文章编号：1003-0077(2019)05-0001-16

文本摘要常用数据集和方法研究综述

侯圣峦[1,2]，张书涵[1,2]，费超群[1,2]

(1. 中国科学院 计算技术研究所 智能信息处理重点实验室，北京 100190；
2. 中国科学院大学，北京 100049)

摘 要：文本摘要成为人们从互联网上海量文本信息中便捷获取知识的重要手段。现有方法都是在特定数据集上进行训练和效果评价，包括一些公用数据集和作者自建数据集。已有综述文献对现有方法进行全面细致的总结，但大多都是对方法进行总结，而缺少对数据集的详细描述。该文从调研数据集的角度出发，对文本摘要常用数据集及在该数据集上的经典和最新方法进行综述。对公用数据集的综述包括数据来源、语言及获取方式等，对自建数据集的总结包括数据规模、获取和标注方式等。对于每一种公用数据集，给出了文本摘要问题的形式化定义。同时，对经典和最新方法在特定数据集上的实验效果进行了分析。最后，总结了已有常用数据集和方法的现状，并指出存在的一些问题。

关键词：文本摘要；自然语言处理；机器学习；人工智能

中图分类号：TP391 **文献标识码**：A

A Survey to Text Summarization: Popular Datasets and Methods

HOU Shengluan[1,2], ZHANG Shuhan[1,2], FEI Chaoqun[1,2]

(1. Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Text summarization has become an essential way of knowledge acquisition from mass text documents on the Internet. The existing surveys to text summarization are mostly focused on methods, without reviewing on the experimental datasets. This survey concentrates on evaluation datasets and summarizes the public and private datasets together with corresponding approaches. The public datasets are recorded for the data source, language and the way of access, and the private dataset are recorded with the scale, access and annotation methods. In addition, the formal definition of text summarization by each public dataset are provided. We analyze the experimental results of classical and latest text summarization methods on one specific dataset. We conclude with the present situation of existing datasets and methods, and some issues concerning them.

Keywords: text summarization; natural language processing; machine learning; artificial intelligence

0 引言

文本摘要任务旨在从一篇或多篇相同主题的文本中抽取能够反映主题的精简压缩版本[1-2]，可以帮助用户快速形成对特定主题文本内容的全面了解，提高浏览信息和获取知识的效率。随着互联网上文本数量的爆炸式增长，对文本摘要的需求也越来越大，近十几年来，许多准确而高效的文本摘要算法被提出。

文本摘要方法的分类方式有多种。根据输入文本的数量，文本摘要方法可以分为单文本摘要方法和多文本摘要方法。根据不同的标准，文本摘要方法又有不同的分类体系。表 1 总结了现有的主流文本摘要方法分类体系，可以看出，针对不同的文本摘要任务需求可以使用不同的方法，以达到更好的

收稿日期：2018-06-28 **定稿日期**：2018-09-21

基金项目：国家重点研发计划项目(2016YFB1000902)；国家自然科学基金(61232015, 21472412, 61621003)

效果。

已有工作都通过特定数据集来训练和评估提出方法的性能，有些使用公用数据集，有些数据集则是作者根据互联网上的文本资源自建的。目前关于文本摘要综述的文献较多[2-4]，但多是针对不同类别方法，从不同维度的分析，缺少对方法用到的实验数据集的总结描述。另一方面，虽然已有少量工作面向跨语言的文本摘要方法研究[5]，但仍处于初步阶段。已有综述文献主要是对于英文文本摘要方法的总结综述，缺少对中文文本摘要方法的综述和面向英文文本摘要方法对中文文本的可适用性分析。

表1 主流文本摘要方法分类体系

分类标准	分类方法	解释
文本数量	{单文本摘要(single-document summarization)，多文本摘要(multi-document summarization)}	二者的区别在于是从一篇文本还是多篇文本中生成摘要，多文本摘要还要保证生成的摘要不矛盾、无冗余等
抽取方式	{抽取式摘要(extractive summarization)，生成式摘要(abstractive summarization)}	抽取式摘要直接从原文本中不加修改地抽取文本片段(通常是句子)组成摘要；生成式摘要是重新组织句子形成比抽取式摘要更加精简的形式
面向语言	{面向中文的摘要(Chinese text summarization)，面向英文的摘要(English text summarization)，跨语言摘要(multilingual text summarization)，……}	大部分方法都是面向一种语言提出的。随着跨语言文本的出现，一些跨语言的文本摘要方法也被提出，但研究处于初步阶段
摘要目标	{面向查询的摘要(query-oriented summarization)，一般总结性摘要(generic summarization)}	面向查询的摘要是只抽取与查询问题相关的成分组成摘要；一般总结性摘要则是考虑所有文章成分的重要程度，抽取重要部分组成摘要
领域需求	{医学摘要(medical summarization)，邮件摘要(email summarization)，……}	领域摘要方法是对应特定需求的摘要方法[3]。例如医学摘要更重视信息的精度；邮件摘要需要注重单次出现的词(unique word)来进行新邮件提醒
……	……	……

本文从文本摘要相关技术和所用到的数据集出发，对已有工作进行调研，总结了目前常用的数据集和方法。我们将文本摘要常用数据集分为两种，一种是公用的、专门用于测试文本摘要方法性能的数据集，我们称之为公用数据集；另一种是在文献中作者为验证方法独立构建的数据集，我们称之为自建数据集。本文内容主要包括以下几个方面：

(1) 文本摘要常用数据集总结。

- 对于公用数据集，包括来源、语言、规模和获取方式等；
- 对于自建数据集，包括来源、规模、获取方式和标注方法。

(2) 对于每一种公用数据集，给出了文本摘要问题的形式化定义，并对经典和最新方法进行综述。选定一种数据集，对已有方法在该数据集上的实验效果进行了总结分析。

(3) 总结了现有常用数据集和对应方法的研究现状、存在的问题。

本文剩余部分组织结构如下：第1节是文本摘要常用数据集总体概览；第2～8节是常用公用数据集的介绍及在该数据集上几种典型方法的详述；第9节是对自建数据集及对应方法的综述；第10节总结了经典算法和最新方法用到的数据集；第11节分析了经典方法在数据集上的实验效果；最后一节总结了发展趋势，指出了存在的问题。

1 文本摘要常用数据集总体概况

文本摘要常用数据集包括两部分：一是公用数据集，二是作者自建数据集。本节总结了中英文文本摘要中常用公用数据集，这些数据集的概览如表2所示。

表 2　中英文文本摘要方法中常用公用数据集概览

数据集名称	语言	适用方法	摘要方式	数据规模	获取方式
DUC/TAC	英文	单文本摘要/多文本摘要	抽取式摘要/生成式摘要	每期百篇规模	网页申请
Gigaword	英文	单文本摘要	生成式摘要	4 000 000	付费申请
CNN/Daily Mail	英文	单文本摘要	抽取式摘要/生成式摘要	300 000	免费获取
NYTAC	英文	单文本摘要	生成式摘要	650 000	付费申请
Amazon SNAP Review Dataset	英文	多文本摘要	生成式摘要	35 000 000	免费获取
LCSTS	中文	单文本摘要	生成式摘要	2 400 000	网页申请
NLPCC	中文	单文本摘要	生成式摘要	50 000	网页申请

从表 2 可知，面向英文的文本摘要方法中用到的公用数据集较多，面向中文文本摘要方法的公用数据集包括两个：LCSTS 和 NLPCC，且都是用于生成式摘要方法的性能测评。

从适用方法来看，已有公用数据集大都用于单文本摘要方法。DUC/TAC 可以用于普通文本的多文本摘要方法，Amazon SNAP Review Dataset 常用于评论和情感的多文本摘要方法。

就摘要方式来说，用于生成式摘要方法的数据集较多。为了解决抽取式摘要方法缺少训练数据的问题，已有方法通常将用于生成式文本摘要的数据集进行简单转换，例如，Cheng 等[6] 将 CNN/Daily Mail 数据集中的每篇文本中句子与生成式摘要句计算匹配度，匹配度较高的句子作为抽取式摘要句，构成抽取式摘要方法的数据集。

诸多工作尝试深度神经网络模型在文本摘要中的应用。但由于深度学习模型复杂，待学习参数较多，因此需要较大规模的训练数据。Gigaword、CNN/Daily Mail、LCSTS 等都是十万级规模，可满足深度神经网络训练的需求。

2　DUC/TAC

文本理解会议（Document Understanding Conference，DUC）[①] 主要面向英文文本摘要的评估，从 2001 年到 2007 年每年发布 1 次测评数据集。从 2008 年开始，DUC 成为了文本分析会议（Text Analysis Conference，TAC）[②] 中的一个文本摘要任务。自 2003 年起，DUC/TAC 主要面向多文本摘要任务，所以对单文本摘要方法来说，测试数据集更少。

TAC2014 提出了面向生物医学领域的科技文献文本摘要任务，其余 DUC/TAC 数据集面向新闻类文本摘要任务。此处我们随机选择 DUC2004 和 TAC2009 数据集进行分析：DUC2004 包括单文本摘要和多文本摘要两个任务，其中单文本摘要任务包括 500 篇文本；多文本摘要任务包括 50 个文本簇，每一个文本簇中有 10 篇文本。TAC2009 中的多文本摘要任务数据集包括 44 个主题，每个主题有两个文本集，分别包括 10 篇新闻文本，用于文本摘要生成。

从以上分析可知，DUC/TAC 是人工标注的生成式摘要数据集。由于 DUC/TAC 数据集在百篇规模，不适用于训练深度神经网络模型，常用于传统文本摘要方法的性能评估。

定义 1　DUC/TAC 数据集上的文本摘要

给定 $k(0<k\leqslant 50)$ 组英文新闻类文本集合 $\{T_1, T_2, \cdots, T_k\}$，其中 $T_i(1\leqslant i\leqslant k)$ 包含描述同一主题的 $p(0<p\leqslant 25)$ 篇文本，主要包括以下几个任务：

（1）单文本摘要。对于 T_i 中的每一篇文本 $D_j(0<j\leqslant 25)$，生成长度限制在 L_s 的语法通顺、结构连贯的摘要。

（2）多文本摘要。对于每一个 T_i，生成内容覆盖 T_i 中所有重要文本信息，并且长度限制在 L_m 的摘要。

（3）面向查询的摘要。对于每一个 T_i，生成可以回答问题 Q 并且长度限制在 L_q 的摘要。其中 Q 是例如“Who is X？”形式的问题。 □

经典的 DUC/TAC 数据集上的方法主要包括基于图模型的方法和基于传统机器学习的方法。

① http://duc.nist.gov/

② http://www.nist.gov/tac/

2.1 基于图模型的方法

基于图模型的方法是将文本单元(如句子或者词)作为节点,文本单元间关系作为边构建图模型,通过图挖掘等算法从图中抽取重要成分组成摘要。典型方法包括LexRank[1]和TextRank[7]。

LexRank将句子作为节点、句子间的语义相似关系作为边构建图模型。从图模型中根据节点间的边及权重抽取重要句子作为摘要句。LexRank在DUC2003和DUC2004上测试了方法的性能。

TextRank的基本思想则是PageRank[8],通过对待处理文本建立图模型,利用投票机制对文本中的重要成分进行排序。TextRank可以用于关键词提取和摘要句的抽取。其优点是简洁高效,不需要事先对模型进行训练,属于一种无监督方法。TextRank在DUC2002上验证了方法的有效性。

Baralis等[9]提出了一种改进的算法GraphSum。GraphSum利用关联规则挖掘来计算句子间的相似度,然后构建文本摘要图模型,利用PageRank算法迭代计算得到摘要句。在DUC2004数据集上取得了更好的效果。

虽然LexRank和TextRank在英文文本数据集上进行了性能评估,但同样适用于中文文本摘要的提取。由于缺少面向中文的公用数据集,面向中文的基于图模型的方法都采用自建数据集进行方法测评,具体方法将在第9节详述。

2.2 基于传统机器学习的方法

随着机器学习技术的发展和在自然语言处理等各个领域的成功应用,越来越多的工作将机器学习算法应用到文本摘要中。本文将利用朴素贝叶斯、支持向量机等理论的文本摘要方法归为基于传统机器学习的方法,将利用深度神经网络模型的文本摘要方法归为基于深度学习的方法。其中,基于传统机器学习的文本摘要方法效果取决于特征提取、模型选择及训练数据规模。

Gillick和Favre[10]将多文本摘要看成优化问题,利用整数线性规划构建文本摘要模型。实验结果表明,在TAC2008数据集上利用二元词组(bigram)特征较利用一元词组(unigram)和三元词组(trigram)特征的效果更好。

为了取得更好效果,Fattah[11]结合多种机器学习模型,考虑以下几种特征:词的相似度、文本格式、中心段、整篇文本中词频统计分值、标题、句子位置和无关信息是否出现等,并利用这些特征提出了一种结合最大熵、朴素贝叶斯分类器和支持向量机三种模型的多文本摘要方法。由于每个模型都可以看成是一个二分类器,通过引入联合概率分布函数来判断句子的重要程度,最后实现了一种基于混合机器学习模型的多文本摘要方法。方法在DUC2001和DUC2002数据集上取得了较好效果,但缺点在于方法的复杂度太高。

3 Gigaword

Gigaword① 是一个由英文新闻文章组成的数据集,共包括接近950万来自纽约时报(New York Times)等多个新闻源的新闻语料,其中部分文章包含一句话的简短新闻提要(headline)。将新闻提要与文章的首句话组成生成式摘要平行语料库,用于深度神经网络模型的训练与测试。Gigaword用于生成式文本摘要方法的数据规模见表3。

表3 Gigaword数据集规模

数据集	规模
训练集	3 800 000
验证集	189 000
测试集	2 000

定义2 Gigaword数据集上的文本摘要

给定文本集D,包含k个英文新闻类文本摘要对$<h,\text{first}>$,其中h表示单句话的新闻提要,由新闻作者给出,first表示新闻正文的首句话。将D分成三份,分别是训练集D_{train}、验证集$D_{\text{validation}}$和测试集D_{test}。将h作为摘要句,first作为原句子,在D_{train}和$D_{\text{validation}}$上训练方法提出的模型,在D_{test}上测试方法的性能。□

在Gigaword数据集上,Rush等[12]首次将神经网络用于生成式文本摘要,利用了"编码器—解码器"(encoder-decoder)模型,作者尝试了三种"编码"的方式:分别是"词袋"(bag-of-words)模型、卷积神经网络和基于"注意力"(attention)机制的方式。在Gigaword和DUC2004数据集上的实验结果表明,基于"注意力"机制的"编码器"效果最好。方法取得的效果使得基于深度神经网络的模型成为可能,后续工作大都基于此方法。

① https://catalog.ldc.upenn.edu/ldc2003t05

Chopra 等[13]同样利用“编码器—解码器”模型，在“解码器”中使用一种条件循环神经网络（Conditional RNN），在 Gigaword 和 DUC2004 数据集上的实验结果要优于 Rush 等的方法。

为进一步提升效果，Nallapati 等[14]引入了 TF-IDF、命名实体等语言学特征来强化句子中的关键信息。将这些特征显式地作为神经网络的输入，在 Gigaword 和 DUC2004 数据集上提高了生成式文本摘要效果。同时，作者在 CNN/Daily Mail 数据集上也进行了方法的性能测试。

Zhou 等[15]提出了一种“选择性编码”（selective encoding）模型，将文本摘要问题看成是一个序列标注的任务，建模的方法是基于一个已经“编码”好的句子，利用句子信息来判断句中的词是否重要，由此来构建一个输入句子中词的新的表示。在 Gigaword 数据集上得到了较已有工作更好的效果。

Cao 等[16]认为已有方法得到的结果虽然测评结果较高，但是不够可靠（faithful），往往无法直接用于实际应用。为此，提出了一种提升信息量的方法：利用 Stanford CoreNLP[17]提取“主谓宾”三元组作为输入。论文中的“编码器—解码器”模型包括两个“编码器”和一个“双注意力”机制的“解码器”，两个“编码器”分别用于句子本身和三元组的语义表示。在 Gigaword 数据集上的结果表明，该方法在提高“可靠性”的同时，准确率较已有方法也有所提升。

受传统的基于模板的生成式文本摘要的启发，Cao 等[18]提出了一种新的“端到端”的模型。将已有的摘要句看作是“软模板”（soft template），作为参考来指导摘要的生成。提出的模型包括检索（retrieving）、重排序（reranking）和重写（rewriting）三个模块，称之为 Re3Sum。

Gigaword 数据集的特点在于原句和摘要句都是单个句子，而在实际应用中，除了对单个句子生成摘要的情形之外，还存在对由多个句子组成的整篇文本生成摘要的情形。

4　CNN/Daily Mail

与 Gigaword 和部分 DUC/TAC 数据集只包含单句话的摘要不同，CNN/Daily Mail（简称 CNN/DM）作为单文本摘要语料库，每篇摘要包含多个摘要句。CNN/DM 最初是 Hermann 等[19]发布的机器阅读理解语料库。作者从美国有线新闻网（CNN）①和每日邮报网（Daily Mail）②中收集了约 100 万条新闻数据作为机器阅读理解语料库。在 CNN 和 Daily Mail 的新闻数据中，每篇新闻包括一条或者多条人工要点，将隐藏一个命名实体的要点作为填空题的问题，将新闻内容作为回答填空题的阅读文字。表 4 是语料库的详细统计信息。

表 4　Hermann 等[19]文献中 CNN/DM 数据规模

数据集		文本数量	问题数量	总文本数量
CNN	训练集	90 266	380 298	92 579
	验证集	1 220	3 924	
	测试集	1 093	3 198	
Daily Mail	训练集	196 961	879 450	219 506
	验证集	12 148	64 853	
	测试集	10 397	53 182	

Nallapati 等[14]进行简单改动，形成用于单文本生成式摘要的语料库。将每篇新闻的要点按原文中出现的顺序组成多句的摘要，每个要点看成是一个句子。表 5 给出了用于单文本摘要的 CNN/DM③数据集规模。

表 5　用于单文本摘要的 CNN/DM 数据集规模

训练集大小	286 817
验证集大小	13 368
测试集大小	11 487
训练集中平均摘要句子数	3.72

定义 3　CNN/DM 数据集上的文本摘要

给定文本集 D，包含 k 个文本摘要对 $<H,S>$，其中 H 表示新闻提要，S 表示新闻正文。将 D 分成三份，分别是训练集 D_{train}、验证集 $D_{\text{validation}}$ 和测试集 D_{test}。将 H 作为摘要句集合，S 作为原文本，在 D_{train} 和 $D_{\text{validation}}$ 上训练方法提出的模型，在 D_{test} 上进行方法的性能测试。□

在 CNN/DM 数据集上，See 等[20]认为传统的循环神经网络用于文本摘要，在执行序列数据计算时，会存在两个问题：一是摘要不能准确复制事实细节，二是存在多次重复同样内容。作者提出用指针生成网络（pointer-generator network）来解决问题一，利用汇聚（coverage）技术来解决问题二。这种方法在 CNN/DM 数据集上取得了较好的效果。

① https://edition.cnn.com/
② http://www.dailymail.co.uk/home/index.html
③ https://github.com/deepmind/rc-data

典型方法还包括Cheng和Lapata的方法[6]，即一种数据驱动的基于深度神经网络的摘要句抽取方法。这种方法面向单文本抽取式摘要任务，包括一个结合卷积神经网络、循环神经网络的“编码器”和基于“注意力”机制的“解码器”，也称为摘要提取器。基于卷积神经网络得到句子的表示，将句子的表示作为输入，基于循环神经网络得到文本的表示。

作者将CNN/DM数据集中的Daily Mail部分进行了转换，计算原文中句子与已有生成式摘要的匹配度，匹配度较高的句子作为抽取式摘要句。通过这种方式将已有数据集转换为用于抽取式摘要的数据集。利用转换后的数据集进行模型训练和测试，同时在DUC2002数据集上进行了方法测评，取得了更好的效果。

5 NYTAC

纽约时报标注数据集(New York Times Annotated Corpus，NYTAC)①包括了从1987年1月到2007年6月的《纽约时报》大约180万篇英文文章，其中65万篇文章包括人工摘要。NYTAC可用于文本摘要、信息检索和信息抽取等自然语言处理任务。

定义4 NYTAC数据集上的文本摘要

NYTAC数据集中的文本与CNN/DM数据集中的文本类似：每篇文本对应的摘要对$<H,S>$中，新闻提要H包含多句，S表示整篇新闻正文，只是来自不同的新闻源。因此在NYTAC数据集上的文本摘要问题定义同定义3。 □

在NYTAC数据集上的代表性方法是Durrett等[21]的方法，即一种用于单文本摘要的判别式模型，模型基于结构化支持向量机(structured SVM)。作者考虑了比以往方法更多的特征，通过丰富的稀疏特征来提取文本摘要。作者在两个数据集上进行方法训练和测试，从NYTAC中选取了3 000篇文本进行方法模型的训练，然后在英文修辞结构理论标注数据集(RST Discourse Treebank，RST-DT)②上进行测试。

6 Amazon SNAP Review Dataset

亚马逊在线评论数据集(Amazon SNAP Review Dataset，ASNAPR)③包括从1995年到2013年接近0.35亿用户的评论数据，每条评论数据包括用户ID、评论内容、评论摘要和评论时间等内容。由于ASNAPR是商品评论数据，因此都是短文本。ASNAPR数据集的特点是文本篇幅较短，常用于评论和情感的多文本摘要。

定义5 ASNAPR数据集上的文本摘要

给定数据集D，包含k组亚马逊英文在线评论数据(x,y,l)，其中x表示评论原文，y表示评论的摘要，l表示商品的情感标签。将D分成训练集D_{train}、验证集$D_{\text{validation}}$和测试集D_{test}三部分。从D_{train}和$D_{\text{validation}}$中学习评论原文到评论摘要及评论原文到情感标签的映射，在D_{test}上验证方法的有效性。

在ASNAPR数据集上，经典方法包括Ma等[22]的方法。作者认为文本摘要和情感分析都是提取文章中的主要内容，只是提取的层次不同，他们提出了一种分层式“端到端”模型，整合文本摘要和情感分类。模型包括一个摘要层(将源文本压缩成短句子)和一个情感分类层(给文本打一个情感类别标签)。这种分层结构会使两个任务彼此提升：通过摘要层压缩文本，情感分类器可以更加轻松地预测情感标签；同时文本摘要还能标记出重要和有信息的词，并移除对预测情感有害的冗余和误导性信息，提升文本摘要的性能。作者从ASNAPR中选取了部分数据(约110万条评论)，用到了每条评论中的摘要和情感标签元数据。

7 LCSTS

随着微博等社交媒体软件的普及，部分工作提出了面向社交媒体文本的文本摘要算法。由于中文社交媒体文本大都是短文本，具有篇幅较短、存在较多噪声等特点，传统的文本摘要方法在这类文本上往往效果较差。

LCSTS(large scale Chinese short text summarization dataset)④是Hu等[23]从新浪微博⑤获取的短文本新闻摘要数据库，规模超过200万。详细数据规模见表6。图1是一个数据样例，将中括号中

① https://catalog.ldc.upenn.edu/LDC2008T19/

② RST-DT是人工标注的篇章结构树，共包括385篇来自华尔街日报(Wall Street Journal，WSJ)的新闻文章，具体数据在https://catalog.ldc.upenn.edu/LDC2002T07。

③ http://snap.stanford.edu/data/web-Amazon.html

④ http://icrc.hitsz.edu.cn/Article/show/139.html

⑤ http://weibo.com/

的要点看成是后面一段文本新闻的摘要。

表 6　LCSTS 数据规模

数据集	数据规模
训练集大小	2 400 000
验证集大小	10 000
测试集大小	1 000

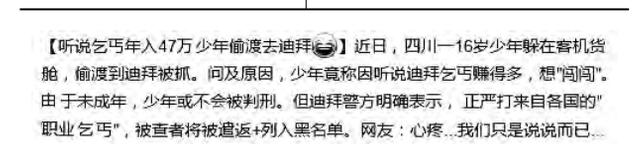
【听说乞丐年入47万 少年偷渡去迪拜】近日，四川一16岁少年躲在客机货舱，偷渡到迪拜被抓。问及原因，少年竟称因听说迪拜乞丐赚得多，想"闯闯"。由于未成年，少年或不会被判刑。但迪拜警方明确表示，正严打来自各国的"职业乞丐"，被查者将被遣返+列入黑名单。网友：心疼...我们只是说说而已...

图 1　LCSTS 数据样例

对于验证集和测试集，作者手工标注了正文和标题之间的相关性，相关性分值区间是[1，5]，分值越高表示越相关。LCSTS 数据集的特点是文本篇幅较短，并且存在噪声。

在发布 LCSTS 中文数据集的同时，作者提出了一种利用循环神经网络提取生成式摘要的方法，给出了在 LCSTS 数据集上的基准方法，后续相关工作都将该方法作为基准方法进行方法效果的比较。

定义 6　LCSTS 数据集上的文本摘要

给定文本集 D，包含 k 个中文短文本新闻摘要对$<h,S>$，其中 h 表示单句话的新闻提要，由新闻作者给出，S 表示短文本的新闻正文，由多个短句子组成。将 D 分成三份，分别是训练集 D_{train}、验证集 $D_{\text{validation}}$和测试集 D_{test}。将 h 作为摘要句，S 作为原文本，在 D_{train}和 $D_{\text{validation}}$上训练方法提出的模型，在 D_{test}上测试方法的性能。□

Ma 等[24]提出了一种面向中文社交媒体短文本摘要的方法。这是一种基于深度学习的抽取式摘要方法，他们提出的模型基于循环神经网络的"编码器—解码器"和"注意力"机制。这种方法在 LCSTS 数据集上的效果较 Hu 等[23]的方法有所提升。

8　NLPCC

自然语言处理与中文计算会议(CCF Conference on Natural Language Processing & Chinese Computing，NLPCC)是由中国计算机学会(CCF)举办的自然语言文本测评会议，包括文本摘要、情感分析、自动问答等任务。NLPCC 于 2012 年开始举办，每年一届。在过去的 NLPCC 测评任务中，NLPCC 2015①、NLPCC 2017②和 NLPCC 2018③包括文本摘要任务，且都是单文本抽取式摘要。NLPCC 数据集的特点是新闻文本不分领域、不分类型，篇幅相对较长。

定义 7　NLPCC 数据集上的文本摘要

给定文本集 D，包含 k 个中文新闻类文本摘要对$<H,S>$，其中 H 表示新闻提要，由新闻作者给出，S 表示新闻正文。通常情况下，H 包含多个摘要句，S 是整篇新闻文本，新闻源自头条新闻、财经网等中文新闻网站。学习由 S 到 H 的映射，在测试集 D_{test}($D_{\text{test}}\subseteq D$)上测试方法的有效性。

在 NLPCC 数据集上，与经典图模型的方法不同，莫鹏等[25]提出了一种基于超图的文本摘要和关键词生成方法。将句子作为超边(hyperedge)，将词作为节点(vertice)构建超图(hypergraph)。利用超图中句子与词之间的高阶信息来生成摘要和关键词。方法在 NLPCC2015 数据集上取得较好效果。

Xu 等[26]针对已有的利用极大似然估计来优化的生成式摘要模型存在的准确率低的问题，提出了一种基于对抗增强学习的中文文本摘要方法，提升了基于深度学习方法在中文文本摘要上的准确率。方法在 LCSTS 和 NLPCC2015 数据集上进行了测评。

LCSTS 和 NLPCC 是目前面向中文的文本摘要公用数据集。可以作为未来更多的面向中文的文本摘要方法的训练和测试数据集，同时，可以在 LCSTS 数据集上验证已有面向英文的基于深度学习的方法对中文文本摘要的适用性。

9　自建数据集及其对应方法

由于文本摘要公用数据集较少，除了上述在公用数据集上进行训练和测试的工作之外，还有大量自建数据集的方法。对于用户自建数据集的文本摘要任务，常用方法可分为基于统计的方法、基于图模型的方法、基于词法链的方法、基于篇章结构的方法和基于机器学习的方法，本节对每种类别的几种典型方法中作者自建的数据集和方法进行总结。

9.1　基于统计的方法

基于统计的方法通过一些统计特征来辅助摘要

① http://tcci.ccf.org.cn/conference/2015/pages/page05_evadata.html

② http://tcci.ccf.org.cn/conference/2017/taskdata.php

③ http://tcci.ccf.org.cn/conference/2018/taskdata.php

句的选取，常用的特征包括句子所在的位置、TF-IDF、n-gram等。这种方法不需要额外的语言学知识和复杂的自然语言处理技术，实现较为简单。已有方法的主要区别在于特征类型和特征数量的选取。

Ko和Seo[27]提出一种基于上下文特征和统计特征的摘要句提取方法，将每两个相邻的句子合并为一个二元语言模型伪句子(Bi-Gram pseudo sentence，BGPS)，BGPS包含比单个句子更多的特征。根据统计方法对BGPS进行重要程度打分，选取分值较高的BGPS对应的句子作为摘要句。

对于单文本摘要，作者用到了韩国研究与发展信息中心的(KOrea Research and Development Information Center，KORDIC)数据，包括841篇新闻文章，手工标注压缩率为10%和30%的摘要句；对于多文本摘要，作者选取了5个主题共55篇新闻文章自建数据集，手工标注摘要句。方法在两个数据集上都取得了较好的结果。

基于统计的文本摘要方法较为直观，抽取的特征相对简单，因此方法较易实现，但准确率较低。这类方法同样适用于中文文本摘要任务。

9.2 基于图模型的方法

部分基于图模型的方法也在自建数据集上进行了测试。Hu等[28]认为，对于Web文本来说，读者的评论对于文本摘要等信息检索任务是有价值的。提出的方法不仅考虑文本内容本身，还将读者的评论信息加入文本摘要抽取中，将评论作为节点，将评论之间的关系作为边，利用图模型对评论的重要程度进行打分。他们提出了两种文本摘要方法：一种通过评论中的关键词来对候选摘要句进行打分；另一种将原文本和评论组成一个“伪文本”，对该“伪文本”进行摘要句的抽取。作者从两大英文博客网站Cosmic Variance① 和IEBlog② 中分别获取了50篇文章作为实验语料，4个标注者人工标注摘要句。由于他们的方法结合了文章的评论，因此要求标注者分别读取博文和评论后再标注出摘要句。

Lin等[29]提出了一种基于情感信息的PageRank多文本情感摘要方法，作者同时考虑了情感和主题这两方面的信息，提升了算法的准确率。由于针对中文文本情感摘要的研究较少，公共语料缺乏，作者从亚马逊中文网③中收集了15个产品的评论语料，每个产品包括200条评论，自建了包括15个主题的多文本摘要数据集。挑选出3名标注者从每个主题的评论中抽取48个句子作为该主题的摘要句。

9.3 基于词法链的方法

词法链(lexical chain)[30]是一种描述篇章衔接性的理论体系，常用于文本摘要、情感分析等自然语言处理应用中。Chen等[31]首次将词法链方法应用到中文文本摘要中，提出了一种基于词法链的中文文本摘要方法。首先利用HowNet作为词法链构建知识库，然后识别强词法链，最后基于启发式规则选取摘要句。从互联网上随机选取100篇中文新闻语料自建数据集。对每篇文本，标注压缩率分别为10%和20%的摘要句。

Yu等[32]在词法链的基础上，结合一些结构特征，提出了一种基于词法链和结构特征的中文文本摘要方法。同样利用HowNet构建词法链，结构特征包括句子的位置(如是否是首句)等。利用词法链特征和结构特征进行加权对句子重要程度进行打分，选取摘要句。作者从互联网上随机选取50篇不同类别的中文新闻语料自建数据集。对每篇文本，标注压缩率分别为10%，20%和30%的摘要句。

Wu等[33]提出了个性化Web新闻的过滤和摘要系统PNFS。PNFS的新闻摘要是总结并提取能够刻画新闻主题的关键词。关键词的提取是利用基于词法链的方法[34]，利用词之间的语义相关性进行词义消歧并构建词法链。构建的关键词一方面可以提供给用户一种精简的阅读形式，节省阅读时间，另一方面可以用于构建用户兴趣模型。作者从163新闻网站④获取了120篇中文新闻文章自建数据集，然后利用ICTCLAS⑤进行中文分词。

传统词法链主要由名词和名词短语构成，缺少了动词等所包含的语义信息。Hou等[35]提出了全息词法链(holographic lexical chain)并将其应用到中文的单文本摘要中。全息词法链包括名词、动词和形容词三类词法链，这三类词法链包括了文章的主要语义信息，因此称为全息词法链。根据句子中包含全息词法链中词的特征，利用Logistic回归、支持向量机等机器学习方法学习摘要句。作者从互联网上选取159篇外贸领域中文新闻语料自建数据集。对每篇文本，人工标注摘要句，进行模型的训练

① http://blogs.discovermagazine.com/cosmicvariance #.WyyfqadLjIU/

② https://blogs.msdn.microsoft.com/ie/

③ https://www.amazon.cn

④ http://news.163.com

⑤ http://ictclas.nlpir.org

和测试。

9.4　基于篇章结构的方法

基于篇章结构的方法是利用篇章结构信息指导文本摘要的生成，典型方法包括 Cheng 等[36]提出的中文 Web 文本自动摘要方法。作者首先分析段落之间的语义关联，将语义相近的段落合并，划分出主题层次，进而得到篇章结构。在篇章结构的指导下，使用统计方法，结合一些启发式规则进行关键词和关键句子的提取，最终生成中文 Web 文本的摘要。作者从新浪①、计算机世界报②等网站获取了 IT 类文章，随机选取了 228 篇文本自建语料库。人工对其理解和分析，得到文本包含的主题及子主题、关键词。作者认为此方法人工分析工作量大，仅能选取少量文本进行方法验证。

这类方法利用了篇章结构的信息，可以得到结构上连贯、准确率相对较高的结果。但是模型复杂度较高，并且缺少规模较大的篇章结构数据集来进行机器学习模型的训练和测试，已有方法都是在自建数据集上进行提出方法的测评。

9.5　基于机器学习的方法

大部分基于机器学习的文本摘要方法是有监督的方法，即需要有标注的训练集和测试集。Hu 等[37]提出了一种基于主题的中文单文本摘要方法。首先通过段落聚类发现文本所反映的主题，然后从每一个主题中选取与主题语义相关性最大的一句话作为摘要句，最后根据选取的摘要句在原文本中的顺序组成最终的摘要。随机选取 200 篇不同类型的中文文章自建语料库，进行提出方法的效果评估。

Baumel 等[38]提出了一种基于 LDA 主题模型(topic model)[39]的新型文本摘要任务：面向查询的更新摘要方法(query-chain focused summarization)。更新摘要是假设已经提取出部分摘要句，在避免冗余的前提下，将新内容加入摘要中；而面向查询的摘要是提取出与查询相关的重要句子作为摘要句。结合这两种任务，将用户多次查询的结果生成更新摘要。也就是说，用户的第 n 条查询语句得到的结果要在前 $n-1$ 条查询语句结果摘要基础上进行更新摘要，最终生成的摘要是所有查询语句得到的结果的摘要。

选取来自“消费者健康(consumer health)”领域的语料自建数据集。针对面向查询的摘要，首先从 PubMed③中选取包括“气喘(asthma)”、“肺癌(lung cancer)”、“肥胖症(obesity)”和“老年痴呆(alzheimer)”四个关键词的查询语句，然后从英文 Wiki④，WebMD⑤等网上资源中获取与查询语句相关的文本，找医学专业学生标注文本摘要。最终得到人工标注摘要 186 篇，作为训练和测试数据集。

庞超等[40]结合循环神经网络的“编码器—解码器”结构和基于分类的结构，提出一种理解式文本摘要方法。同时，在“编码器—解码器”结构中使用了“注意力”机制，提升了模型对于文本内容的表达能力，进一步提升了文本摘要的性能。作者从中国新闻网⑥获取新闻内容，自建语料库。共包括 120 万条语料，其中训练集 90 万条，验证集 20 万条，测试集 10 万条。每条语料包括新闻标题、新闻内容和新闻类别(分时政、国际、社会、财经、金融、汽车、能源、文化、娱乐、体育、健康共 11 个类别)。

10　经典算法和最新方法用到的数据集

本节调研了 ACL、AAAI、EMNLP、ICJNLP 和 COLING 等自然语言处理相关国际会议和部分期刊中的文本摘要方法相关文献，表 7 总结了经典算法和最新方法相关文献中用到的数据集。

从表 7 可知，经典算法和最新方法大都是基于深度学习的方法，也包括 LexRank、TextRank 等经典方法。

已有工作提出面向中英文文本摘要的通用方法，Lin 等[43]的工作分别在 LCSTS 和 Gigaword 数据集上进行了测评。

当前的深度神经网络模型中，最常用的数据集是 Gigaword、CNN/DM 和 LCSTS 等大规模数据集。文本摘要数据集 DUC/TAC 的规模较小，但不适用于深度神经网络模型的训练，已有深度神经网络模型通常在大规模数据集上进行训练和测试。模型训练完成后，DUC/TAC 数据集也是重要的测评标准。因此，DUC/TAC 也是一种常用的文本摘要方法测评数据集。

---

① http://www.sina.com.cn

② http://www.ccw.com.cn

③ 医学、生命科学领域的科研文献检索数据库，https://www.ncbi.nlm.nih.gov/pmc/

④ https://en.wikipedia.org/wiki/Wiki

⑤ https://www.webmd.com

⑥ http://www.chinanews.com

表 7 文献用到的数据集总结

文献标题	文献来源	数据集
A unified model for extractive and abstractive summarization using inconsistency loss[41]	ACL2018	CNN/DM
Extractive summarization with SWAP-NET: Sentences and words from alternating pointer networks[42]	ACL2018	CNN/DM
Global encoding for abstractive summarization[43]	ACL2018	LCSTS,Gigaword
Autoencoder as assistant supervisor: Improving text representation for Chinese social media text summarization[44]	ACL2018	LCSTS
Neural document summarization by jointly learning to score and select sentences[45]	ACL2018	CNN/DM
Retrieve,rerank and rewrite: Soft template based neural summarization[18]	ACL2018	Gigaword
Learning to extract coherent summary via deep reinforcement learning[46]	AAAI2018	CNN/DM
Faithful to the original: Fact-aware neural abstractive summarization[16]	AAAI2018	Gigaword
Sequential copying networks[47]	AAAI2018	Gigaword
Generative adversarial network for abstractive text summarization[48]	AAAI2018	CNN/DM
Unity in diversity: Learning distributed heterogeneous sentence representation for extractive summarization[49]	AAAI2018	CNN/DM,DUC2001,DUC2002,DUC2004
Selective encoding for abstractive sentence summarization[15]	ACL2017	Gigaword,DUC2004
Supervised learning of aautomatic pyramid for optimization-based multi-document summarization[50]	ACL2017	TAC2009
Oracle summaries of compressive summarization[51]	ACL2017	DUC2004
Improving semantic relevance for sequence-to-sequence learning of Chinese social mediatext summarization[24]	ACL2017	LCSTS
Get to the point: Summarization with pointer-generator networks[20]	ACL2017	CNN/DM
Extract with order for coherent multi-document summarization[52]	Text Graphs-11	DUC2004
Revisiting the centroid-based method: A strong baseline for multi-document summarization[53]	New Frontiers in Summarization 2017	DUC2004
SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents[54]	AAAI2017	CNN/DM,DUC2002
Improving multi-document summarization via text classification[55]	AAAI2017	DUC2001,DUC2002,DUC2004
Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction[56]	ACL2007	DUC2002
Affinity-preserving random walk for multi-document summarization[57]	EMNLP2017	DUC2003,DUC2004

续表

文献标题	文献来源	数据集
Cascaded attention based unsupervised information distillation for compressive summarization[58]	EMNLP2017	DUC2006，DUC2007
Deep recurrent generative decoder for abstractive text summarization[59]	EMNLP2017	DUC2004，Gigaword，LCSTS
Extractive summarization using multi-task learning with document classification[60]	EMNLP2017	NIKKEI Financial Report Corpus①，NYTAC
Generating coherent summaries of scientific articles using coherence patterns[61]	EMNLP2016	PLOS Medicine②，DUC2002
A neural attention model for abstractive sentence summarization[12]	EMNLP2015	Gigaword，DUC2003，DUC2004
Cascaded filtering for topic-driven multi-document summarization[62]	DUC2007	DUC2007
Textrank：Bringing order into text[7]	EMNLP2004	DUC2002
Abstractive multi-document summarization by partial tree extraction，recombination and linearization[63]	IJCNLP2017	TAC2011，DUC2004，DUC2005
Towards abstractive multi-document summarization using submodular function-based framework，sentence compression and merging[64]	IJCNLP2017	DUC2004，DUC2007
A general optimization framework for multi-document summarization using genetic algorithms and swarm intelligence[65]	COLING2016	DUC2002，DUC2003
Exploring text links for coherent multi-document summarization[66]	COLING2016	DUC2004
Abstractive news summarization based on event semantic link network[67]	COLING2016	DUC2006，DUC2007
Extractive summarization using supervised and semi-supervised learning[68]	COLING2008	DUC2001
Coherent narrative summarization with a cognitive model[69]	Computer Speech & Language	DUC2001，DUC2002
Event-based extractive summarization[70]	Text Summarization Branches Out	DUC2001
A scalable global model for summarization[10]	Workshop on ILP for NLP	TAC2008
Lexrank：Graph-based lexical centrality as salience in text summarization[1]	Journal of AI Research	DUC2003，DUC2004
Multi-document summarization using bipartite graphs[71]	TextGraphs-9	DUC2005，DUC2006，DUC2007
A study of global inference algorithms in multi-document summarization[72]	ECIR2007	DUC2002
Multi-topic based query-oriented summarization[73]	SIAM Conference on DM 2009	DUC2005，Epinions③

① 日文语料库。日本上市公司的财务报告语料，日本经济新闻网 https://www.nikkei.com 有 3 911 篇对这些报告的摘要。

② 共包括 50 篇科技文章，每篇文章包括编辑写的人工摘要。

③ 从网站 www.epinions.com 爬取的英文产品评论数据集，共包括 44 个不同商品的评论。

11 经典方法在数据集上的实验效果分析

为了对比经典方法在数据集上的实验效果，本节以Gigaword数据集为例，分析对比了如下7种单文本生成式文本摘要方法在Gigaword数据集的训练集上进行模型训练，在测试集上进行测试的结果。

ABS：Rush等[12]的基于“注意力”机制的“编码器”和基于标准前馈神经网络语言模型(NNLM)的“解码器”。

ABS+：Rush等[12]在ABS的基础上进行了模型改进，利用DUC 2003数据集进一步调整了参数。

Luong-NMT：Chopra等[13]在ABS和ABS+基础上进行了改进，同样利用了“编码器—解码器”模型，只是在“解码器”中使用了一种条件循环神经网络。

Feats2s：Nallapati等[14]在ABS+和Luong-NMT的基础上，引入了传统的TF-IDF、命名实体等语言学特征作为神经网络的输入。

SeqCopyNet：Zhou等[15]提出的“选择性编码”模型，基于一个已经“编码”好的句子，利用句子信息来判断句中的词是否重要，由此来构建一个输入句子中词的新的表示。

FTSum：Cao等[16]提出的提升信息量的“编码器—解码器”模型，两个“编码器”分别用于句子本身和“主谓宾”结构三元组的语义表示。

Re3Sum：Cao等[18]提出的新的“端到端”的模型，将已有的摘要句看作是“软模板”(soft template)，作为参考来指导摘要的生成。

表8是各种经典模型在Gigaword数据集上的实验效果，其中评估标准采用ROUGE[74]，一种通用的文本摘要评估标准。ROUGE计算模型输出的摘要与参考摘要之间的一元词、二元词、三元词及最长公共子串(longest common subsequence，LCS)等字符串的重合度。单文本摘要中常用的有ROUGE-1、ROUGE-2和ROUGE-L，分别表示模型输出的摘要和参考摘要的一元词、二元词和LCS之间的重合度，本文也采用了这三种标准。

从实验效果看，在大规模训练数据上，基于“注意力”机制的循环神经网络模型体现出了在单文本生成式文本摘要方面的有效性，在引入了传统的人工语义特征后，效果进一步提升。为了进一步提升生成摘要的质量，已有方法在网络结构及信息输入上进行了改进。例如，“SeqCopyNet”提出了选择性门网络，可以选择输入句子中的重要部分。“FTSum”引入了“主谓宾”结构，在ROUGE-1指标上取得了当前最好的结果。Re3Sum受传统的基于模板的生成式摘要的启发，将已有的摘要句作为参考来指导摘要的生成，在ROUGE-2和ROUGE-L这两个指标上都取得了最好的效果。

表8 经典方法在Gigaword上的实验效果

模型	实验效果		
	ROUGE-1	ROUGE-2	ROUGE-L
ABS	29.55	11.32	26.42
ABS+	29.76	11.88	26.96
Luong-NMT	33.10	14.45	30.71
Feats2s	32.67	15.59	30.64
SeqCopyNet	35.93	17.51	33.35
FTSum	**37.27**	17.65	34.24
Re3Sum	37.04	**19.03**	**34.46**

12 结论

在文本摘要领域，目前已有多个公用数据集可用于方法的训练、验证和测试。通过对常用数据集的分析，可以得到如下结论：

(1) 英文数据集较多，既包括百篇规模的DUC/TAC数据集，可以用于单文本摘要、多文本摘要等多种任务，又包括Gigaword和CNN/DM等大规模数据集。中文数据集较少，目前中文只有LCSTS和NLPCC，并且LCSTS是短文本数据集，NLPCC规模较小，不适用于神经网络方法的训练。因此，缺少大规模中文长文本数据集。

(2) 已有数据集中，除了DUC/TAC数据集可用于多文本摘要任务之外，其他数据集只适用于单文本摘要任务。

(3) 就摘要方式来说，大部分数据集只适用于生成式摘要方法的训练和测试，只有CNN/DM和DUC2002可用于抽取式摘要任务。

(4) 随着文本数量的激增，各领域对文本摘要的需求也越来越多。已有数据集中，除ASNAPR和TAC2014，其余都是新闻类文本。因此，未来应有更多其他领域的文本摘要数据集被提出。

从提出的文本摘要方法来看，除了已有的基于

统计的方法、基于图模型的方法和基于传统机器学习的方法之外，随着对神经网络和深度学习的研究不断深入，越来越多的工作提出了基于神经网络和深度学习的方法。但由于深度学习模型相对复杂，待学习参数较多，因此需要在大规模数据集上进行模型训练，这类方法对于数据集的规模要求较高。

对于 Gigaword 和 LCSTS 等大规模数据集，虽然在这些数据集上训练出的模型显示出较好的效果，但是这些方法是数据驱动的，对于数据的依赖性较强。未来研究中，不依赖训练数据特点的通用方法将更具实用性和可扩展性。

由于公用数据集较少，并且不同的任务需要有不同的数据集。对于一些特定任务（例如，对于评论的文本摘要，基于篇章结构的文本摘要）的公用数据集更少。部分面向中英文的文本摘要方法通过自建数据集进行方法的训练和测试，尤其是面向中文的文本摘要方法。

随着机器学习和深度学习技术的不断深入，对高质量标注数据的需求和依赖也越来越高。不单单是对文本摘要任务，对于其他自然语言处理任务如命名实体识别、情感分析，甚至计算机视觉领域，标注数据也是不可或缺的。在缺少公用数据集的情况下，除了在自建数据集上进行性能测试之外，半自动的数据集构建方法[75]会成为一个新的研究方向。

参考文献

[1] Erkan G, Radev D R. Lexrank: Graph-based lexical centrality as salience in text summariza-tion[J]. Journal of Artificial Intelligence Research. 2004,22: 457-479.

[2] Gambhir M,Gupta V. Recent automatic text summarization techniques: a survey[J]. Artificial Intelligence Review,2017,47(1): 1-66.

[3] Nenkova A, McKeown K. Automatic summarization [J]. Foundations and Trends in Information Retrieval, 2011,5(2-3): 103-233.

[4] Nenkova A,McKeown K. A survey of text summarization techniques [M]. Mining Text Data. Boston: Springer,2012: 43-76.

[5] Baralis E,Cagliero L,Fiori A,et al. Mwi-sum: A multilingual summarizer based on frequent weighted itemsets[J]. ACM Transactions on Information Systems (TOIS),2015,34(1): 5.

[6] Cheng J,Lapata M. Neural summarization by extracting sentences and words[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics,2016: 484-494.

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

[8] Page L,Brin S,Motwani R,et al. The PageRank citation ranking: Bringing order to the web[R]. Stanford InfoLab,1999.

[9] Baralis E,Cagliero L,Mahoto N,et al. GRAPHSUM: Discovering correlations among multiple terms for graph-based summarization[J]. Information Sciences, 2013,249: 96-109.

[10] Gillick D,Favre B. A scalable global model for summarization[C]//Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing. Association for Computational Linguistics, 2009: 10-18.

[11] Fattah M A. A hybrid machine learning model for multi-document summarization[J]. Applied in-telligence,2014,40(4): 592-600.

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

[13] Chopra S,Auli M,Rush A M. Abstractive sentence summarization with attentive recurrent neural networks[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 93-98.

[14] Nallapati R,Zhou B,dos Santos C,et al. Abstractive Text Summarization using Se-quence-to-sequence RNNs and Beyond [C]//Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. 2016: 280-290.

[15] Zhou Q,Yang N,Wei F,et al. Selective Encoding for Abstractive Sentence Summarization [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1095-1104.

[16] Cao Z,Wei F,Li W,et al. Faithful to the original: Fact aware neural abstractive summarization[C] // Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018.

[17] Manning C,Surdeanu M,Bauer J,et al. The stanford CoreNLP natural language processing toolkit[C]// Proceedings of 52nd annual meeting of the association

for computational linguistics: system demonstrations. 2014: 55-60.

[18] Cao Z, Li W, Li S, et al. Retrieve, rerank and rewrite: Soft template based neural summarization[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 152-161.

[19] Hermann K M, Kocisky T, Grefenstette E, et al. Teaching machines to read and comprehend[C]//Proceedings of the 29th Annual Conference on Neural Information Processing Systems. 2015: 1693-1701.

[20] See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 1073-1083.

[21] Durrett G, Berg Kirkpatrick T, Klein D. Learning-based Single-document summarization with compression and anaphoricity constraints[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 1998-2008.

[22] Ma S, Sun X, Lin J, et al. A hierarchical End-to-End model for jointly improving text summarization and sentiment classification [C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018.

[23] Hu B, Chen Q, Zhu F. LCSTS: A large scale Chinese short text summarization dataset[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1967-1972.

[24] Ma S, Sun X, Xu J, et al. Improving semantic relevance for Sequence-to-Sequence learning of Chinese social media text summarization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 635-640.

[25] 莫鹏，胡珀，黄湘冀，等. 基于超图的文本摘要与关键词协同抽取研究[J]. 中文信息学报，2015，29(06)：135-140.

[26] Xu H, Cao Y, Shang Y, et al. Adversarial reinforcement learning for Chinese text summarization[C]//Proceedings of the 18th International Conference on Computational Science. 2018: 519-532.

[27] Ko Y, Seo J. An effective sentence-extraction technique using contextual information and statistical approaches for text summarization[J]. Pattern Recognition Letters, 2008, 29(9): 1366-1371.

[28] Hu M, Sun A, Lim E P. Comments-oriented document summarization: understanding documents with readers' feedback[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2008: 291-298.

[29] 林莉媛，王中卿，李寿山，等. 基于 PageRank 的中文多文档文本情感摘要[J]. 中文信息学报，2014，28(2)：85-90.

[30] Barzilay R, Elhadad M. Using lexical chains for text summarization[J]. Advances in automatic text summarization, 1999: 111-121.

[31] Chen Y, Wang X, Guan Y. Automatic text summarization based on lexical chains[C]//Proceedings of the 1st International Conference on Natural Computation. Springer, 2005: 947-951.

[32] Yu L, Ma J, Ren F, et al. Automatic text summarization based on lexical chains and structural features [C]//Proceedings of the 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2007, 2: 574-578.

[33] Wu X, Xie F, Wu G, et al. PNFS: personalized web news filtering and summarization[J]. International Journal on Artificial Intelligence Tools, 2013, 22(05): 1360007.

[34] Ercan G, Cicekli I. Using lexical chains for keyword extraction[J]. Information Processing & Management, 2007, 43(6): 1705-1714.

[35] Hou S, Huang Y, Fei C, et al. Holographic Lexical Chain and Its Application in Chinese Text Summarization[C]//Proceedings of the 2nd Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data. Springer, 2017: 266-281.

[36] 王继成，武港山，周源远，等. 一种篇章结构指导的中文 Web 文档自动摘要方法[J]. 计算机研究与发展，2003，3：398-405.

[37] Hu P, He T, Ji D. Chinese text summarization based on thematic area detection[C]//Proceedings of the ACL-04 Workshop: Text Summarization Branches Out Text Summarization Branches Out, 2004: 112-119.

[38] Baumel T, Cohen R, Elhadad M. Query-chain focused summarization[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014: 913-922.

[39] Blei D M. Probabilistic topic models[J]. Communications of the ACM, 2012, 55(4): 77-84.

[40] 庞超，尹传环. 基于分类的中文文本摘要方法[J]. 计算机科学，2018，45(01)：144-147，178.

[41] Hsu W T, Lin C K, Lee M Y, et al. A unified model for extractive and abstractive summarization using inconsistency loss[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 132-141.

[42] Jadhav A, Rajan V. Extractive summarization with SWAP-NET: Sentences and words from alternating pointer networks[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 142-151.

[43] Lin J, Sun X, Ma S, et al. Global encoding for abstractive summarization[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 163-169.

[44] Ma S, Sun X, Lin J, et al. Autoencoder as assistant supervisor: improving text representation for Chinese social media text summarization[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 725-731.

[45] Zhou Q, Yang N, Wei F, 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. 2018: 654-663.

[46] Wu Y, Hu B. Learning to extract coherent summary via deep reinforcement learning[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018.

[47] Zhou Q, Yang N, Wei F, et al. Sequential copying networks[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018.

[48] Liu L, Lu Y, Yang M, et al. Generative adversarial network for abstractive text summarization[C] // Proceedings of 32nd AAAI Conference on Artificial Intelligence. 2018.

[49] Singh A K, Gupta M, Varma V. Unity in Diversity: Learning distributed heterogeneous sentence representation for extractive summarization[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018.

[50] Peyrard M, Eckle-Kohler J. Supervised learning of automatic pyramid for optimization-based multi-document summarization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1084-1094.

[51] Hirao T, Nishino M, Nagata M. Oracle summaries of compressive summarization[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 275-280.

[52] Nayeem M T, Chali Y. Extract with order for coherent multi-document summarization[C]//Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing. 2017: 51-56.

[53] Ghalandari D G. Revisiting the centroid-based method: A strong baseline for multi-document dummarization[C]//Proceedings of the EMNLP 2017 Workshop on New Frontiers in Summarization. 2017: 85-90.

[54] Nallapati R, Zhai F, Zhou B. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. 2017: 3075-3081.

[55] Cao Z, Li W, Li S, et al. Improving Multi-document summarization via text classification[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. 2017: 3053-3059.

[56] Wan X, Yang J, Xiao J. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extrac-tion[C]//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. 2007: 552-559.

[57] Wang K, Liu T, Sui Z, et al. Affinity preserving random walk for multi-document summarization[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 210-220.

[58] Li P, Lam W, Bing L, et al. Cascaded attention based unsupervised information distillation for compressive summarization[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2081-2090.

[59] Li P, Lam W, Bing L, et al. Deep recurrent generative decoder for abstractive text summarization[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2091-2100.

[60] Isonuma M, Fujino T, Mori J, et al. Extractive summarization using multi-task learning with document classification[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2101-2110.

[61] Parveen D, Mesgar M, Strube M. Generating coherent summaries of scientific articles using coherence patterns[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.

2016：772-783.

[62] Filippova K，Mieskes M，Nastase V，et al. Cascaded filtering for topicdriven multi-document summarization[C]//Proceedings of the 7th Document Understanding Conference. 2007：26-27.

[63] Kurisinkel L J，Zhang Y，Varma V. Abstractive Multi-document summarization by partial tree extraction，recombination and linearization[C]//Proceedings of the 8th International Joint Conference on Natural Language Processing. 2017：812-821.

[64] Chali Y，Tanvee M，Nayeem M T. Towards abstractive Multi-document summarization using submodular function-based framework，sentence compression and merging[C]//Proceedings of the 8th International Joint Conference on Natural Language Processing. 2017：418-424.

[65] Peyrard M，Eckle-Kohler J. A general optimization framework for Multi-document summarization using genetic algorithms and swarm intelligence[C]//Proceedings of the 26th International Conference on Computational Linguistics：Technical Papers. 2016：247-257.

[66] Wang X，Nishino M，Hirao T，et al. Exploring text links for coherent multi-document summarization[C]//Proceedings of the 26th International Conference on Computational Linguistics：Technical Papers. 2016：213-223.

[67] Li W，He L，Zhuge H. Abstractive news summarization based on event semantic link network[C]//Proceedings of the 26th International Conference on Computational Linguistics：Technical Papers. 2016：236-246.

[68] Wong K F，Wu M，Li W. Extractive summarization using supervised and semi-supervised learning[C]//Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics，2008：985-992.

[69] Zhang R，Li W，Liu N，et al. Coherent narrative summarization with a cognitive model[J]. Computer Speech & Language，2016，35：134-160.

[70] Filatova E，Hatzivassiloglou V. Event-based extractive summarization[C]//Proceedings of Text Summarization Branches Out，2004.

[71] Parveen D，Strube M. Multi-document summarization using bipartite graphs[C]//Proceedings of TextGraphs-9：the workshop on Graph-based Methods for Natural Language Processing. 2014：15-24.

[72] McDonald R. A study of global inference algorithms in multi-document summarization[C]//Proceedings of the 29th European Conference on Information Retrieval. Berlin：Springer，Heidelberg，2007：557-564.

[73] Tang J，Yao L，Chen D. Multi-topic based query-oriented summarization[C]//Proceedings of the 2009 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics，2009：1148-1159.

[74] Lin C Y. Rouge：A package for automatic evaluation of summaries[C]//Proceedings of the ACL-04 Workshop：Text Summarization Branches Out，2004.

[75] Yang Y S，Zhang M，Chen W，et al. Adversarial Learning for Chinese NER from Crowd Annotations[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. 2018.

侯圣峦(1989—)，**通信作者，博士研究生，主要研究领域为自然语言处理、文本摘要、机器学习。**
E-mail：houshengluan1989@163.com

费超群(1992—)，**博士研究生，主要研究领域为自然语言处理、机器学习。**
E-mail：feichaoqun15@mails.ucas.ac.cn

张书涵(1991—)，**博士研究生，主要研究领域为自然语言处理、机器学习。**
E-mail：zhangshuhan@ict.ac.cn