	基于线性规划求解,文本单元以词和短语为主	求解时存在维数灾,由于文本单元粒度变小导致参数量增加,计算复杂性更高	其他
	语义表示	将原文表示为深层语义形式,计算深层语义子图,由其生成摘要	仍属于探索阶段,其效果还不尽如人意	无监督学习
	模板	通过观察人工摘要总结模板,填充模板框架转换为摘要	摘要的语言千篇一律,过于呆板	无监督学习
	深度学习	利用神经网络模型进行文本理解,端到端生成摘要	对数据要求较高,参数量较多且训练较慢,易出现梯度消失或爆炸等影响效果	监督学习

训练数据训练模型,训练数据的质量决定了模型性能的峰值,并且训练过程普遍耗时较长,部分重要的模型参数需要人工设置、优化.相关研究者在生成式方法上做出了大量的创新工作,取得了显著成绩.下面我们将具体介绍这些自动文摘算法的技术以及研究成果.

2.1 抽取式方法

抽取式方法主要考虑摘要的相关性和句子的冗余度2个指标[11-12].相关性衡量摘要所用的句子是否能够代表原文的意思,冗余度是用来评估候选句子包含冗余信息的多少.大多数现有的抽取式摘要系统使用句子作为提取的基本单位,因为它们是可以表达为语句的最小语法单位[13].该方法通常面临2个难题:一方面是如何对划分的文本单元进行排序;另一方面是如何选择排序后的文本单元[14].

2.1.1 基于主题模型的方法

自然语言处理最需要解决的任务之一是如何使计算机可以真正地理解文本.因此涌现出一些基于主题模型的方法,如潜在语义分析(latent semantic analysis, LSA)[15]、隐狄利克雷分布模型(latent Dirichlet allocation, LDA)[16]等来挖掘词句隐藏信息,该类方法的效果依赖训练数据质量和领域等情况.

LSA是一种数据模型,核心思想是将词和文章映射到矢量语义空间,通过降维去除部分噪声,在低维空间中提取文档中词的概念.不足之处是它虽然可以解决一义多词(synonymy)问题,但对于一词多义(polysemy)问题还不能很好地处理.LSA的处理流程为:

1) 分析文档集并建立词汇-文本矩阵;

2) 对词汇-文本矩阵进行奇异值分解(singular value decomposition, SVD);

3) 对SVD分解后的矩阵进行降维;

4) 使用降维后的矩阵构建潜在语义空间.

Gong等人[17]第1次提出使用LSA用于自动文摘任务,文档 D 由 m 个词和 n 个句子组成,构建句子矩阵 $\boldsymbol{A}=(\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_n)$,每个列向量 $\boldsymbol{A}_i$ 代表文档中句子 i 加权的词频(term-frequency)向量,那么该文档可以表示为 $m\times n$ 的矩阵 $\boldsymbol{A}$,然后利用SVD分解该矩阵:

$$\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}},$$

其中,$\boldsymbol{U}$ 是矩阵 $\boldsymbol{A}$ 的特征向量组成的 $m\times n$ 矩阵,$\boldsymbol{U}$ 中的每个特征向量被称为 $\boldsymbol{A}$ 的左奇异向量;$\boldsymbol{\Sigma}$ 是 $n\times n$ 的对角矩阵,对角元素是降序的非负奇异值;$\boldsymbol{V}$ 是 $n\times n$ 的正交矩阵,$\boldsymbol{V}$ 中的每个特征向量被称为 $\boldsymbol{A}$ 的右奇异向量.然后从每个右奇异向量矩阵中选择排名最高的句子组成摘要.Steinberger等人[18]利用指代消解提升基于LSA的自动文摘系统的性能,他们使用指代消解系统GUITAR[19]解析表达式,发现当添加词典信息作为SVD的输入时会取得较好的效果.

LDA主题模型的主要思想是通过对文字建模发现隐含的主题,其是由Blei等人[16]在pLSA[20]的基础上进行了扩展,pLSA参数过多时会导致过拟合问题,在此基础上LDA加入了超参数,并使用Dirichlet分布作为文档-主题和词-主题的先验分布.LDA实现过程如图1所示,首先从Dirichlet分布 α 中采样生成文档-主题分布 θ_m,在主题分布中生成第 m 篇文档的第 n 个词的主题 $Z_{m,n}$,在Dirichlet词-主题分布 β 中采样生成主题 $Z_{m,n}$ 对应的词分布 ϕ_k,然后从词分布中得到词 $W_{m,n}$.Kar等人[21]提出了一种在任何用户定义的时间段内利用LDA模型发现隐含主题结构变化的方法,在动态文本集合中生成摘要,在当时取得了优于基线的效果.

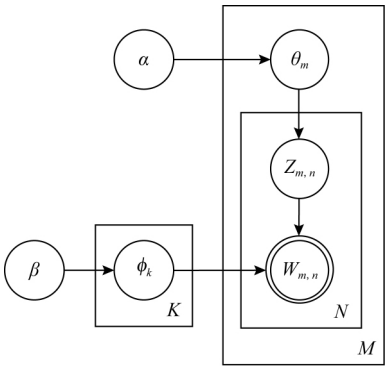


Fig. 1 LDA topic model

图1 LDA主题模型

2.1.2 基于图的方法

基于图的方法是通过全局信息确定文本单元(单词、句子),将文本单元构成图的顶点,2个相似的点用边连接起来,将文本构建成拓扑结构图,利用图排序算法TextRank或LexRank等对包含文本自身的结构信息的词句进行排序.该方法只依赖于句子相似度,由于存在任意句子相似性计算和迭代计算,所以会导致运行速度相对比较慢,也无法避免选出的句子之间具有极高的相似度.

TextRank算法基于PageRank. Mihalcea等人[22]介绍了通过TextRank抽取文本中重要度较高的句子形成文本摘要,主要步骤有5个:

1) 将输入的文本分割成句子并建立有向加权图 $G=(V,E)$,由点集合 V 和边集合 E 组成,其中 $E\in V\times V$.

2) 图 G 中节点 V_i,V_j 之间边的权重为 ω_{ji},权重的计算基于 2 个句子 S_i,S_j 之间的相似度:

$$sim(S_i,S_j)=\frac{|\{w_k|w_k\in S_i \ \&\ w_k\in S_j\}|}{\mathrm{lb}(|S_i|)+\mathrm{lb}(|S_j|)},$$

w_k 表示句子中的单词,如果 S_i,S_j 之间的相似度大于给定的阈值,则认为 2 个句子语义相关,并将其连接起来,边的权重为

$$\omega_{ji}=sim(S_i,S_j).$$

3) 对顶点 V_i 计算得分,$In(V_i)$ 为指向该点的点集合,$Out(V_i)$ 为点 V_i 指向的点集合,d 为阻尼系数,取值范围为 0～1,代表从图中某一特定点指向其他任意点的概率,一般取值为 0.85[23],对图中的节点指定任意的初值,并递归计算直到收敛:

$$Score(V_i)=1-d+d\sum_{V_j\in In(V_i)}\frac{\omega_{ji}}{\sum_{V_k\in Out(V_j)}\omega_{jk}}Score(V_j).$$

4) 根据 V_i 的得分进行排序,抽取重要度最高的 T 个句子形成候选集合.

5) 根据字数或句子数量要求,从候选集合中抽取句子组成文摘.

TextRank 不需要训练数据,只利用单篇文章本身的信息即可实现自动文摘,节省大量计算资源.它属于无监督算法,因其简洁有效、速度快等优点而得到广泛应用.2004 年,密西根大学的 Erkan 等人[24]提出了一种与 TextRank 类似的图排序算法 LexRank 用于多文档摘要,他们认为文档集中与很多句子相似的句子被认为是该文档集的主题中心.与 TextRank 不同的是,LexRank 是一个无向无权图.首先对文档集分句后的结果利用余弦相似度计算相似度,当 2 个句子之间的相似度超过给定的阈值,代表这 2 个句子语义相关,将它们代表的节点连接起来.每个节点的度是指与其相连的边的数量,度越大代表该句子包含的信息越重要.为了避免将每条边同等对待,需要考虑节点的权威性,如果一个节点的度较大,那么认为与其相连的句子相应地也比较重要.然后根据句子间的连接矩阵迭代计算句子所包含的信息量,进行排序并根据需求选择句子组成文摘.

Leiva[25]利用 TextRank 算法应用在网页的响应式文本摘要上,网页设计人员可以在各种设备上为广泛的用户创建自定义阅读解决方案.Fang 等人[26]提出了 CoRank 的单词-句子共同排序模型,它将单词-句子关系与基于图的无监督排序模型相结合.从矩阵运算的角度来看,CoRank 理论上可以保证其收敛性.Parveen 等人[27]针对学术论文的摘要任务提出由句子和实体节点组成的二分图来表示输入文档,基于 HITS 图排序算法对句子进行排名,在 DUC-2002 数据上取得了当时最先进的结果.

2.1.3 基于特征评分的方法

在研究的早期,大部分研究工作通过分析原文的特征来提取摘要,特征包括词频、首句与标题相似度,以及句子长度、句子中心性等因素,常见的评分特征见表 2,通过对特征评分来判断文本单元是否属于摘要.这种方法简单、速度快,但效果容易受到异常数据影响生成与主题无关的摘要,且存在内容不全面、语句冗余、不连贯等问题.Luhn[9]的工作就是使用词频特征来解决自动文摘任务,他认为文章的信息都应包含在句子中,该任务的目标是找出那些包含信息最多的句子来组成摘要.

Ferreira 等人[28]分析了 15 种句子评分算法(针对词:词频、TF-IDF、大写字母、专有名次、词共现、词汇相似性;针对句子:提示语、包含数字的句子、句子长度、句子位置、句子中心性、句子与标题相似性;图排序:TextRank、Bushy 路径、集合相似性)对抽取文本摘要进行定量和定性的评估.Wang 等人[29]提出了 9 种启发式方法(冗余句子删除法、基于完整摘要的句子评分、基于具有不同单词的完整摘要的句子评分、基于摘要句子的句子评分、基于具有不同单词的摘要句子的句子评分、基于具有不同单词的反向摘要句子的句子评分等)为抽取式摘要来构造近似理想的抽取和上界,用 6 种评分方法(词频、标题词、句子长度、句子位置、Bushy 路径和 TextRank)和 5 种不同的语料库来证明所提出方法的有效性.Oliveira 等人[30]分析了 18 种评分方法(集合相似性、Bushy 路径、提示语、词汇相似性、命名实体、动名词短语、数字数据、公开关系、专有名词、句子中心性、句子长度、句子位置、句子与标题的相似性、词频-逆句子频率指数、TextRank、大写、词共现、词频)和 4 种组合策略(平均组合、加权平均组合、基于投票的组合、Condorcet 排名)对单文档和多文档自动摘要性能的影响,发现语料库的特征会影响所研究的技术和组合的性能,给出了技术和方法组合等进行自动文摘句子选择的建议.

Table 2 The Features Related of Score

表 2 评分相关特征

评分对象	特征名称	计算公式
单词级别	词频	$TF(w,d)=f(w,d)/d$.其中,$f(w,d)$表示词 w 在文档中出现的次数,d 表示文档中包含的单词数量.
	逆文档频率	$IDF(w,D)=\mathrm{lb}(D/df(w,d))$.其中,$df(w,d)$表示文档集 D 中包含 w 的文档数,即文档频率.
	词频-逆文档频率	$TF\text{-}IDF(w,d,D)=TF(w,d)\times IDF(w,D)$.
	大写字母	$UpperCase(S_i)=upp(S_i)/\lvert S_i\rvert$.其中,$upp(S_i)$表示 S_i 中首字母大写单词的数量,$\lvert S_i\rvert$表示 S_i 中所有单词的数量.
	专有名词	$WordPro(S_i)=pro(S_i)/\lvert S_i\rvert$.其中,$pro(S_i)$表示 S_i 中专有名词的数量,$\lvert S_i\rvert$表示 S_i 中所有单词的数量.
	词共现	$WordCoo(S_i)=sum(sim(S_i,S_j))$,$i,j=1,2,\cdots,n$.其中,$sim()$表示 2 个句子之间的 n-gram 相似度.
句子级别	提示短语	$SentCue(S_i)=Cue(S_i)/cue$.其中,$Cue(S_i)$表示句子 S_i 中提示短语数量,cue 表示文档 d 中提示短语数量.
	动名词短语	$SentNV(S_i)=NV(S_i)/nv$.其中,$NV(S_i)$表示句子 S_i 中动名词短语数量,nv 表示单句中最大动名词短语数.
	命名实体	$SentNer(S_i)=Ner(S_i)/e$.其中,$Ner(S_i)$表示句子 S_i 中命名实体数量,e 表示单句中最大命名实体数量.
	长度	$SentLen(S_i)=Len(S_i)/\max(S)$.其中,$Len(S_i)$表示句子 S_i 的长度,$\max(S)$表示最大句子的长度.
	位置	$SentPos(S_i)=1-i/S_n$.其中,i 表示句子 S_i 的位置,S_n 表示文档中句子的数量.
	中心性	$SentCen(S_i)=(w_i\cap w_o)/(w_i\cup w_o)$.其中,$w_i$ 表示句子 S_i 中关键词的数量,w_o 表示其他句子中关键词的数量.
	与标题相似性	$SentSim(S_i)=sim(S_i)/T$.其中,$sim(S_i)$表示句子 S_i 与标题单词相同的数量,T 表示标题中所有单词的数量.

2.1.4 基于序列标注的方法

对抽取式自动文摘而言,以前大多数的监督学习都将任务视为二分类问题,每个句子相互独立,没有利用句子之间的联系.无监督学习使用一些启发式的规则来提取有信息量的句子.因此结合上面 2 种方法的优势,可以将自动文摘看成一个序列标注问题,如统计概率图方法利用朴素贝叶斯(naive Bayesian, NB)、隐马尔可夫模型(hidden Markov model, HMM)或者条件随机场(conditional random field, CRF)来抽取文本组成摘要.该方法将自动文摘问题看作序列标注问题,原文是句子的序列,序列标注问题就是将原文序列打上 0,1 的标签.标签为 1 代表为文本的摘要,反之为 0,该方法需要质量较高的数据,执行速度较慢.

贝叶斯网络是使用有向图表示变量之间的依赖关系,朴素贝叶斯是特殊的贝叶斯网络,其假设特征之间相互独立,这与在自动文摘任务中假设摘要的句子之间相互独立的特点相符合[31-32].马尔可夫模型是一种简单的动态贝叶斯网络,在马尔可夫模型中状态不可见,并且当前状态只依赖于前一时刻的状态,并且满足观测独立性假设.HMM 是对 NB 的改进,因为在 NB 中的独立性假设不符合实际情况.在文摘摘要任务中将是否为摘要的标注视为 HMM 中的状态是不可见的,观察变量为文本的一些特征,如文本的句子、句子的位置等[33].HMM 在一定程度上解决了特征独立性的问题,但是在特征空间很大甚至特征之间有重叠的情况下,HMM 的观察独立性条件就不再满足;且用一个联合随机变量模型来解决给定观测序列的判别问题也是不太合适的.因此提出条件随机场来解决以上的问题.条件随机场这里专指 CRF 线性链,在 CRF 中特征可随意组合,不需要特征独立性假设,解决了文本上下文相关的问题.CRF 还是判别式模型,更适合序列标注问题.在 CRF 中可以使用一些简单的特征,如单词或者句子的位置、长度信息、和附近句子的相似程度,或一些更加复杂的特征如隐藏主题特征、句子的打分信息等[34].将 CRF 应用在自动文摘的任务上在各种特征和训练数据下的实验结果都优于上述 2 种结果.

2.1.5 基于分类的方法

分类方法利用 SVM、贝叶斯等分类模型判断句子是否属于摘要,该方法的效果同样依赖训练数据质量和领域等情况.Louis[35] 在贝叶斯惊奇(Bayesian surprise)模型的基础上结合背景知识来形成摘要,并基于此方法在通用摘要和更新摘要任务上进行了实验.贝叶斯惊奇由 Itti 等人[36] 提出,用于量化在输入新的数据(新闻报道)前后,用户背景知识不同假设的概率分布之间的差异.在 Louis 等人的模型中,H 是编码背景知识的所有假设的集合空间,每个假设 $h\in H$ 采用多项分布的形式表示.$P(h)$是基于背景语料库中的信息计算得出的每个假设的先验概率,符合狄利克雷分布(Dirichlet distribution).背景语料库的词汇量大小为 V,$w_1,w_2,\cdots,w_v$ 表示其中的单词,$P(h)=Dir(\alpha_1,\alpha_2,\cdots,\alpha_v)$,其中 $\alpha_i(1\leqslant i\leqslant v)$ 是狄利克雷分布的浓度参数,I 表示新输入的文档

中的文本单元,I 中单词的频数表示为 $c_1,c_2,\cdots,c_v$,则 h 的后验概率为

$$P(h\mid I)=Dir(\alpha_1+c_1,\alpha_2+c_2,\cdots,\alpha_V+c_V),$$

I 在假设空间 H 上创建的惊奇 $S(I,H)$ 表示假设的先验分布和后验分布之间的差异,使用 KL-散度进行计算:

$$S(I,H)=K_{\mathrm{KL}}(P(h\mid I),P(h))=\int_H P(h\mid I)\ \mathrm{lb}\frac{P(h\mid I)}{P(h)}.$$

该算法的主要步骤为:

1) 单词评分.为输入的文档中的每一种单词类型计算 1 个分数.设单词 w_i 在输入 I 中出现了 c_i 次,则 $P(h\mid w_i)=Dir(\alpha_1,\alpha_2,\cdots,\alpha_i+c_i,\cdots,\alpha_V).w_i$ 的分数由 $P(h\mid\omega_i)$ 和 $P(h)$ 之间的 KL-散度计算得到.

2) 句子评分.根据单词分数的平均值和总和的组合函数为句子打分.

3) 选择句子.利用贪心算法选择高分句子,为避免冗余,在选择某个句子之后,将该句子中单词的分数设置为 0,重新计算剩余句子的得分,重复上述选择过程,直到摘要达到长度约束.

Abdi 等人[37]对 4 种特征(信息增益、增益比、对称不确定性、Relief-F[38])选择技术和 7 种著名的分类方法(决策树、朴素贝叶斯、支持向量机、k-最近邻、随机森林、逻辑回归、人工神经网络)进行了性能研究.其中特征选择是减少原始特征集并移除不相关特征的过程,对于分类过程至关重要,消除不相关、噪声、冗余、无价值的特征可以提高分类准确度并改善分类的运行时间,减小特征空间的大小,提高分类方法的质量.实验结果表明,将基于支持向量机的情感分类方法与信息增益作为特征选择技术相结合,在总结评论中表达的观点时性能最好.

2.1.6 基于启发式算法

启发式算法是相对于最优化算法提出的基于直观或经验构造的算法,通常在可接受的时间和空间花费下给出待解决组合优化问题每个实例的一个可行解,该可行解与最优解的偏离程度一般不能被预计.现阶段,启发式算法以仿自然体算法为主,在自动文摘领域,主要利用遗传算法、蚁群算法等将文本摘要问题形式化表示为优化问题,提取最优句子形成摘要.该方法运算复杂,参数设置和迭代停止条件等相当重要,但是却只能依赖经验调整.

Sanchez-Gomez 等人[39]针对多文档摘要任务首次设计并实现了多目标人工蜂群优化算法.该算法主要有 2 个流程.

1) 初始化,随机生成种群规模为 n 的雇佣蜂,每个雇佣蜂代表一个从原始文档集中随机抽取句子形成的摘要,即一种解决方案.

2) 在设定的最大循环次数 K 之间重复执行以下步骤:

① 发送雇佣蜂.利用突变机制(在摘要中添加或删除句子)形成新的摘要,如果突变后的摘要能够支配(在改进某些目标的同时不会使其他目标恶化)原摘要,则使用突变后的摘要,否则保留原摘要.

② 利用定义的排名(rank)和拥挤度(crowding)模块确定最佳摘要.前者根据主导关系对不同帕累托前沿(Pareto fronts)的解决方案进行排序;后者根据解决方案的拥挤距离(crowding distance)评估密度指标,倾向于更多样化的解决方案.基于这 2 种操作计算每个摘要可能被选择的概率,更好的摘要将被分配更高的概率.

③ 发送跟随蜂.跟随蜂根据上一步计算得出的概率选择 1 只雇佣蜂,即选择 1 个摘要,选择完成后,与发送雇佣蜂阶段类似,利用突变机制在新旧摘要间选择更优的摘要.

④ 发送侦察蜂.侦察蜂验证耗尽(在预设次数的突变之后效果没有改进)的解决方案,并以随机方式生成新的摘要,取代与该解决方案相关联的雇佣蜂或跟随蜂.同时,侦察蜂应进行一定数量的突变,从而能够有机会与现有的解决方案竞争.突变的规模与当前的循环次数成正比,即循环次数越多,现有的解决方案应该越好,因此需要更多的突变.

⑤ 将当前的种群规模缩小至原始规模 n,再次利用排名和拥挤模块选择最佳摘要,如果生成的摘要不符合预先设定的长度约束,则对此摘要进行修复(删除影响摘要质量的句子),然后进行下一次循环.

Mosa 等人[40]在短文本摘要(short text summarization, STS)领域里,对于社交平台上的评论提取摘要,能够使用户在不阅读整个评论列表的前提下获取评论简报.算法以混合蚁群优化(ant colony optimization, ACO)为基础,采用局部搜索机制(local search, LS),即 ACO-LS-STS,以产生最优或接近最优的摘要.首先使用图着色算法缩小解的范围,然后将不同的评论组合在一起标记上相同的颜色,同时保留原评论列表中信息的比例,利用 ACO-LS-STS 算法,以并行形式从每种颜色中提取最具交互性的评论,最后从最佳颜色中选择最佳摘要.Peyrard 等人[41]将自动金字塔(automatic pyramid)作为遗传算法的适应度函数,提出了自动生成训练数据的

方法,并在此基础上提出了新的监督框架,该框架学习自动评估金字塔分数,并将其应用于基于优化的多文档摘要的提取中.Litvak 等人[42]基于多种单文档摘要方法的变体开发了多语言提取和压缩(MUSEEC)的摘要工具,其中 MUSE 方法是基于遗传算法的监督摘要生成器,该方法对文档中的句子进行排序并提取排名靠前的句子组成摘要.

2.1.7 基于线性规划的方法

基于线性规划的方法将自动文摘任务看作是基于 0-1 二值变量的求解全局最优解的问题[43-47].整数线性规划(integer linear programming, ILP)在计算复杂性上一般为 NP-难问题,求解过程在实际应用中会表现较慢,并不适合实时性较高的应用场景,需要采用一些技巧解决这个问题.

早先,研究人员使用较为简单的去除冗余机制最大边缘相关法(maximal marginal relevance, MMR)[11]选择合适的内容组成摘要,后来 McDonald[46]针对多文档摘要提出用全局最优方法替代 MMR,其中一种方法是将多文档摘要问题表示为整数线性规划问题,采用高效的分支界定算法解决 NP-难问题:

$$\begin{aligned}&\max\ \sum_i a_i Rel(i)-\sum_{i<j} a_{ij} Red(i,j).\\&\text{s.t. 1)}\ a_i,a_{ij}\in\{0,1\};\\&\quad 2)\ \sum_i a_i l(i)\leqslant K;\\&\quad 3)\ a_{ij}-a_i\leqslant 0;\\&\quad 4)\ a_{ij}-a_j\leqslant 0;\\&\quad 5)\ a_i+a_j-a_{ij}\leqslant 1.\end{aligned}$$

其中,a_i,a_j 和 a_{ij} 称为指示变量,当文本单元 i 或者文本单元对 i 和 j 在摘要中时值为 1.ILP 的目标是通过设置这些指示变量的值,在保证解是有效的前提下满足约束条件并最大化回报,$Rel(i)$是它的相关性,$Red(i,j)$是句子 i 与句子 j 的冗余度.约束 1)表明指示变量是二值的,约束 2)是摘要中句子的长度之和必须小于我们预先设定的最大值,约束 3)~5)保证解是有效的,约束 3)4)简单地表明若摘要中包含文本单元对 i 和 j,则 i 和 j 也应被单独包含在其中,约束 5)刚好与之相反. McDonald 从 ROUGE 值和可扩展性 2 方面对贪心算法、整数线性规划、基于背包问题解决方案的动态算法进行了对比,整数线性规划的方法取得了比较高的 ROUGE 分数,但基于背包问题的动态规划算法比其有更好的扩展性.

为了提升 ILP 的扩展性,2009 年 Gillick 等人[47]提出了基于 ILP 的可扩展全局模型,它在子句(sub-sentence)或者说概念级(concept-level)上操作,假设概念是独立的,其可以是单词、命名实体、语法子树、语义关系.该工作可更有效地扩展到更大的问题是因为它不需要二次变量处理冗余项,公式为:

$$\begin{aligned}&\max\ \sum_i \omega_i a_i.\\&\text{s.t. 1)}\ \sum_j l_j S_j\leqslant L;\\&\quad 2)\ s_j\ Occ_{ij}\leqslant a_i;\\&\quad 3)\ \sum_j S_j Occ_{ij}\geqslant a_i;\\&\quad 4)\ a_i\in\{0,1\};\\&\quad 5)\ S_j\in\{0,1\}.\end{aligned}$$

其中,a_i 和 Occ_{ij} 为指示变量,a_i 指示概念 i 是否存在于摘要中,其权重为 ω_i.Occ_{ij} 则指示概念 i 是否存在句子 j 中.约束 1)保证了摘要的长度,约束 2)3)确保了求解的逻辑一致性,选择某个句子就要选择其包含的所有概念,约束 2)同时也阻止选择概念少的句子.除此之外,Boudin 等人[48]通过使用近似算法来消除 NP-难问题以及由于剪枝带来的多个最优解问题,取得了理想的效果.

2.1.8 基于次模函数的方法

随着自动文摘技术研究的发展,研究人员根据贪心选择目标函数都具有次模性的特点使用次模函数来处理自动文摘任务.次模函数(submodular function)具有次模性,是边际效益递减(property of diminishing returns)现象的形式化描述.对于一个函数 $f(\cdot)$来说,若 $A\subseteq B\subseteq V$,那么对于 $\forall e\in V-B$ 都满足:

$$f(A\cup\{e\})-f(A)\geqslant f(B\cup\{e\})-f(B),$$

若它还满足 $f(A)\geqslant f(B)$则称它是单调函数.

Lin 和 Bilmes[49]是最早将次模函数引入自动文摘的研究者之一,他们提出将自动文摘定义为预算约束(budget constraint)下次模函数最大化问题,即每个文本单元都有一个预算.在此基础上,Lin 等人[50]设计了一类适用于抽取式自动文摘任务的次模函数.这些函数由 2 部分组成:第 1 部分用于鼓励摘要包含更多的信息;第 2 部分用于鼓励内容的多样性,降低冗余度.这些函数是单调不减的,这意味一个高效可伸缩的贪婪最优化方案具有常数因子最优性保证.Wu 等人[51]使用次模函数的方法解决特定领域问题,从大量有关灾害管理的新闻和报告中抽取简明扼要的摘要报告,以帮助专家分析灾难的趋势,实验表明他们的方法是具有竞争力的.虽然将次模函数应用于自动文摘任务取得了一定的效果,

但是到目前为止如何设计最适合任务模型的次模函数仍然没有一个统一的标准.

仍有一些工作虽产生了新句子,但次模函数的作用仍然是用来做句子抽取.Chali 等人[52]定义了3个单调的次模函数,即重要性、覆盖率和非冗余度,目标函数是次模函数的线性组合,将产生摘要的过程形式化表示为在长度约束下将目标函数最大化的问题,通过次模函数对压缩后的句子进行抽取.该方法首先对多文档中主语相同但动词短语不同的句子进行合并,然后通过依存树对句子进行压缩,生成更加简明且信息量更大的新的摘要候选句,从该句子集合中选择最佳句子使目标函数最大化,最后使用贪心算法获得近似最优的摘要.Bairi 等人[53]基于预先给定的层次性 DAG 主题结构,从中选择规模更小但信息量更大的主题子集用于生成原始文档集合的摘要.通过引入一系列单调的次模函数(如主题的覆盖范围、相似性、特异性、清晰度、相关性和一致性)衡量主题的适用性,目标函数是上述次模函数的凸组合,在预测框架下优化目标函数中各个次模函数的权重系数,最后通过贪心算法对目标函数优化,得到一组能够对原始文档集合进行分类概括的主题子集.

2.1.9 基于深度学习的方法

深度学习方法利用受限玻尔兹曼机(restricted Boltzmann machine, RBM)、卷积神经网络(convolutional neural network, CNN)、循环神经网络(recurrent neural network, RNN)等神经网络模型对原文建模得到文本单元表示后进行文本单元的抽取形成摘要.

Liu 等人[54]基于 RBM 提出了面向查询的多文档摘要的深度学习模型.该模型分为3个部分,分别是面向观点的提取、重构验证和摘要生成.第1部分使用贪心的分层提取算法;第2部分最小化重构信息损失获得全局最优参数;第3部分根据第2部分获得的参数使用动态规划算法获得最后满足长度的摘要.Cao 等人[55]提出不需要手动提取特征,利用 CNN 对文本进行分类的方法,然后通过文档的表示和文档的类别来生成不同类型的摘要.Yin 等人[56]先利用 CNN 语言模型训练出句子的表示,然后利用 PageRank 算法算出句子的重要程度,迭代地选出重要的句子. Singh 等人[57]提出利用同文档内容相关/无关的特征来更好地表示文档,从而提取出信息量更大的句子.模型分为3个部分:第1部分是 CSTI 利用 CNN 来获得句子本身的特征,利用 Bi-LSTM Tree Indexer 来获取和文档无关的句子语义和组成的特征;第2部分是 Extractor,利用一些简单的文档相关的简单特征(如句子的位置、在文档中出现的频率)来表示句子;第3部分是 Regression,将前2部分的句子的表示连接并且回归得到句子的打分. Cheng 等人[58]提出一种提取式的文本自动摘要模型,模型框架分为2个大的子结构:一部分是对文档的读取,相当于传统的编码器-解码器框架中的编码器部分,区别在于句子级别的编码使用卷积神经网络;另一部分是提取器,相当于编码器-解码器框架中的解码器.同时由于文本自身就是分层架构的,所以网络也设计为分层架构,读取器先从单词到句子进行编码,然后对句子到文档进行编码,提取器先从文档提取合适的句子,再从句子中提取合适的单词.Nallapati 等人[59]将抽取式摘要看作是序列分类问题,采用 GRU 作为基本序列分类器的基本模块,取得了比较不错的效果.另外这篇工作基于 CNN/Daily Mail 数据集利用无监督学习构造抽取式摘要的数据集.Chen 等人[60]通过观察人类生成摘要时对文档阅读及理解多遍的事实,提出了基于交互式文本摘要技术的抽取式摘要生成模型.考虑到当前摘要生成技术局限于对待生成摘要文本只处理1遍,多数文本表达无法得到全局最优的结果.针对这种情况,采用通过不断迭代来更新相应文本及优化相应的文本表征,使用所有迭代的输出表示来为原文中句子集合打标,抽取相关句组成摘要,取得了不错的效果.

2.2 生成式方法

生成式方法属于自然语言处理的文本生成领域,它产生的摘要不是来自原文中的句子拼接,而是利用生成技术通过对原文语义的理解后生成的.目前自然语言的理解和生成是比较困难和复杂的,因此生成式摘要尚需要富有建设性的创新和大量的工作来提升性能.在生成式自动文摘方法中也存在一些同抽取式方法类似的工作,例如基于线性规划、基于图等的方法.他们的核心思路是相同的,区别在于在生成式任务中不再是简单的为文本单元打分、排序,而是对其进行改进更适合自动文摘生成任务.本节将具体介绍生成式自动文摘的算法.

2.2.1 基于图的方法

Mehdad 等人[61]针对基于图排序的生成式方法提出基于图排序算法的最佳路径排名策略,该方法在根据查询短语对原文进行句子抽取的基础上,利用词汇相似性对选择的句子进行聚类,在每类句子集合中构造以单词为结点的有向图,并用有向边连接

相邻的单词.在摘要生成阶段,从构建的单词图中选择所有至少包含一个动词的路径,根据流畅性、查询短语的相关性和整体内容定义排序函数来选择最佳路径,作为每个原始句子集合中生成的摘要句,组成最终摘要.

2.2.2 基于线性规划的方法

Banerjee 等人[62]首先从多文档集合中识别出最重要的文档,该文档中的每个句子都被初始化为一个单独的聚类,然后将其他文档中的句子分别聚合到与其相似性最高的聚类中.在摘要生成阶段,针对每个聚类生成一个单词图(word graph)结构,并从图的起始结点到结束结点之间构造路径,然后采用整数线性规划(ILP)模型,将信息量和语言质量结合在一个优化框架中组成目标函数,同时在 ILP 模型中加入约束条件:确保每个聚类只生成 1 个句子;避免使用来自不同聚类的具有相同或相似信息的冗余句子.将上述构造的路径表示为二元变量,其值表示该路径是否包含在生成的摘要中,从路径集合中选择最佳句子来最大化目标函数,使得生成的摘要包含的信息内容最多、可读性最强.该优化问题的解所包含的路径集合即为原始多文档集合的摘要.Durrett 等人[63]将句子中的词组作为基本单位对文档进行细粒度文本单元的提取,采用整数线性规划方法,在长度约束下,根据在训练数据上学习的模型参数选择文本单元使目标函数最大化,由上述文本单元组成摘要.同时,基于句法和修辞理论结构(rhetorical structure theory, RST)对句子进行压缩,保证摘要的语法性.针对摘要中代词指代不明的问题加入回指约束,利用加入先行词或用指代的短语替换代词的方法进行指代消解,保证摘要的连贯性.

2.2.3 基于语义的方法

Liu 等人[64]提出了基于语义信息生成摘要模型,如图 2 所示.首次利用抽象语义表示(abstract meaning representation, AMR)将源文本解析为一组 AMR 图,将图转换为摘要图,然后从摘要图生成文本.随后,Takase 等人[65]将 AMR 信息纳入标准编码器-解码器以改善结果,这些方法与提取式方法相比是有竞争力的,但它们在摘要生成中仍远未达到人类水平的质量.这些方法的问题是无法保证它们处理语言细节的程度,例如具有否定全文含义的单词或共同引用的单词等.Dohare 等人[66]在文献[65]的基础上开发了基于共指消解和元节点的方法生成故事 AMR,取得了优于基线的效果.

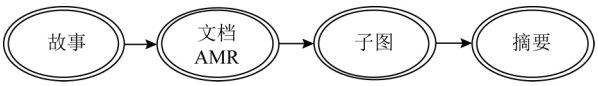


Fig. 2 The pipeline proposed by Liu et al.[64]

图 2 Liu 等人提出的方法步骤[64]

Li[67]提出从文本中提取语义信息来生成多文档摘要的方法,构建基本语义单元上的语义链接网络以捕获文本的语义信息.基本语义单元是描述事件或动作的语义,摘要由语义链接网络生成的句子构成.

2.2.4 基于模板的方法

基于模板的方法将原文中的关键内容填充到提前定义好的模板,一般来说模板是个不完整的句子.Zhou 等人[68]首次使用全局选择的标题短语填充到预先指定的标题模板中生成标题.Oya 等人[69]通过调整字图算法从人工编写的摘要中生成模板,进而通过会议记录自动生成摘要,他们创建了包含 2 个组件的框架:一个离线模板生成模块,从人工编写的摘要中创建模板;另一个是在线生成摘要模块,根据主题对会议记录分段并从中提取重要短语,填充到适当的模板中生成摘要.

此前,序列到序列的自动文摘方法只依赖原文本来产生摘要.Cao 等人[70]受到基于模板的自动文摘方法的启发,将已有摘要作为软模板(soft templates)来指导文本摘要的生成.如图 3 所示,该方法由 3 个模块组成:1) Retrieve.利用常用的信息检索平台 Lucene 从训练语料库中找出候选模板,然后应用递归神经网络(RNN)编码器将输入语句和每个候选模板转换为隐藏状态.2) Rerank.根据隐藏状态与输入句子的相关性来衡量一个候选模板的信息量,具有最高预测信息量的候选模板被视为实际的软模板.

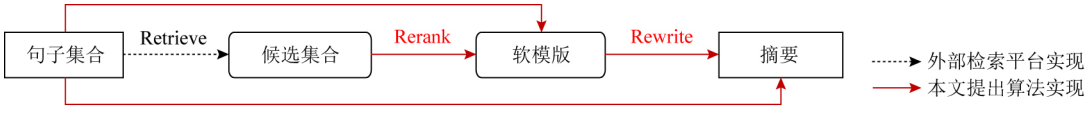


Fig. 3 Flow chat of the proposed method by Cao et al.[70]

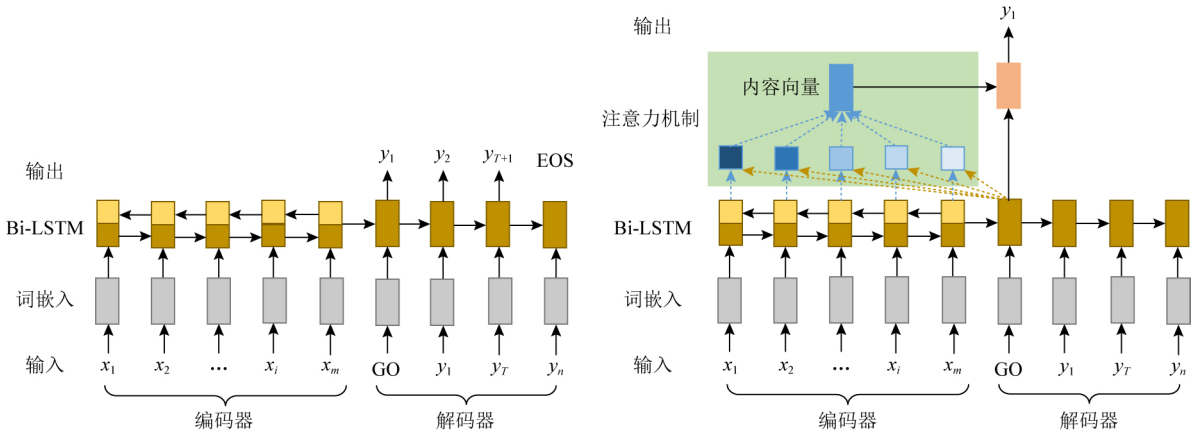
图 3 Cao 等人提出方法的流程[70]

3)Rewrite.根据句子和模板的隐藏状态生成摘要.软模板方法具有很强的竞争力,高质量外部摘要的导入提高了生成摘要的稳定性和可读性.

由于模板是人工编写的,因此生成的摘要通常是流畅并包含信息的.但模板的构建非常耗时,并且需要大量的领域知识,生成的语言千篇一律,显得呆板.而且不可能为各种领域的摘要开发所有模板.目前在金融领域上应用较多,例如股票市场的报价形式较统一,对实时性要求较高,因此基于模板生成摘要是一个不错的选择.

2.2.5 基于深度学习的方法

近年来随着深度学习在图像、文本处理等领域的发展,尤其是基于深度学习的机器翻译模型在多种语言和评价指标上超过了传统的算法模型,因此也涌现出越来越多基于深度学习的自动文摘生成式方法[71-95],目前最为流行的是基于序列到序列(sequence-to-sequence, Seq2Seq)框架的模型,如图 4 所示,因其可以避免繁琐的人工特征提取,也避开了权重计算、内容选择等模块,只需要足够的输入、输出即可开始训练模型.相关研究者提出了许多有趣的技术来改进 Seq2Seq 模型,提升模型的性能.在本文中,基于深度学习的生成式方法主要关注基于 Seq2Seq 展开的工作,图 5 展示了该框架下生成式自动文摘研究工作的经典发展历程.



(a) Seq2Seq模型 (b) 基于注意力机制的Seq2Seq模型

Fig. 4 Two models

图 4 两种模型

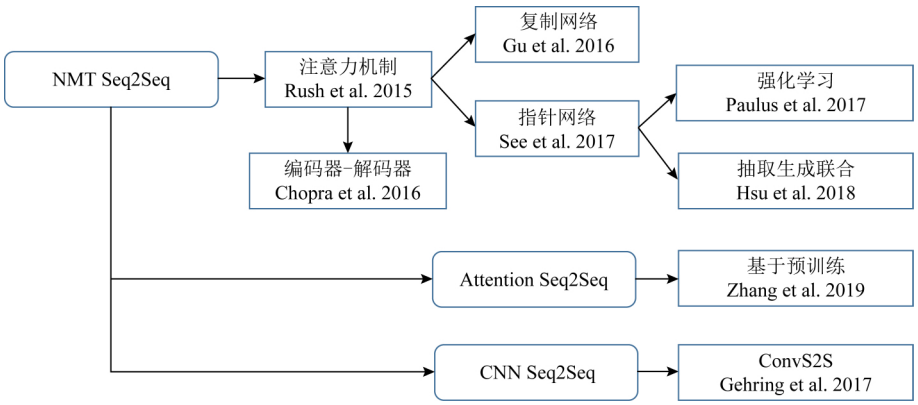


Fig. 5 Classical development of abstractive summarization based on deep learning Seq2Seq model

图 5 基于深度学习 Seq2Seq 模型的生成式文本摘要经典发展历程

2.2.5.1 基于 RNN 结构

Rush 等人[71]受到神经机器翻译(neural machine translation, NMT)[72]研究的启发,首次提出基于注意力(attention)机制的编码器、神经网络语言模型(neural network language model, NNLM)解码器的模型用于生成式摘要任务,与传统的方法相比,性能取得了显著的提升.随后,Chopra 等人[73]对其进行了扩展,基于循环神经网络构造解码器,在 Gigaword 数据集上效果优于其他先进的模型.之后的很多工作都以此为基线模型,Nallapati 等人[74]为了解决

生成式摘要容易遇到3个关键问题,在RNN编码器-解码器的架构上引入一些新技术:

1)在编码器加入丰富的文本特征捕获关键词;

2)加入生成器指针来解决词典外词汇(out-of-vocabulary, OOV)和低频词的问题;

3)利用层级注意力机制来捕获不同级别文档结构信息.

除此之外,他们还在阅读理解数据集CNN/Daily Mail的基础[75]上构建了英文长文本摘要数据集,在此之前大量的工作都是在英文短文本摘要数据集Gigaword上进行的.这些模型不能简单地直接应用到长文本摘要上,最明显的问题是他们没办法有效地解决OOV问题.Gu等人[76]提出复制网络(copy network, CopyNet)结构,CopyNet有2个主要的优势:一是通过复制有效保留源文中的重要信息;二是输出端可以生成一些和源文措辞不同的摘要,因此也可以把CopyNet认为是抽取式和生成式摘要的结合.然而CopyNet有一个比较大的局限是它原封不动地复制输入端的信息,不能灵活调整.同一时期,See等人[77]提出了指针生成器网络用于摘要生成,它可以通过指针自动地选择是从原文中复制摘要所需要的单词还是由词表生成新的单词,有效地缓解了该问题.值得一提的是See等人还提出使用覆盖(coverage)机制[78]处理解码端生成摘要过程中的重复问题,这项工作取得了当时最先进的结果.

尽管之前的研究已经取得了不错的结果,但Seq2Seq模型仍存在曝光偏差(exposure bias)和训练与评估不匹配的问题,前者是说在训练时使用Teacher-Forcing[79]的方式,即解码端上一时刻输入的单词是来自训练集的正确目标单词,但在测试时的输入是模型生成的单词,这会导致误差的积累,使得随着序列长度的增加而生成越来越差的摘要.后者指模型在训练阶段使用交叉熵损失优化模型,评价模型时常使用不可微分的ROUGE和BLEU等指标进行评价.Paulus等人[80]首先提出使用强化学习来应对自动文摘中的这个问题,他们应用自批评(self-critical)策略梯度算法[81]训练模型,提出了一种混合目标函数,它将强化学习损失与传统的交叉熵损失相结合.因此,他们的方法既可以利用不可微分的评价指标,又可以提高可读性.

Cao等人[82]为避免模型生成的摘要中存在不符事实的信息,通过使用开放的信息抽取和依存分析技术从源文中提取实际的事实描述,还提出Dual-Attention序列到序列的框架使得模型必须以原文本和提取的事实描述为条件来生成摘要.实验结果证明他们的方法可以减少80%的虚假事实出现.Hsu等人[83]提出了一种抽取式与生成式相结合的方式,先利用抽取模块对句子的重要程度打分,在该基础上使用生成模块更新对原始文章中每个单词的注意力权值,然后逐词生成得到该文的摘要.Zhou等人[84]在编码器加入Selective门控网络,将词的隐层状态与句子的隐层状态拼接到一起,输入到前馈网络里生成新的语义向量.Li等人[85]借鉴应用在图像领域的VAE(variational auto-encoder)[86],将句子潜在的结构信息融入到生成摘要模型中,进而提高模型生成摘要的质量.Jiang等人[87]认为Seq2Seq模型应具有强大的编码器,它可以从输入的文本中提取和记忆重要信息,他们通过增加一个不需要注意力机制和指针网络的Closed-book解码器来提高指针生成器模型编码器的记忆能力.这样的解码器迫使编码器在其存储状态下编码的信息更具选择性,因为解码器不能依赖注意力和复制模块提供的额外信息,因此改进了整个模型.Gehrmann等人[88]发现现有模型在内容选择上表现不佳,提出通过内容选择器来过度确定源文档中应成为摘要一部分的短语.他们使用此选择器作为Bottom-up attention步骤,将模型约束为可能的短语.实验表明,这种方法提高了压缩文本的能力,同时仍能生成流畅的摘要.Lin等人[89]针对Seq2Seq模型生成的摘要经常会存在重复或者无语义的问题,提出了基于源文本上下文的全局信息的Global Encoding框架,负责控制编码器到解码器的信息流.

2.2.5.2 基于其他结构

此前主流的Seq2Seq模型的编码器和解码器主要使用的是循环神经网络、长短期记忆网络(LSTM)和门控循环单元(GRU).但基于RNN结构的解码器和编码器因为具有顺序依赖性,不可避免的问题是不能并行计算,长序列需要大量的计算资源,导致在训练过程中训练时间和难度会随着序列长度的增加而不断提升.

针对这个问题, Vaswani等人[90]提出一种新型的Seq2Seq网络结构Transformer,只依赖前馈网络和注意力机制实现Seq2Seq架构.该模型可以并行计算,并且在提升机器翻译性能的同时也可加快训练速度.Zhang等人[91]将预训练语言模型Bert与Transformer结构相结合提出2阶段解码模型,其在CNN/Daily Mail数据集上取得了领先的效果.

Gehring等人[92]则提出完全使用卷积神经网络

来构成 Seq2Seq 模型(ConvS2S)用于机器翻译任务,超越了谷歌创造的基于 LSTM 机器翻译的效果.除此之外,ConvS2S 在自动文摘任务上也取得了不错的效果.基于卷积神经网络的序列到序列模型结构可以准确地控制上下文的长度,有效地处理句子的结构信息,同时可以并行计算提高效率.Fan 等人[93]将 ConvS2S 模型进一步应用于生成式文本摘要,可以关注用户的个人风格来生成摘要,包括摘要长度、行文风格、用词等,并在 CNN/Daily Mail 数据集上取得了优于指针生成网络的结果.Wang 等人[94]提出将 ConvS2S 模型结合主题信息并使用强化学习优化摘要任务中的 ROUGE 分数,取得了理想的效果.Transformer 和 ConvS2S 的出现为自动文摘的发展提供了新的技术路线.

我们将不同的基于深度学习的模型在各个数据集上的 ROUGE 分数展示在了表 3 和表 4 中.

Table 3 ROUGE Scores of Different Models on the English Dataset

表 3 英文数据集上不同模型的 ROUGE 分数

模型	Gigaword			Non-ano CNN/Daily Mail		
	RG-1	RG-2	RG-L	RG-1	RG-2	RG-L
ABS	29.55	11.32	26.42			
ABS+[71]	29.76	11.88	26.96			
RAS-Emlan[73]	33.78	15.97	31.15			
words-lvt2k[74]	32.67	15.59	30.64	32.49	11.84	29.47
pointer-generator+coverage[77]				36.44	15.66	33.42
				39.53	17.28	36.38
ML+RL[80]				39.87	15.82	36.90
SEASS[84]	36.15	17.54	33.63			
FTSum[82]	37.27	17.65	34.24			
ConvS2S[93]				39.75	17.29	36.54
End2end w/inconsistency loss[83]				40.68	17.97	37.13
Seq2Seq+CGU[89]	36.30	18.00	33.38			
Reinforced-Topic-ConvS2S[94]	36.92	18.29	34.58			
RL+pg+cbdec[87]				40.66	17.87	37.06
Bottom-Up Summarization[88]				41.22	18.68	38.34

Table 4 ROUGE Scores of Different Models on the Chinese Dataset LCSTS

表 4 中文数据集 LCSTS 上不同模型 ROUGE 分数

模型	RG-1	RG-2	RG-L
RNN	21.50	8.90	18.60
RNN-context[95]	29.90	17.40	27.20
CopyNet[76]	34.40	21.60	31.30
DRGD[85]	36.99	24.15	34.21
Seq2Seq+CGU[89]	39.40	26.90	36.50

2.3 小 结

自动文摘技术的更迭经历了起步期—探索期——发展期 3 个阶段.起步期主要基于利用计算机自动地收集统计数据,通过特征评分的方法简单产生摘要.该方法不能适应复杂多变的非结构化数据.因此探索期涌现出大量主题模型、线性规划、次模函数、启发式算法等经典算法的研究工作,这期间产生的摘要可能在某些小领域取得不错的效果,无法广泛使用并落地.近年来由于神经网络的发展取得了重大进展,自动文摘的研究重点也逐渐从传统算法转向了深度学习的方法,进入一个高速发展期.相关研究者利用深度学习技术在抽取式方法和生成式方法上都取得了显著的进展.在抽取式方法中,深度学习的作用主要体现在分类模型上性能的提升,尽可能使输出结果拟合标准数据的分布.对于生成式方法来说取得了突破性的进展,改变了生成式自动文摘的研究思路,基于深度学习的生成方式模拟人类写作的习惯,其输出的结果包含了不存在原始文本中的表达方式.深度学习端到端的训练方式正式使自动文摘任务向人工智能迈出了重要一步.但不可避免的是,深度学习方法同样存在一些缺陷,如需要大量高质量标注数据、调参缺乏理论指导等问题,未来还需要研究者设计出更高效的算法来满足大数据下的自动文摘需求.

3 自动文本摘要数据集

3.1 中文数据集

3.1.1 LCSTS

LCSTS[95]是由哈尔滨工业大学智能计算中心发布的中文短文本摘要数据集,该数据集采集于新浪微博认证用户发布的超过 200 万个中文短文.作者将整个数据集分成了 3 个部分,2.4×10^6 个文本对的训练集、1×10^4 个文本对的验证集和 1.1×10^3 个文本对的测试集.其中验证集和测试集增加了摘要和原文之间的相关程度打分,分数越高代表相关程度越高,方便了研究者根据不同任务特点调整数据集的使用.

3.1.2 NLPCC

NLPCC 是由 CCF 中文信息技术专委会组织的中文计算会议.其中一项任务为面向中文微博的新闻摘要,在官网上提供了所需的实验数据.NLPCC-2015 包含从主要新闻门户网站收集的 140 篇带标题的新闻文章,每篇文章对应 2 篇人工生成的标准摘要,数据集中不同样例的原文长度之间差异较大,但提供的标准摘要的长度均不超过 140 个汉字.NLPCC-2017 提供了包含标准摘要和不包含标准摘要的 2 个训练数据集,每个数据集都包含 5 000 篇新闻文档,其中包含标准摘要的数据集中每篇文档对应 1 个摘要,摘要长度均不超过 60 个汉字.

3.1.3 搜狐新闻数据集

搜狐新闻数据集来自 2012 年 6—7 月间搜狐新闻网上国际、体育、社会 、娱乐等 18 个频道的新闻数据.根据不同的预处理方法,该数据集可分别用于文本分类、事件检测跟踪、新词发现、命名实体识别、自动文摘等任务.该数据集包含 140 万条新闻正文和新闻标题.

3.2 英文数据集

3.2.1 CNN/Daily Mail

CNN/Daily Mail 数据集是 Hermann 等人[75]从美国有线新闻网和每日邮报网中收集的大约 100 万条新闻数据作为机器阅读理解语料库,在这个数据集中的每个文章具有 1 个人工写作的多句摘要.该数据集有 2 个版本 :匿名版和非匿名版.匿名版的新闻故事里所有的命名实体已经被特殊的标签替代(例如@entity2);非匿名版的新闻故事包含原始的命名实体内容.Nallapati 等人[74]在 Hermann 的基础上构建了 CNN/Daily Mail 文本摘要数据集,包含 286 817 个训练对、13 368 个验证对和 11 487 个测试对.See 等人[77]对原始数据进行去标签等预处理后得到非匿名版数据,包含 287 226 个训练对、11 490 个验证对和 13 368 个测试对[77].

3.2.2 Gigaword

Gigaword 语料数量较大,约有 950 万篇新闻文章,数据集用第 1 句话作为输入,用标题作为文本的摘要,也属于单句摘要的数据集.英文 Gigaword 数据集最早在 2003 年由 Graff 等人[96]提出,数据是由法新社(Agence France Press)、美联社(Associated Press)、纽约时报(The New York Times)、新华社(The Xinhua News Agency)中的英文新闻文本组成.后来 Rush 等人[71]在带注解的英文 Gigaword 数据集进行了整理,得到了 3.8×10^6 个文本对的训练集、1.89×10^5 个文本对的验证集和 1951 个文本对的测试集.

3.2.3 DUC/TAC

DUC(Document Understanding Conference)是仅供测评用的小规模数据集,在 2001—2007 年 DUC 提供了自动文摘的比赛,2008 年之后更改为 TAC(Text Analysis Conference).目前常用的摘要数据集是 DUC-2002,DUC-2003,DUC-2004.DUC-2002 包含 567 篇文档,每篇文档有 2 个人工生成的 100 词的摘要;DUC-2003 包含 624 个文章-摘要对;DUC-2004 包含 500 篇文档,每篇新闻都有对应的 4 篇不同的人工生成的截取 75B 的参考摘要.

3.2.4 New York Times

New York Times 数据集[97]是经纽约时报的文章预处理后构成,它包含了 1987—2007 年间数百万篇文章,约有超过 65 万篇工作人员撰写的摘要和 150 万篇人工标注的文章,并有人、组织、位置和主题等内容的归一化索引表,可用于自动文摘、文本分类、内容提取等任务.对自动文摘任务来说,由于摘要的风格偏向于抽取式策略的结果,因此其更适合作为抽取式自动文摘的数据集.

3.2.5 Newsroom

Newsroom 数据集[98]是可用于训练和评价自动文摘系统的大型数据集,它收录了 38 个主要新闻出版社人工撰写的 130 万篇文章和摘要.这些数据是从 1998—2017 年间的搜索和社交媒体中获取得到,并使用了多种抽取式和生成式结合的策略进行摘要预处理,这使得 Newsroom 可以作为 2 种摘要产生方法的数据集.

3.2.6 Bytecup

Bytecup 数据集由 2018 Byte Cup 国际机器学习竞赛公布,由 130 万篇新闻文章组成,其中 110 万篇作为训练集.这些文章来自一站式内容消费平台 Topbuzz,每篇文章包含文章 ID、文章内容和文章标题,由于标题较短,因此该数据集更适合作为生成式自动文摘的数据集.

3.2.7 其他数据集

多年来,部分研究工作也发布了一些自动文摘数据集,其中使用较多的数据集主要包括:会议摘要数据集 AMI[99]、雅思摘要数据集 LELTS[100]、学术论文数据集[101]等.这些数据集的涌现对自动文摘任务的发展起到了很好的促进作用.

3.3 小 结

国内自动文摘起步较晚,公开数据集匮乏.中文数据集主要有源于微博的 LCSTS 和源于新闻的 NLPCC、搜狐新闻数据集,它们属于标题或单句摘要,即短文本数据集.该类型数据更适用生成式自动文摘任务的评价,不适用于抽取式方法,目前学术界缺乏大规模中文长文本摘要数据集.英文自动文摘数据集因不断有研究者贡献新的数据集,数量和种类远多于中文.CNN/Daily Mail 属于多句子摘要数据集,Newsroom 使用多种抽取式和生成式结合的摘要策略进行预处理,因此都可用于抽取式和生成式任务的评价.Gigaword 和 DUC 属于短文本数据集,主要适用于生成式任务的训练和评价;Bytecup 虽然原文较长但其面向的任务是标题生成,因此摘要较短更适用于生成式任务.New York Times 的摘要主要使用抽取式策略产生,因此比较适合抽取式任务.此外,尚有研究工作围绕细分场景构造了数据集,如科技、法律、医学等领域.高质量的自动文摘数据集可有效地促进自动文摘模型性能的提升.但随着技术的发展,在信息数据爆炸的时代我们不能过分依赖高质量的数据集,这促使科研工作者在弱监督方法上尝试新的突破.

4 自动文本摘要评价方法

自动文摘技术在各个领域得到了广泛的应用,模型的评价手段对提升文本摘要的研究结果具有重要意义.目前的评价方法根据是否有人工参与分为自动评价方法和人工评价方法,自动评价方法中常用的指标主要有 ROUGE 和 METEOR.

4.1 自动评价

4.1.1 ROUGE

ROUGE 是 Lin[102]提出的自动文摘评价方法,被广泛用于自动文摘模型性能的评价.其基本思想是将模型产生的系统摘要和参考摘要进行对比,通过计算它们之间重叠的基本单元数目来评价系统摘要的质量.常用评价指标为 ROUGE-1,ROUGE-2,ROUGE-L 等,其中 1,2,L 分别代表基于 1 元词、2 元词和最长子字串.该方法是摘要评价系统的通用标准之一,但该方法只能评价参考摘要和系统摘要的表面信息,不涉及到语义层面的评价.计算公式为

$$R_{\text{ROUGE-}N}=\frac{\sum_{S\in\{Ref\}}\sum_{N_{n\text{-gram}}\in S}Count_{\text{match}}(N_{n\text{-gram}})}{\sum_{S\in\{Ref\}}\sum_{N_{n\text{-gram}}\in S}Count(N_{n\text{-gram}})},$$

其中 n-gram 表示 n 元词,$\{Ref\}$ 表示参考摘要,$Count_{\text{match}}(N_{n\text{-gram}})$ 表示系统摘要和参考摘要中同时出现 n-gram 的个数,$Count(N_{n\text{-gram}})$ 表示参考摘要中出现 n-gram 的个数.ROUGE 还有 3 项评价指标:准确率 P(precision)、召回率 R(recall)和 F 值.ROUGE 的公式即是由召回率的计算公式演变而来.在评价阶段,研究人员常使用工具包 pyrouge 计算模型的 ROUGE 分数.

4.1.2 METEOR

Denkowski 等人[103]发现评价指标中召回率的意义后提出 METEOR 度量方法.该方法是对 BLEU[104]的改进,同时考虑了对整个语料库上的准确率和召回率,因此可信度更高.早期经常用作机器翻译的评价方法,后也被研究人员用作自动文摘任务的评价.METEOR 基于单精度的加权调和平均数以及单字召回率,P,R 分别表示系统摘要和参考摘要计算的准确率和召回率,F 值计算为

$$F_{\text{mean}}=\frac{P\times R}{\alpha P+(1-\alpha)R}.$$

为了解释单词顺序之间的差异,使用 2 个摘要文本匹配的单词总数 m 和连续有序的块 ch 的数量来计算惩罚系数 P_{Pen}:

$$P_{\text{Pen}}=\gamma\left(\frac{ch}{m}\right)^{\theta},$$

因此,METEOR 的分数是基于块的分解匹配和表征分解匹配质量的调和平均:

$$M_{\text{Score}}=(1-P_{\text{Pen}})F_{\text{mean}}.$$

α,γ,θ 是通过人工调整的参数,使其最大化与人类判断的相关性.

4.2 人工评价

因为现阶段的自动评价方法只能刻画句子之间的表层关系,不能通过语义区分摘要的质量,因此人工评价的出现在某种程度上弥补了自动评价方法的不足.但人工评价方式受母语、教育程度等因素影响较大,略显主观且效率太低.根据不同的问题,人工评价的侧重点也不同.通常会根据句子的可读性、与原文的相关性、流畅度、是否满足语法限制等属性人为地对摘要进行打分,具体细则有:

1) 可读性.摘要的书写应该是流利的,拼写应该是正确的.

2) 相关性.摘要应和原文的主题信息密切相关,不应该偏离原意.

3) 信息性.摘要应该包含原文的大部分重要信息,如果从摘要中获得的信息很少,那么这个摘要很可能是不合格的.

4) 连贯性.摘要的逻辑和语法应该是正确的.

5) 简洁性.摘要的长度尽可能精简,不能为提升其他指标而过多重复,冗余信息尽可能少.

4.3 小 结

由于缺乏原始文档或文档集合的理想参考摘要,自动文摘的性能评价一直以来是项困难的任务.理想的摘要在一定程度上是很难定义的,人类根据不同主题和角度对同一原始文档或文档集合可以撰写出不同的正确摘要,然而现有数据集普遍都是单一参考摘要,缺乏准确性和多样性.而人工评价方法受教育背景等因素影响缺乏客观性,在对比工作中可信度较低.因此,虽然自动评价 ROUGE 和 METEOR 等基于 n-gram 的方法具有无法评价语义、多样性的问题,但是其具有很高的客观性,所以被研究者广泛地作为评价模型性能的指标.近年来出现一些围绕自动文摘评价方法的研究工作,但进展缓慢.缺失标准而有效的评价方法导致自动文摘的评价面临极大挑战,这亟待相关从业者解决.

5 自动文本摘要面临的挑战及其发展趋势

目前,自动文摘技术已应用在某些特定领域.但整体来看,近年大量的工作将研究重点放在了抽取或生成的算法上,数据集与评价指标的研究工作较少.除此之外,关于自动文摘的研究工作缺乏针对性的跨越式进步,还需要突破性的创新工作提升性能才能更广泛地适应各个场景,所以自动文摘任务的质量和性能还面临诸多挑战:

1) 数据集.高质量的自动文摘数据集较少,甚至中文长文本数据集缺失,限制了中文文本摘要技术的研究.

2) 评价指标.自动评价方法过于死板,人工评价方法较主观,缺乏被学术界广泛认可并切实可行的评价方法,这减缓了该任务的发展.

3) 语义表达.文档的摘要应有多种表达方式,但是目前来说同一语义的不同表达、重复表达同一语义的问题还需要相应的工作来解决.

自动文摘的研究已经有近 60 年的历史,由于该任务的难度导致初期的效果并不理想,随着深度学习的快速发展才使得人们看到自动文摘广泛应用的希望.长期看来,自动文摘的发展有 6 个趋势:

1) 数据集.中文、英文和其他语言的高质量自动文摘数据集将有可能推动自动文摘任务的发展,若仅依靠人工参与构建数据集将是项耗时耗力的工作,因此如果可以通过计算机自动地构建高质量数据集将是非常有意义的.

2) 评价指标.目前有工作提出通过计算文本之间语义相似度、改进的 ROUGE 等对自动文摘进行评价,但尚不能有效地扩展,因此更加完善的自动文摘评价指标必然是长期研究的重点问题.

3) 方法融合.新技术的探索是永远的话题,对传统算法与深度学习的结合,或抽取式方法与生成式方法进一步融合将是学术界乃至工业界必然的趋势.

4) 借助外部知识.机器效仿人类生成摘要的过程时需要背景知识的辅助(如纳入背景知识库),对于深度学习方法来说还可用预训练的模型为自动文摘模型提供强有力的外部知识.

5) 弱监督或无监督发展.由于缺乏高质量的自动文摘数据集,一种有效可靠的方法是通过少量的训练数据或无训练数据使用高效的算法处理自动文摘任务.

6) 应用场景.研究人员的重心将会慢慢从普适性的工作转移到特定细分场景上,针对不同的子任务场景提出更加具有针对性的算法,如新闻标题、自动对联、评论摘要、会议摘要、金融快报等.

6 总 结

自动文摘技术自 20 世纪 50 年代末提出,经历了一段缓慢的发展历程,如今深度学习所展现的优秀表现给自动文摘的研究带来了新的机会,使其近年来快速发展,进入高速发展期.自动文摘属于自然

语言处理领域中文本生成的范畴,其社会价值促使自动文摘在自然语言处理领域占有重要的地位.目前该技术不仅在金融、新闻、媒体等领域表现出优秀的性能,还在信息检索、舆情分析、内容审查等方面展现出重要的作用.本文通过对众多研究工作的回顾和分析,对自动文摘技术算法进行了分类梳理,从抽取式方法和生成式方法2个角度介绍了常见的自动文摘算法,并对与之紧密相关的数据集和评价指标进行了详细介绍.最后本文对自动文摘面临的挑战和未来的发展趋势做出了预测和展望.可以预见,随着新技术的发展、模型性能的提升,其应用将越来越广泛,在不远的将来可显著地提高人们在海量数据中的信息获取效率,为人类的生活带来更多便利.

参 考 文 献

[1] Liu Chengying, Chen M, Tseng C. Incrests: Towards real-time incremental short text summarization on comment streams from social network services [J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(11): 2986-3000

[2] Marujo L, Ling Wang, Ribeiro R, et al. Exploring events and distributed representations of text in multi-document summarization [J]. Knowledge-Based Systems, 2016, 94: 33-42

[3] Codina-Filbà J, Bouayad-Agha N, Burga A, et al. Using genre-specific features for patent summaries [J]. Information Processing & Management, 2017, 53(1): 151-174

[4] Condori R E L, Pardo T A S. Opinion summarization methods: Comparing and extending extractive and abstractive approaches [J]. Expert Systems with Applications, 2017, 78: 124-134

[5] Mishra R, Bian J, Fiszman M, et al. Text summarization in the biomedical domain: A systematic review of recent research [J]. Journal of Biomedical Informatics, 2014, 52: 457-467

[6] Wan Xiaojun, Yao Jinge. Research progress and trend of automatic summarization [EB/OL]. 2016 [2018-06-01]. http://www.cipsc.org.cn/qngw/?p=979 (in Chinese)
(万小军, 姚金戈. 自动文摘研究进展与趋势[EB/OL]. 2016 [2018-06-01]. http://www.cipsc.org.cn/qngw/?p=979)

[7] Wang Junli, Wei Shaochen, Guan Min. Survey on graph model-based document summarization [J]. Computer Science, 2015, 42(12): 1-7 (in Chinese)
(王俊丽, 魏绍臣, 管敏. 基于图排序算法的自动文摘研究综述[J]. 计算机科学, 2015, 42(12): 1-7)

[8] Cao Yang, Cheng Ying, Pei Lei. A review on machine learning oriented automatic summarization [J]. Library and Information Service, 2014, 58(18): 122-130 (in Chinese)
(曹洋, 成颖, 裴雷. 基于机器学习的自动文摘研究综述[J]. 图书情报工作, 2014, 58(18): 122-130)

[9] Luhn H P. The automatic creation of literature abstracts [J]. IBM Journal of Research and Development, 1958, 2(2): 159-165

[10] Qin Bing, Liu Ting, Chen Shanglin, et al. Sentences optimum selection for multi-document summarization [J]. Journal of Computer Research and Development, 2006, 43(6): 1129-1134 (in Chinese)
(秦兵, 刘挺, 陈尚林, 等. 多文档文摘中句子优化选择方法研究[J]. 计算机研究与发展, 2006, 43(6): 1129-1134)

[11] Carbonell J G, Goldstein J. The use of MMR, diversity-based reranking for reodering documents and producing summaries [C] //Proc of the 21st Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 1998: 335-336

[12] Erkan G, Radev D R. LexRank: Graph-based lexical centrality as salience in text summarization [J]. Journal of Artificial Intelligence Research, 2004, 22(1): 457-479

[13] Ceylan H, Mihalcea R, Özertem U, et al. Quantifying the limits and success of extractive summarization systems across domains [C] //Proc of the 2010 Conf of the NAACL. Stroudsburg: ACL, 2010: 903-911

[14] Jin Feng, Huang Minlie, Zhu Xiaoyan. A comparative study on ranking and selection strategies for multi-document summarization [C] //Proc of the 23rd Int Conf on Computational Linguistics. Stroudsburg: ACL, 2010: 525-533

[15] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis [J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407

[16] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022

[17] Gong Yihong, Liu Xin. Generic text summarization using relevance measure and latent semantic analysis [C] //Proc of the 24th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2001: 19-25

[18] Steinberger J, Kabadjov M A, Poesio M, et al. Improving LSA-based summarization with anaphora resolution [C] //Proc of the Conf on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2005: 1-8

[19] Poesio M, Kabadjov M A. A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation [C/OL] //Proc of the LREC. Paris: European Language Resources Association, 2004 [2018-06-03]. http://www.lrec-conf.org/proceedings/lrec2004/summaries/559.htm

[20] Hofmann T. Probabilistic latent semantic indexing [J]. ACM SIGIR Forum, 2017, 51(2): 211-218

[21] Kar M, Nunes S, Ribeiro C. Summarization of changes in dynamic text collections using latent Dirichlet allocation model [J]. Information Processing & Management, 2015, 51(6): 809-833

[22] Mihalcea R, Tarau P. TextRank: Bringing order into text [C] //Proc of the 2004 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2004: 404-411

[23] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine [J]. Computer Networks and ISDN Systems, 1998, 30(1/2/3/4/5/6/7): 107-117

[24] Erkan G, Radev D R. LexPageRank: Prestige in multi-document text summarization [C] //Proc of the 2004 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2004: 365-371

[25] Leiva L A. Responsive text summarization [J]. Information Processing Letters, 2018, 130: 52-57

[26] Fang Changjian, Mu Dejun, Deng Zhenghong, et al. Word-sentence co-ranking for automatic extractive text summarization [J]. Expert Systems with Applications, 2017, 72: 189-195

[27] Parveen D, Strube M. Integrating importance, non-redundancy and coherence in graph-based extractive summarization [C] //Proc of the 24th Int Joint Conf on Artificial Intelligence. Menlo Park: AAAI, 2015: 1298-1304

[28] Ferreira R, Cabral L D, Lins R D, et al. Assessing sentence scoring techniques for extractive text summarization [J]. Expert Systems with Applications, 2013, 40(14): 5755-5764

[29] Wang Waiming, Li Zhi, Wang Jiawen, et al. How far we can go with extractive text summarization? Heuristic methods to obtain near upper bounds [J]. Expert Systems with Applications, 2017, 90: 439-463

[30] Oliveira H, Ferreira R, Lima R, et al. Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization [J]. Expert Systems with Applications, 2016, 65: 68-86

[31] Aone C, Okurowski M E, Gorlinsky J, et al. A scalable summarization system using robust NLP [J/OL]. Intelligent Scalable Text Summarization, 1997 [2018-09-10]. https://www.aclweb.org/anthology/W97-0711/

[32] Kupiec J, Pedersen J, Chen F. A trainable document summarizer [C] //Proc of the 18th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 1995: 68-73

[33] Conroy J M, O'leary D P. Text summarization via hidden Markov models [C] //Proc of the 24th Annual Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2001: 406-407

[34] Shen Dou, Sun Jiantao, Li Hua, et al. Document summarization using conditional random fields [C] //Proc of Int Joint Conf on Artificial Intelligence. Menlo Park: AAAI, 2007: 2862-2867

[35] Louis A. A Bayesian method to incorporate background knowledge during automatic text summarization [C] //Proc of the 52nd Annual Meeting of the ACL. Stroudsburg: ACL, 2014: 333-338

[36] Itti L, Baldi P. Bayesian surprise attracts human attention [J]. Vision Research, 2009, 49(10): 1295-1306

[37] Abdi A, Shamsuddin S M, Hasan S, et al. Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment [J]. Expert Systems with Applications, 2018, 109: 66-85

[38] Kenji K, Larry A. The feature selection problem: Traditional methods and a new algorithm [C] //Proc of the 10th National Conf on Artificial Intelligence. Menlo Park: AAAI, 1992: 129-134

[39] Sanchez-Gomez J M, Vega-Rodríguez M A, Pérez C J. Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach [J]. Knowledge-Based Systems, 2018, 159: 1-8

[40] Mosa M A, Hamouda A, Marei M. Graph coloring and ACO based summarization for social networks [J]. Expert Systems with Applications, 2017, 74: 115-126

[41] Peyrard M, Eckle-Kohler J. Supervised learning of automatic pyramid for optimization-based multi-document summarization [C] //Proc of the 55th Annual Meeting of the ACL. Stroudsburg: ACL, 2017: 1084-1094

[42] Litvak M, Vanetik N, Last M, et al. Museec: A multilingual text summarization tool [C] //Proc of ACL-2016 System Demonstrations. Stroudsburg: ACL, 2016: 73-78

[43] Peyrard M, Eckle-Kohler J. Optimizing an approximation of rouge—A problem-reduction approach to extractive multi-document summarization [C] //Proc of the 54th Annual Meeting of the ACL. Stroudsburg: ACL, 2016: 1825-1836

[44] Huang Xiaojiang, Wan Xiaojun, Xiao Jianguo. Comparative news summarization using linear programming [C] //Proc of the 49th Annual Meeting of the ACL. Stroudsburg: ACL, 2011: 648-653

[45] Galanis D, Lampouras G, Androutsopoulos I. Extractive multi-document summarization with integer linear programming and support vector regression [C] //Proc of Int Conf on Computational Linguistics. Stroudsburg: ACL, 2012: 911-926

[46] McDonald R. A study of global inference algorithms in multi-document summarization [C] //Proc of European Conf on Information Retrieval. Berlin: Springer, 2007: 557-564

[47] Gillick D, Favre B. A scalable global model for summarization [C] //Proc of the Workshop on Integer Linear Programming for Natural Language Processing. Stroudsburg: ACL, 2009: 10-18

[48] Boudin F, Mougard H, Favre B, et al. Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015: 1914-1918

[49] Lin Hui, Bilmes J. Multi-document summarization via budgeted maximization of submodular functions [C] //Proc of the 2010 Annual Conf of the NAACL. Stroudsburg: ACL, 2010: 912-920

[50] Lin Hui, Bilmes J. A class of submodular functions for document summarization [C] //Proc of the 49th Annual Meeting of the ACL. Stroudsburg: ACL, 2011: 510-520

[51] Wu Keshou, Li Lei, Li Jingxuan, et al. Ontology-enriched multi-document summarization in disaster management using submodular function [J]. Information Sciences, 2013, 224: 118-129

[52] Chali Y, Tanvee M, Nayeem M T. Towards abstractive multi-document summarization using submodular function-based framework, sentence compression and merging [C] // Proc of the 8th Int Joint Conf on Natural Language Processing. Stroudsburg: ACL, 2017, 2: 418-424

[53] Bairi R, Iyer R, Ramakrishnan G, et al. Summarization of multi-document topic hierarchies using submodular mixtures [C] //Proc of the 53rd Annual Meeting of the ACL and the 7th Int Joint Conf on Natural Language Processing. Stroudsburg: ACL, 2015: 553-563

[54] Liu Yan, Zhong Shenghua, Li Wenjie. Query-oriented multi-document summarization via unsupervised deep learning [C] //Proc of the AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2012: 1699-1705

[55] Cao Ziqiang, Li Wenjie, Li Sujian, et al. Improving multi-document summarization via text classification [C] //Proc of the AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2017: 3053-3059

[56] Yin Wenpeng, Pei Yulong. Optimizing sentence modeling and selection for document summarization [C] //Proc of Int Joint Conf on Artificial Intelligence. Menlo Park: AAAI, 2015: 1383-1389

[57] Singh A K, Gupta M, Varma V. Unity in diversity: Learning distributed heterogeneous sentence representation for extractive summarization [C] //Proc of the AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2018: 5473-5480

[58] Cheng Jianpeng, Lapata M. Neural summarization by extracting sentences and words [C] //Proc of the 54th Annual Meeting of the ACL. Stroudsburg: ACL, 2016: 484-494

[59] Nallapati R, Zhai Feifei, Zhou Bowen. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents [C] //Proc of the AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2017: 3075-3081

[60] Chen Xiuying, Gao Shen, Tao Chongyang, et al. Iterative document representation learning towards summarization with polishing [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 4088-4097

[61] Mehdad Y, Carenini G, Ng R T. Abstractive summarization of spoken and written conversations based on phrasal queries [C] //Proc of the 52nd Annual Meeting of the ACL. Stroudsburg: ACL, 2014: 1220-1230

[62] Banerjee S, Mitra P, Sugiyama K. Multi-document abstractive summarization using ILP based multi-sentence compression [C] //Proc of Int Joint Conf on Artificial Intelligence. Menlo Park: AAAI, 2015: 1208-1214

[63] Durrett G, Bergkirkpatrick T, Klein D, et al. Learning-based single-document summarization with compression and anaphoricity constraints [C] //Proc of the 54th Annual Meeting of the ACL. Stroudsburg: ACL, 2016: 1998-2008

[64] Liu Fei, Flanigan J, Thomson S, et al. Toward abstractive summarization using semantic representations [C] //Proc of the 2015 Conf of the NAACL. Stroudsburg: ACL, 2015: 1077-1086

[65] Takase S, Suzuki J, Okazaki N, et al. Neural headline generation on abstract meaning representation [C] //Proc of the 2016 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2016: 1054-1059

[66] Dohare S, Gupta V, Karnick H. Unsupervised semantic abstractive summarization [C] //Proc of the ACL 2018, Student Research Workshop. Stroudsburg: ACL, 2018: 74-83

[67] Li Wei. Abstractive multi-document summarization with semantic information extraction [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015: 1908-1913

[68] Zhou Liang, Hovy E. Template-filtered headline summarization [C] //Proc of the ACL Workshop on Text Summarization. Stroudsburg: ACL, 2004:56-60

[69] Oya T, Mehdad Y, Carenini G, et al. A template-based abstractive meeting summarization: Leveraging summary and source text relationships [C] //Proc of the 8th Int Natural Language Generation Conf. Stroudsburg: ACL, 2014: 45-53

[70] Cao Ziqiang, Li Wenjie, Li Sujian, et al. Retrieve, rerank and rewrite: Soft template based neural summarization [C] // Proc of the 56th Annual Meeting of the ACL. Stroudsburg: ACL, 2018: 152-161

[71] Rush A M, Chopra S, Weston J, et al. A neural attention model for abstractive sentence summarization [C] //Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015: 379-389

[72] Bahdanau D, Cho K, Bengio Y, et al. Neural machine translation by jointly learning to align and translate[C/OL]// Proc of the Int Conf on Learning Representations. 2015 [2019-01-15]. https://arxiv.org/abs/1409.0473

[73] Chopra S, Auli M, Rush A M. Abstractive sentence summarization with attentive recurrent neural networks [C] //Proc of the 2016 Conf of the North American Chapter of the ACL: Human Language Technologies. Stroudsburg: ACL, 2016: 93-98

[74] Nallapati R, Zhou Bowen, Santos C N, et al. Abstractive text summarization using sequence-to-sequence RNNs and beyond [C] //Proc of the 20th SIGNLL Conf on Computational Natural Language Learning. Stroudsburg: ACL, 2016: 280-290

[75] Hermann K M, Kocisky T, Grefenstette E, et al. Teaching machines to read and comprehend [C] //Proc of the Advances in Neural Information Processing Systems. New York: Curran Associates, Inc, 2015: 1693-1701

[76] Gu Jiatao, Lu Zhengdong, Li Hang, et al. Incorporating copying mechanism in sequence-to-sequence learning [C] // Proc of the 54th Annual Meeting of the ACL. Stroudsburg: ACL, 2016: 1631-1640

[77] See A, Liu P J, Manning C D, et al. Get to the point: Summarization with pointer-generator networks [C] //Proc of the 55th Annual Meeting of the ACL. Stroudsburg: ACL, 2017: 1073-1083

[78] Tu Zhaopeng, Lu Zhengdong, Liu Yang, et al. Modeling coverage for neural machine translation [C] //Proc of the 54th Annual Meeting of the ACL. Stroudsburg: ACL, 2016: 76-85

[79] Williams R J, Zipser D. A learning algorithm for continually running fully recurrent neural networks [J]. Neural Computation, 1989, 1(2): 270-280

[80] Paulus R, Xiong Caiming, Socher R, et al. A deep reinforced model for abstractive summarization [C] //Proc of the Int Conf on Learning Representations. 2018 [2019-01-20]. https://openreview.net/forum?id=HkACIQgA-

[81] Rennie S J, Marcheret E, Mroueh Y, et al. Self-critical sequence training for image captioning [C] //Proc of the IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society, 2017: 7008-7024

[82] Cao Ziqiang, Wei Furu, Li Wenjie, et al. Faithful to the original: Fact aware neural abstractive summarization [C] // Proc of Int Joint Conf on Artificial Intelligence. Menlo Park: AAAI, 2018: 4784-4791

[83] Hsu W, Lin C, Lee M, et al. A unified model for extractive and abstractive summarization using inconsistency loss [C] // Proc of the 56th Annual Meeting of the ACL. Stroudsburg: ACL, 2018: 132-141

[84] Zhou Qingyu, Yang Nan, Wei Furu, et al. Selective encoding for abstractive sentence summarization [C] //Proc of the 55th Annual Meeting of the ACL. Stroudsburg: ACL, 2017: 1095-1104

[85] Li Piji, Lam W, Bing Lidong, et al. Deep recurrent generative decoder for abstractive text summarization [C] // Proc of the 2017 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2017: 2091-2100

[86] Kingma D P, Welling M. Stochastic gradient VB and the variational auto-encoder [C/OL] //Proc of the Int Conf on Learning Representations. 2014 [2019-02-10]. https://arxiv.org/abs/1312.6114v8

[87] Jiang Y, Bansal M. Closed-book training to improve summarization encoder memory [C] //Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 4067-4077

[88] Gehrmann S, Deng Y, Rush A M, et al. Bottom-up abstractive summarization [C] //Proc of the 2018 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2018: 4098-4109

[89] Lin J, Sun X, Ma S, et al. Global encoding for abstractive summarization [C] //Proc of the 56th Annual Meeting of the ACL. Stroudsburg: ACL, 2018: 163-169

[90] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C] //Proc of the Advances in Neural Information Processing Systems. New York: Curran Associates, Inc, 2017: 5998-6008

[91] Zhang Haoyu, Gong Yeyun, Yan Yan, et al. Pretraining-based natural language generation for text summarization [C] //Proc of the 23rd Conf on Computational Natural Language Learning. Stroudsburg: ACL, 2019: 789-797

[92] Gehring J, Auli M, Grangier D, et al. Convolutional sequence to sequence learning [C] //Proc of the 34th Int Conf on Machine Learning. New York: ACM, 2017: 1243-1252

[93] Fan A, Grangier D, Auli M, et al. Controllable abstractive summarization [C] //Proc of the 56th Annual Meeting of the ACL. Stroudsburg: ACL, 2018: 45-54

[94] Wang Li, Yao Junlin, Tao Yunzhe, et al. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization [C] //Proc of Int Joint Conf on Artificial Intelligence. Menlo Park: AAAI, 2018: 4453-4460

[95] Hu Baotian, Chen Qingcai, Zhu Fangze, et al. LCSTS: A large scale Chinese short text summarization dataset [C] // Proc of the 2015 Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015: 1967-1972

[96] Graff D, Christopher C. English Gigaword [J/OL]. Linguistic Data Consortium, 2003 [2017-09-10]. https://catalog.ldc.upenn.edu/LDC2003T05

[97] Sandhaus E. The New York Times annotated corpus [J/OL]. Linguistic Data Consortium, 2008 [2017-10-20]. https://catalog.ldc.upenn.edu/docs/LDC2008T19/new_york_times_annotated_corpus.pdf

[98] Grusky M, Naaman M, Artzi Y. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies [C] //Proc of the 2018 Conf of the NAACL. Stroudsburg: ACL, 2018: 708-719

[99] Carletta J, Ashby S, Bourban S, et al. The AMI meeting corpus: A pre-announcement [C] //Proc of the Int Workshop on Machine Learning for Multimodal Interaction. Berlin: Springer, 2005: 28-39

[100] Fang Yimai, Zhu Haoyue, Muszyńska E, et al. A proposition-based abstractive summariser [C] //Proc of the 26th Int Conf on Computational Linguistics. Stroudsburg: ACL, 2016: 567-578

[101] Yasunaga M, Kasai J, Zhang Rui, et al. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks [C] // Proc of the AAAI Conf on Artificial Intelligence. Menlo Park: AAAI, 2019: 7386-7393

[102] Lin C. ROUGE: A package for automatic evaluation of summaries [C] //Proc of the Workshop of ACL 2004. Stroudsburg: ACL, 2004: 74-81

[103] Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language [C] // Proc of the 9th Workshop on Statistical Machine Translation. Stroudsburg: ACL, 2014: 376-380

[104] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation [C] //Proc of the 40th Annual Meeting of the ACL. Stroudsburg: ACL, 2002: 311-318

Li Jinpeng, born in 1994. Master. His main research interests include natural language processing and deep learning.

李金鹏,1994 年生.硕士.主要研究方向为自然语言处理和深度学习.

Zhang Chuang, born in 1982. PhD, associate professor. Member of CCF. His main research interests include natural language processing and cloud computing.

张　闯,1982 年生.博士,高级工程师,CCF 会员.主要研究方向为自然语言处理和云计算.

Chen Xiaojun, born in 1979. PhD, professor senior engineer. Member of CCF. His main research interests include natural language processing, collaborative intelligence, data security and privacy preserving.

陈小军,1979 年生.博士,正高级工程师,CCF 会员.主要研究方向为自然语言处理、协同智能、数据安全和隐私保护.

Hu Yue, born in 1963. PhD, professor. Member of CCF. Her main research interests include natural language processing and deep learning.

胡　玥,1963 年生.博士,研究员,CCF 会员.主要研究方向为自然语言处理和深度学习.

Liao Pengcheng, born in 1996. Master candidate. Student member of CCF. His main research interests include natural language processing and deep learning.

廖鹏程,1996 年生.硕士研究生,CCF 学生会员.主要研究方向为自然语言处理和深度学习.