

文本摘要研究进展与趋势

明拓思宇, 陈鸿昶

(国家数字交换系统工程技术研究中心, 河南 郑州 450002)

摘要: 随着互联网上的信息呈爆炸式增长, 如何从海量信息中提取有用信息成了一个关键的技术问题。文本摘要技术能够从大数据中压缩提炼出精炼简洁的文档信息, 有效降低用户的信息过载问题, 成为研究热点。分类整理分析了近些年来国内外的文本摘要方法及其具体实现, 将传统方法和深度学习摘要方法的优缺点进行了对比分析, 并对今后的研究方向进行了合理展望。

关键词: 大数据; 文本摘要; 机器学习; 传统方法; 深度学习

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2018048

Research progress and trend of text summarization

MING Tuosiyu, CHEN Hongchang

National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002, China

Abstract: With the explosive growth of information on the Internet, how to extract useful information from massive information has become a key technical issue. The text summarization technology can compress and extract refined and concise document information from big data, effectively reducing the user information overload problem, and it has become a research hotspot. The domestic and foreign text summarization methods and their concrete realization in recent years were analyzed, the advantages and disadvantages between traditional methods and deep learning summary methods were compared, and a reasonable outlook for future research directions was made.

Key words: big data, text summarization, machine learning, traditional methods, deep learning

1 引言

文本摘要是一种从一个或多个信息源中抽取关键信息的方法, 它帮助用户节省了大量时间, 用户可以从摘要获取到文本的所有关键信息点而无需阅读整个文档。文本摘要是一个典型的文本压缩任务。随着信息化时代的到来, 人们变得越来越依赖互联网获取所需要的信息。但是随着互联网的发展, 其上的信息呈现爆炸式增长,

如何有效地从海量信息中筛选出所需的有用信息成了关键性的技术问题。因为涉及深层次的自然语言处理的能力, 所以一直以来它都是个任务难点。自动文本摘要技术对文档信息进行有效的压缩提炼, 帮助用户从海量信息中检索出所需的相关信息, 避免通过搜索引擎来检索可能产生过多冗余片面信息的问题, 有效地解决了信息过载的问题。

文本摘要有多种分类方法, 按照摘要方法划

收稿日期: 2018-05-02; 修回日期: 2018-06-01

通信作者: 明拓思宇, 1139446336@qq.com

基金项目: 国家自然科学基金青年科学资助项目 (No.61601513)

Foundation Item: The National Natural Science Foundation of Youth Science (No.61601513)

分可以分为抽取式摘要^[1]方法和生成式摘要^[2]方法。抽取式摘要方法通过抽取文档中的句子生成摘要,通过对文档中句子的得分进行计算,得分代表重要性程度,得分越高代表句子越重要,然后通过依次选取得分最高的若干个句子组成摘要,摘要的长度取决于压缩率。生成式摘要方法不是单纯地利用原文档中的单词或短语组成摘要,而是从原文档中获取主要思想后以不同的表达方式将其表达出来。生成式摘要方法为了传达原文档的主要观点,可以重复使用原文档中的短语和语句,但总体上来说,摘要需要用作者自己的话来概括表达。生成式摘要方法需要利用自然语言理解技术对原文档进行语法语义的分析,然后对信息进行融合,通过自然语言生成的技术生成新的文本摘要。

按照文档数量划分,可以分为单文档摘要^[3]方法和多文档摘要^[4]方法。单文档摘要方法是指针对单个文档,对其内容进行抽取总结生成摘要;多文档摘要方法是指从包含多份文档的文档集合中生成一份能够概括这些文档中心内容的摘要。

按照文本摘要的学习方法可分为有监督^[5]方法和无监督^[6]方法。有监督方法需要从文件中选取主要内容作为训练数据,大量的注释和标签数据是学习所需要的。这些文本摘要的系统在句子层面被理解为一个二分类问题,其中,属于摘要的句子称为正样本,不属于摘要的句子称为负样本。机器学习中的支持向量机(SVM, support vector machine)和神经网络也会用到这样分类的方法。无监督^[7-8]的文本摘要系统不需要任何训练数据,它们仅通过对文档进行检索即可生成摘要。

2 研究现状

本文按照时间的顺序,对研究文本摘要方法的相关文献进行收集整理,对文本摘要方法的研究进展进行简要的概括总结和分析。

2.1 基于统计学方法

抽取式文本摘要方法的关键问题是从原文档中抽取出重要性高的若干句子^[9]。研究初期是以统计学为支撑,统计学的方法是基于统计特征,如词频、句子中心性(即与其他句子的相似性)、

句子位置、句子与标题的相似性、句子的相对长度等统计特征来生成摘要的。首先使用“词频”这一简单的文本特征对文档的重要句子和词组进行抽取生成,根据经验可知,除去停用词以外,文中出现频率越高的单词,其重要性也就越高。根据单词的词频高低分别设置相应的词权重,词频越高,对应的权重也就越高;句子的权重是组成句子单词的权重之和。然后从文档中抽取权重高的单词和句子组成摘要,这就是简单的基于词频的文本摘要方法。

句子在段落中的位置对句子的重要性影响很大,根据经验可知,一个段落的首尾句成为概括段落中心思想主题句的概率比其他位置句子的概率要高。Baxendale 等^[10]从句子位置的特征出发,通过计算文档段落首尾句能够概括主题的主题句概率,选取得分最高的若干句子生成文本摘要。Edmundson 等^[11]在研究中也指出,线索词(如“几乎不”“不可能”等类似的特定词汇)对句子重要性也起到了一定的影响作用。他们将词频、句子位置、线索词、标题词等多个特征组合起来作为衡量句子重要性的综合指标,在质量上取得了一定的改善。

其他著名的方法有 TF-IDF(词频-逆向文件频率),它是由 Salton 最先提出的^[12],这种统计的方法通过词频和逆向文件频率共同评估一个词在一个文件集或语料库中的重要程度,字词的重要性随着它在文件中出现的次数成正比增加,但同时会随着它在语料库中出现的频率成反比下降。利用 TF-IDF 生成文本摘要的基本思路是通过背景语料统计出各个词的 TF-IDF 值,将其作为重要性得分的权值,接着根据各个词的 TF-IDF 值累加求和计算得到段落中各个句子的 TF-IDF 值,将其作为该句的重要性得分,然后将句子按照重要性得分排序,依据文本摘要的大小依次选取重要性得分从高到低的若干个句子组成摘要。这种算法在词频算法的基础上引入一个逆文档因子,它能够有效地区分出高频无效词,在一定基础上改进了一般词频算法。其后,基于 TF-IDF 的方法被广泛应用于单词的重要性识别^[13-16]。

El-Beltagy 等^[17]提出了提取关键词性能最好的系统 SemEval2010,它利用词频等统计特征过滤

不可能是关键词的短语, 然后结合 TF-IDF 模型减小对特殊词的影响, 并对候选短语进行排序选出文摘句。Danesh 等^[18]基于统计启发式算法(如 TF-IDF 分数和文档中短语的第一个位置)的组合计算每个短语的初始权重。然后将短语及其初始权重联合, 基于图表的算法产生候选关键词的最终排序。Florescu 等^[19]在 Danesh 等的基础上进行了改进, 它不仅仅考虑文中每个词在句子中首次出现的位置, 还将每个单词在句子中出现的所有位置信息进行聚合来计算相应的词权重, 进一步改进统计学特征中词句位置因素对其重要性的影响, 在一定程度上改良了文本摘要的质量水平。Padmalahari 等^[20]综合统计特征和语言特征对句子赋予权重, 使用连续阈值从给定的输入文本文档中找出重要句子构成文本摘要, 摘要质量得到了有效提升。

统计学的方法主要适用于格式相对比较规范的文档摘要。此类方法比较经典, 首次在自动文本摘要方面取得了重大突破, 但由于此类方法只是基于句子和单词本身的表层特征进行统计, 未能充分利用词义关系、词间关系等特征, 所以还有很大的局限性, 因此针对这些问题, 一些改进的方法随后被接连提出。

2.2 基于外部语义资源的方法

1995 年, Miller^[21]提出了利用生成词汇链的方法来生成摘要, 这种方法主要通过对文章的内容进行自然语言分析处理来生成摘要。首先选定候选词的集合, 然后根据词汇链之间的相关程度, 为每一个候选词选择相应的词汇链, 如果候选词与该词汇链之间的相关程度比较高, 则将候选词添加到该词汇链中。最后依据词汇链的长度与一致性对其进行综合评分, 通过一些启发式方法选取若干词汇链生成最后所需要的摘要。Barzilay 等^[22]基于词汇链的摘要方法做出了创新改进, 不再用单个词作为分析单元, 而是通过 WordNet^[21,23]、词性标注工具等对词义进行分析, 将原文档中与某个主题相关的词汇集合起来, 构成词汇链, 词汇链重要性表达式如下。

$$Score(Chain) = Length \cdot HI \quad (1)$$

其中, $Length$ 表示词汇链中所有词汇的词频之和, HI 是均一度指数, 计算如式(2)所示。

$$HI = 1 - \frac{D}{Length} \quad (2)$$

其中, D 是词汇链中去重后词汇个数, 如果原文档中每个词均只出现一次, 则 $D = Length$, 均一度指数 HI 为 0。计算出各个词汇链的得分后, 根据

$$Score(Chain) > Average(Scores) + 2 \cdot StandardDeviation(Scores) \quad (3)$$

选择出强词汇链。确定强词汇链后, 通过制定的 3 条启发式的规则, 为每一个强词汇链抽取一个句子形成最后的摘要。在此基础上, 后来还提出了结合修辞结构、修辞学理论等一系列方法对词汇链方法进行改进^[24-26]。由于传统的词汇链只考虑名词而忽略其他语法部分的信息导致准确率较低, Hou 等^[27]在此基础上做了进一步改进优化, 分别引入谓词和形容词(副词)的词汇链, 三者一起构成全息词汇链(holographic lexical chain)用于中文文本摘要, 准确率取得了显著性的提高。Lynn 等^[28]通过对常规的词汇链进行增强改进, 引入文本中关键词的 3 个不同文本特征来产生更好的文本相关信息, 取得了良好的提升效果。

基于 WordNet 语义词典对单词的语义进行分析的方法, 通过 Lesk 算法从特定的情景中找到单词的实际意思^[29], 接着使用简化 Lesk 算法分别分析文本中所有句子的权重, 并根据其权重按降序排列。然后根据给定的摘要比例, 从有序列表中选择特定数量的句子生成摘要。虽然此类方法基于词的全局性和语义关系等多种属性来判断句子的重要性, 改进了简单统计方法未充分利用词间关系、词义关系等特征的不足, 但由于受限于当时的技术, 该方法对原文档的分析还停留在比较浅的层面。

2.3 基于图排序的方法

互联网网页上的文档具有较松散且涉及主题较多的结构特点, 在此基础上, 研究出了一些专门针对生成网页文档摘要的自动摘要技术, 就是基于图排序的文本摘要的方法。基于图排序的文本摘要生成的主要方法是通过把文章分成若干个段落或句子的集合, 每个集合对应一个图的顶点, 集合之间的关系对应边, 最后通过图排序的算法(如 PageRank^[30]、HITS^[31]等)计算各个顶点最后的得分, 然后依据得分高低生成文本摘要。

最先, Mihalcea^[32]使用 PageRank 算法及其改进的算法 TextRank 提取关键词和关键句子来生成文档摘要。在这个方法中, 他们将单词或句子作为图的顶点, 将词或句子之间的相似度作为图顶点之间的连边。句子之间的相似度单纯由句子之间内容重叠部分的多少决定, 相似度通过计算 2 个句子之间共同单词的数量得到。对于提取关键句子的部分, 由于考虑到长句所包含的信息量大, 覆盖的内容较多, 导致长句得分普遍比短句高, 因此对长句进行相应的惩罚, 将计算的得分值除去句子的长度得到相似度值, 即边的权值。当 2 个顶点之间相似度大于零时, 则存在一条边将 2 个顶点连接起来, 边的权值也就是顶点之间的相似度数; 当 2 个顶点之间相似度为零时, 则顶点之间不存在连边的关系。几种基本的图模型包括有向无权图、有向加权图、无向加权图、无向无权图。对于有向图, 边的方向是依据文档句子的顺序, 对于所有的图, 边的权值都是句子或词之间的相似度。基于这些图模型, 依据 PageRank、HITS 以及 TextRank 等算法对句子的得分进行计算, 按照句子得分的高低选择合适数量的句子生成摘要。Wan 等^[33]将 TextRank 扩展到 SingleRank, 通过在可变大小 $w \geq 2$ 的窗口中共同出现的单词之间引入加权边。ExpandRank 中包含类似的相邻文档以计算更准确字共现信息。Gollapalli 等^[34]扩展了 ExpandRank, 整合来自引用网络的信息来研究关键词的提取, 取得了超越当前最先进基线的良好实验效果。

Khan 等^[35]提出了一种多文档生成式摘要的改进语义图排序算法, 它是从源文档中以图节点表示的方式构建谓词变元结构 (PAS, predicate argument structures) 的句子语义结构, 图边代表相似性权重, 根据 PAS 语义相似度计算相似性权重, 用改进的图排序算法对重要图节点排序, 从中选取排名高的 PAS 生成摘要句子。Al-Khassawneh 等^[36]提出了一种基于混合图的单文档抽取摘要技术, 将余弦相似度 (sim1)、Jaccard 相似度 (sim2)、基于词对齐的相似度 (sim3) 和基于窗口的相似性度量 (sim4) 4 种不同的相似性度量结合创新了一种混合相似函数来用于计算图的权重, 一定程度上改良了单纯基于 TextRank 字

内容重叠计算权重的不足。其他图排序算法^[37-39]都是基于各种方法计算图节点的权重, 通过筛选出重要性程度高的节点来生成摘要, 都在一定程度上提高了摘要的质量。

耿焕同等^[40]利用句子之间的共同词汇, 在节点关系图的基础上提出了一种基于词共现图的文本摘要生成的方法, 并将条件概率的概念引入词共现矩阵中, 用以构造词共现矩阵。该方法通过词共现图形成的主题信息以及不同主题之间的连接特征信息自动提取生成摘要。

基于 PageRank、HITS 以及 TextRank^[41]改进的图排序算法都是无监督的, 它具有简单、语言弱相关及同时适合单文本和多文本等众多优点, 无需手动标注数据集, 相比有监督或半监督的深度学习的文本摘要方法, 节省了一定的时间和人力。虽然它具有语言独立、不需要大量语料训练、节约时间等优点, 但是由于无监督学习方法还没有充分挖掘和利用数据之间的关系, 所以虽然文本摘要的效果和质量有所提升, 但是对数据利用不够充分, 存在局限性, 之后的基于有监督、半监督学习的文本摘要方法对数据进行了合理利用, 相比无监督方法有显著的性能提升。

2.4 基于统计机器学习的方法

有监督的机器学习的方法在自然语言处理领域也得到了极其广泛的利用。监督、半监督的机器学习方法通过对数据集的语料进行人工标注, 人为地判定划分得到句子的文本特征以及句子重要性的关系模型, 在此模型的基础上, 对未被标注的语料进行训练, 预测未被标注语料中句子的重要性排序, 然后依据句子重要性排序, 依次选取若干句子生成最后所需要的文本摘要。

这种方法类似于一个二分类问题, 对于某一个句子, 它只有 2 种可能性: 要么是摘要句, 要么是非摘要句。首先通过人工标注的方法, 把重要句子手动标注出来, 接着在此基础上提取这些重要句子间的特征, 通过统计分析学习的方法, 找到句子重要性与这些特征之间的关系, 最后选择合适的分类器对其进行分类, 然后对分类器参数进行调整和对模型加以改进得到最佳的分类器结构, 但是这个过程中要防止过拟合的出现。对分类器输入一个句子, 在分类器的输出端将得到

该句子相应的重要性得分,依据重要性得分的高低抽取所需的句子生成摘要。

机器学习的方法很好地利用了计算机强大的计算性能^[42-43],借助统计学的知识理论,对海量的文本信息进行高效合理的建模,能够挖掘出隐藏在海量文本信息中的隐藏属性。这些机器学习的方法中使用较为广泛的有朴素贝叶斯算法^[44-45]、隐马尔可夫算法^[46]、决策树算法^[47]等。

Kupiec 等^[48]首次将统计机器学习的方法应用于文本摘要领域,他们选取了主题词特征、大写词特征、线索短语特征、句子长度特征、段落特征五大类特征,并认为这些特征之间是相互独立互不影响的。该方法在 21 种科技期刊中随机选择了 188 篇文章进行人工撰写摘要作为训练所需的语料,使用贝叶斯分类的方法训练得到一个分类器用以给句子的重要性进行评分。Kupiec 提出了一种通过朴素贝叶斯分类模型的新方法判定文档中的某一个句子是否应该被选取作为摘要。在他提出的方法中,假设 s 是其中某一个句子, S 是组成摘要的句子的集合, $F_1, \dots, F_k$ 是文本中的 k 个特征,假设这些特征两两之间是相互独立的,表达式如下。

$$P(s \in S | F_1, F_2, \dots, F_k) = \frac{\prod_{j=1}^k P(F_j | s \in S) P(s \in S)}{\prod_{j=1}^k P(F_j)} \quad (4)$$

通过上述表达式,可以为每一个句子计算它可能作为文本摘要的概率,然后依次从得分顺序由高到低的句子中顺次抽取前若干个句子生成摘要。

Conroy 等^[49]提出了隐马尔可夫模型的摘要算法,该算法通过使用文档中的一些特征(如句子位置、句内词数以及句子中词语与文档中词语的相似度等)来计算句子得分,然后依据句子得分生成文本摘要。Lin 等^[50]假设用于生成摘要的文档的各种特征之间是相互的,他们提出了决策树的算法来替代之前提出的贝叶斯分类模型来对句子进行评分,从中依次选取评分最高的若干句子生成摘要。

机器学习方法中特征选择的范围是相当广泛

的,由于训练分类器特征的可供选择性非常广泛,在本文看来,任意能够表示句子重要性的特征都可以用来训练分类器。这里所提及的特征有前文中提到的词频、线索词、句子位置、TF-IDF 值等简单的特征。此外,研究者还尝试了一些开放性的特征,如标题词长,平均句子长度,平均词汇连接度,是否包含数词、时间等信息,是否包含代词、形容词等词汇,是否包含命名实体等特征信息。基于这些开放性特征的机器学习的方法在提高生成摘要的可读性和准确度上起到了一定的效果。

2.5 基于深度学习的方法

深度学习是利用多层感知器结构对非线性信息进行处理的一种学习方法。Hinton 等^[51]首先提出了深度置信网络和相应的高效学习算法,该算法具有无监督学习的能力,降低了对人工的依赖,可以较为高效地进行训练,之后深度学习发展都是在此基础上进行改进的。

目前深度学习方法在对自然语言处理方面的基本方向是通过文档上下文进行学习训练,对于中文文档,还需要先进行中文分词处理,然后将文档中的词语、句子分别用连续实值向量进行表示,形成的向量称为嵌入向量,这样做是为了方便处理文本语义特征,将词语、句子用向量表示,在处理文本语义特征时,对词向量、句向量直接进行向量上的计算即可表征它们之间的文本语义关系。

要将自然语言处理的问题转化成为一个机器学习的问题,首先需要让机器能够理解自然语言,所以第一步就是将自然语言转化为机器可以理解的语言,于是想到将它进行符号数学化,为了能表示多维特征,增强其泛化能力,想到用向量对其进行表示,因此也就引出了对词向量^[52-53]、句向量的研究^[54]。但是词向量和句向量的生成仍然具有一定的难度,将文本中不同单词、句子用词向量或句向量进行唯一的表示,这样可以方便其在自然语言处理过程中进行各种操作变换和计算。虽然现在看来将文本中的词、句转化成唯一对应的词、句向量还具有相当大的难度,但由于其在自然语言理解中是关键的一步,因此这将是今后研究中亟待解决的问题和重点研究的方向之一。

Sequence-to-Sequence 模型^[55]广泛应用于机器翻译、语音识别、视频图片处理、文本摘要等多个领域。现在最新的一些基于深度学习研究文本摘要生成方法的也都是基于这个模型进行的。基于 Sequence-to-Sequence 模型的文本摘要需要解决的问题是从原文本到摘要文本的映射问题。摘要相对于原文具有大量的信息损失，而且摘要长度并不会依赖于原文本的长度，所以，如何用简短精炼的文字概括描述一段长文本是文本摘要需要解决的问题。最先是基于递归神经网络（RNN）的 Sequence-to-Sequence 模型用于文本摘要起到了一定的效果^[56]，之后考虑到 RNN 具有长程依赖性，为了减小长程依赖性，提出了基于长短时记忆网络（LSTM）的 Sequence-to-Sequence 模型用于文本摘要的生成。考虑到句子中的某些特定词或特定词性的词更具有影响句子中心意思的作用，引入了广泛应用于机器翻译中的注意力机制（attention mechanism）^[57]对句子的不同部分赋予不同的偏重，即权重。Rush 等^[57]在这个基础上提出基于注意力模型的生成式文本摘要，在 DUC-2004 数据集上显示出了优良的性能提升。Rush 等首次将深度学习应用于文本摘要的生成，他们基于 Sequence-to-Sequence 模型，将卷积模型应用于对原文档进行编码，然后利用上下文之间的注意力前馈神经网络生成摘要，在当时取得了很好的效果。Chopra 等^[58]同样利用卷积模型对原文档进行编码，但使用递归神经网络作为解码器解码输出，取得良好的改进效果，大大提高了文本摘要的质量。Ramesh 等^[59]将综合的 RNN^[58]和 attention 机制的 Sequence-to-Sequence 模型用于生成文本摘要^[60]，在文本摘要的准确性和可读性上取得了很大的提高。See 等^[61]利用注意力模型相对分心机制来缓解重复出现的冗余词句的权重，降低注意力权值大小达到去除冗余信息的效果，从而大大提高了文本摘要的新颖性。2016 年，谷歌也开源了基于 TensorFlow^[62]的一个自动摘要模块 Textsum，该模块也是利用 Sequence-to-Sequence 模型，基于深度学习的方法自动生成新闻标题，在质量和准确度上都取得了良好的效果。Sutskever 等^[63]考虑到 RNN 具有长程依赖性，为了减轻长程依赖的影响，将 RNN 用 LSTM 网络进行了替换，在此基础上取得了良好的改进。

最近 Facebook 尝试把 CNN 引入自然语言处理中，由于 CNN 不能处理变长的序列，在这之前 CNN 都只应用于图片视频领域的处理，恰巧 Sequence-to-Sequence 模型可以很好地处理变长序列问题，所以综合 CNN 与 Sequence-to-Sequence 模型，Gehring 等^[64]提出了基于 CNN 的 Sequence-to-Sequence 模型，成功应用于机器翻译领域，取得了比之前模型应用更优的结果，准确度超过了之前其他的所有模型。这同样也是一个启示，是否能够将基于 CNN 的 Sequence-to-Sequence 模型成功应用到文本摘要的领域取得最优的结果？这也是接下来要研究的一个问题。Liu^[65]将广泛应用于图像领域的生成对抗网络（GAN, generative adversarial networks）^[66]借用于文本摘要领域取得了显著成效，提出了一种生成式文本摘要的生成对抗过程，在这个过程中，同时训练一个生成模型 G 和一个判别模型 D。生成器通过文本的输入来预测生成摘要，判别器则试图将机器生成的摘要与真实摘要进行区分。在这个博弈过程中，双方不断提高性能，最后利用训练得到的生成器生成与真实摘要基本吻合的机器摘要。

最近，Tan 等^[67]提出了一个基于图的注意力神经模型的生成式文本摘要，该文在回顾了基于神经网络的生成式文本摘要难点的基础上，提出了一种新颖的基于图注意力机制的模型框架。这里仍然使用被广泛应用于机器翻译和对话系统中的编码器解码器的框架，需要特别说明的是，这里使用的是一个分层的编码器解码器结构，并引入了基于图的注意力模型。编码器结构由字编码器和句编码器组成，字编码器将句子中的单词编码成句子表示，句编码器将文档中的句子编码成文档表示。这里用 LSTM 结构的变体作为字编码器和句编码器的网络结构。不同于一般的解码器结构，这里使用的是带注意力机制的分层解码器结构。在原始解码器模型中，生成的代表文档整体的语义向量 c 对于所有的输出，字是相同的，这就需要语义向量 c 作为整个输入序列的充分表示。这里引入了注意力机制，对解码器不同发生状态下不同部分的输入赋予了不同的注意力，减轻了需要记忆整个输入序列的负担。分层解码器中还利用了波束搜索的策略对句子中的重复现象

进行缓解。最后基于图的抽取式摘要模型的 TextRank (Mihalcea 和 Tarau 于 2004 年提出)^[68] 和 LexRank (Erkan 和 Radev 于 2004 年提出)^[69] 对文档中句子的重要性进行识别。基于这种改进的图注意力的神经模型的生成式文本摘要方法实现了相较先前神经摘要模型相当大的提升, 效果甚至可以和最先进的抽取式摘要方法进行比较。

总体来说, 基于深度学习的自动摘要方法是目前效果最好的方法, 但其研究深度和研究数量还远远不够, 有待进一步深化。

3 方法比较

针对近些年来国内外文本摘要的各文本摘要方法, 对其优缺点进行了对比分析, 具体如表 1 所示。

4 结束语

文本摘要研究初期, 以原文档中的词、句为基础, 基于词频、句子位置等简单特征从文档中生成摘要。随后借助外部语义资源的方法 (如词汇链、WordNet 语义字典) 帮助选取关键词、关键句, 从而生成相应摘要。接着出现图排序的方法, 将文本单元中的各种关系映射到图论中的顶点与边的关系上, 利用句子中词共现频率、句子相似度、语义关系等特征生成边权重, 从而辅助判断找出句子重要性最高的若干个句子, 较好地解决了结构较为松散且涉及主题较多的文本摘要

生成的问题。之后, 随着计算机性能的显著提升, 基于统计机器学习的文本摘要方法兴起, 朴素贝叶斯算法、隐马尔可夫算法、决策树算法等大量统计机器学习的算法接连应用于文本摘要领域, 并在生成文本摘要方法的效果上取得一定的突破。当前最流行、效果最显著的是基于深度学习的生成文本摘要, 基于深度学习的文本摘要方法很好地利用了计算机强大的性能对文档的局部以及上下文的多维特征同时学习, 对特征进行编码向量化, 使文档的上下文特征、句法特征、语义特征等多维特征转化为能够进行计算的向量特征, 方便利用深层网络对其进行训练学习, 在文本摘要质量上实现了许多最优的实验结果。

展望未来, 可能的发展方向有: 1) 对于文本领域, 词、句的向量表示是极为关键的一个研究点, 相信在今后的研究中, 词、句向量的唯一表示也将是文本领域的重中之重; 2) 模仿人撰写摘要的模式, 融合抽取式和生成式摘要方法生成联合摘要; 3) 基于语义层面研究进一步地深度挖掘句子级、篇章级的语义并加以利用来进一步提高文本摘要的语义一致性和可读性; 4) 研究更好的摘要评估指标; 5) 回望过去文本摘要方法, 预测今后的自动文本摘要方法大趋势还是以深度学习为主进行, 但是传统方法仍有其特点和优势, 如何将深度学习和传统方法结合起来生成文本摘要也将是之后的研究重点所在。

表 1 各文本摘要方法的优缺点

方法	优点	缺点
基于统计学方法	依据文本形式上的规律, 简单直观, 避免考虑复杂的句法、语法结构, 易于实现且应用广泛, 无需训练数据, 执行速度快	只是单纯利用了单词表层特征, 没有充分挖掘词义关系和语义特征, 存在较大局限性
基于外部语义资源方法	在统计学方法的基础上利用词间关系、词义关系进行了改进, 使文本摘要的语义性能得到了一定的提高	受收录词汇的限制比较大; 对于文章题目依赖程度较高; 分词对关键词的影响较大; 相似度阈值的选取对构建词汇链有影响。语法语义结构不连贯
基于图排序方法	适用于结构较为松散且涉及主题较多的结构; 计算句子权重的同时可以充分考虑词汇之间、词组之间或句子之间的全局关系; 无监督, 语言独立, 不需要对大量语料进行处理	通常只考虑了句子节点间的相似性关系, 而忽略了文档篇章结构以及句子上下文的信息; 相似度计算的好坏决定了关键词和句子重要性排序的正确与否; 对数据的利用不够充分; 没有考虑信息冗余
基于统计机器学习方法	特征选择和训练分类器的选择上有较大的可供选择范围, 还可以综合一些开放性特征提高分类的精度	需要人工标注的数据集; 效果严格依赖于训练数据质量的好坏; 监督或半监督, 执行速度较无监督的方法慢
基于深度学习方法	降低了对人工的依赖, 可以高效地进行训练; 可以与多种神经网络结构和 Sequence-to-Sequence 模型结合, 生成文本摘要的可读性和准确度高	可解释性差; 需要大量人工标注的数据集; 由于有复杂的神经网络结构的引入, 执行速度慢, 需要花费相对较长的时间; 对计算机性能有一定的要求

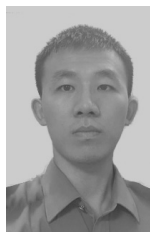
参考文献:

- [1] CHENG J, LAPATA M. Neural summarization by extracting sentences and words[J]. arXiv preprint arXiv:1603.07252, 2016.
- [2] NEMA P, KHAPRA M, LAHA A, et al. Diversity driven attention model for query-based abstractive summarization[J]. arXiv preprint arXiv:1704.08300, 2017.
- [3] LI P, LAM W, BING L, et al. Deep recurrent generative decoder for abstractive text summarization[J]. arXiv preprint arXiv:1708.00625, 2017.
- [4] BING L, LI P, LIAO Y, et al. Abstractive multi-document summarization via phrase selection and merging[J]. arXiv preprint arXiv:1506.01597, 2015.
- [5] LI C, QIAN X, LIU Y. Using supervised bigram-based ilp for extractive summarization[C]//The 51st Annual Meeting of the Association for Computational Linguistics. 2013: 1004-1013.
- [6] VEENA G, GUPTA D, JAGANADH J, et al. A graph based conceptual mining model for abstractive text summarization[J]. Indian Journal of Science and Technology, 2016, 9(S1).
- [7] DANESH S, SUMNER T, MARTIN J H. Sgrank: combining statistical and graphical methods to improve the state of the art in unsupervised keyphrase extraction[C]//The Fourth Joint Conference on Lexical and Computational Semantics. 2015: 117-126.
- [8] FLORESCU C, CARAGEA C. PositionRank: an unsupervised approach to keyphrase extraction from scholarly documents[C]//The 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1105-1115.
- [9] LUHN H P. The automatic creation of literature abstracts[J]. IBM Journal of Research and Development, 1958, 2(2): 159-165.
- [10] BAXENDALE P B. Machine-made index for technical literature—an experiment[J]. IBM Journal of Research and Development, 1958, 2(4): 354-361.
- [11] EDMUNDSON H P. New methods in automatic extracting[J]. Journal of the ACM, 1969, 16(2): 264-285.
- [12] SALTON G, YU C T. On the construction of effective vocabularies for information retrieval[C]//ACM SIGIR Forum. 1973: 48-60.
- [13] 施聪莺, 徐朝军, 杨晓江. TFIDF 算法研究综述[J]. 计算机应用, 2009, 29(B06): 167-170.
SHI C Y, XU C J, YANG X J. Study of TFIDF algorithm[J]. Journal of Computer Applications, 2009, 29(B06): 167-170.
- [14] 徐文海, 温有奎. 一种基于 TFIDF 方法的中文关键词抽取算法[J]. 情报理论与实践, 2008, 31(2): 298-302.
XU W H, WEN Y K. A Chinese keyword extraction algorithm based on TFIDF method[J]. Information Studies: Theory & Application, 2008, 31(2): 298-302.
- [15] SUQIN Z B S H M. An improved text feature weighting algorithm based on TFIDF [J]. Computer Applications and Software, 2011, 2: 7.
- [16] 李静月, 李培峰, 朱巧明. 一种改进的 TFIDF 网页关键词提取方法[J]. 计算机应用与软件, 2011, 28(5): 25-27.
- [17] LI J Y, LI P F, ZHU Q M. An improved TFIDF-based approach to extract key words from Web pages[J]. Computer Applications and Software, 2011, 28(5): 25-27.
- [18] EL-BELTAGY S R, RAFAA A. Kp-miner: participation in semeval-2[C]//The 5th International Workshop on Semantic Evaluation, 2010: 190-193.
- [19] DANESH S, SUMNER T, MARTIN J H. Sgrank: combining statistical and graphical methods to improve the state of the art in unsupervised keyphrase extraction[C]//The Fourth Joint Conference on Lexical and Computational Semantics. 2015: 117-126.
- [20] FLORESCU C, CARAGEA C. PositionRank: an unsupervised approach to keyphrase extraction from scholarly documents[C]//The 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1105-1115.
- [21] PADMALAHARI E, KUMAR D V N S, PRASAD S. Automatic text summarization with statistical and linguistic features using successive thresholds[C]// 2014 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT). 2014: 1519-1524.
- [22] MILLER G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.
- [23] BARZILAY R, ELHADAD M. Using lexical chains for text summarization[J]. Advances in Automatic Text Summarization, 1999: 111-121.
- [24] JAIN A, GAUR A. Summarizing long historical documents using significance and utility calculation using WordNet[J]. Imperial Journal of Interdisciplinary Research, 2017, 3(3).
- [25] SILBER H G, MCCOY K F. Efficient text summarization using lexical chains[C]//The 5th International Conference on Intelligent user interfaces. ACM, 2000: 252-255.
- [26] KOLLA M. Automatic text summarization using lexical chains: algorithms and experiments[D]. University of Lethbridge, 2004.
- [27] POURVALI M, ABADEH M S. Automated text summarization base on lexicales chain and graph using of wordnet and wikipedia knowledge base[J]. arXiv preprint arXiv:1203.3586, 2012.
- [28] HOU S, HUANG Y, FEI C, et al. Holographic lexical chain and its application in chinese text summarization[C]//Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data. 2017: 266-281.
- [29] LYNN H M, CHOI C, KIM P. An improved method of automatic text summarization for Web contents using lexical chain with semantic-related terms[J]. Soft Computing, 2018, 22(12): 4013-4023.
- [30] PAL A R, SAHA D. An approach to automatic text summarization using WordNet[C]//IEEE International Conference on Advance Computing Conference (IACC). 2014: 1169-1173.
- [31] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the Web[R]. Stanford InfoLab, 1999.

- [31] KLEINBERG J M, KUMAR R, RAGHAVAN P, et al. The Web as a graph: measurements, models, and methods[C]//International Computing and Combinatorics Conference, 1999: 1-17.
- [32] MIHALCEA R. Graph-based ranking algorithms for sentence extraction, applied to text summarization[C]//Proceedings of the ACL 2004 on Interactive Poster And Demonstration Sessions. Association for Computational Linguistics, 2004: 20.
- [33] WAN X, XIAO J. Single document keyphrase extraction using neighborhood knowledge[C]//AAAI. 2008, 8: 855-860.
- [34] GOLLAPALLI S D, CARAGEA C. Extracting keyphrases from research papers using citation networks[C]//AAAI. 2014: 1629-1635.
- [35] KHAN A, SALIM N, FARMAN H, et al. Abstractive text summarization based on improved semantic graph approach[J]. International Journal of Parallel Programming, 2018: 1-25.
- [36] AL-KHASSAWNEH Y A, SALIM N, JARRAH M. Improving triangle-graph based text summarization using hybrid similarity function[J]. Indian Journal of Science and Technology, 2017, 10(8).
- [37] WEI F, LI W, LU Q, et al. A document-sensitive graph model for multi-document summarization[J]. Knowledge and Information Systems, 2010, 22(2): 245-259.
- [38] GE S S, ZHANG Z, HE H. Weighted graph model based sentence clustering and ranking for document summarization[C]// 2011 4th International Conference on Interaction Sciences (ICIS). 2011: 90-95.
- [39] NGUYEN-HOANG T A, NGUYEN K, TRAN Q V. TSGVi: a graph-based summarization system for Vietnamese documents[J]. Journal of Ambient Intelligence and Humanized Computing, 2012, 3(4): 305-313.
- [40] 耿焕同, 蔡庆生, 赵鹏, 等. 一种基于词共现图的文档自动摘要研究[J]. 情报学报, 2005, 24(6): 652.
- GENG H T, CAI Q S, ZHAO P, et al. Research on document automatic summarization based on word co-occurrence[J]. Journal of the China Society for Scientific and Technical Information, 2005, 24(6): 652.
- [41] SEHGAL S, KUMAR B, RAMPAL L, et al. A modification to graph based approach for extraction based automatic text summarization[M]//Progress in Advanced Computing and Intelligent Engineering. Singapore Springer Press, 2018: 373-378.
- [42] YOUSEFI-AZAR M, HAMEY L. Text summarization using unsupervised deep learning[J]. Expert Systems with Applications, 2017, 68: 93-105.
- [43] ARRAS L, HORN F, MONTAVON G, et al. What is relevant in a text document? an interpretable machine learning approach[J]. PloS one, 2017, 12(8): e0181142.
- [44] THU H N T. An optimization text summarization method based on naive bayes and topic word for single syllable language[J]. Applied Mathematical Sciences, 2014, 8(3): 99-115.
- [45] SILVA G, FERREIRA R, LINS R D, et al. Automatic text document summarization based on machine learning[C]//2015 ACM Symposium on Document Engineering. ACM, 2015: 191-194.
- [46] NISHIKAWA H, ARITA K, TANAKA K, et al. Learning to generate coherent summary with discriminative hidden semi-markov model[C]//The 25th International Conference on Computational Linguistics: Technical Papers. 2014: 1648-1659.
- [47] ALLAHYARI M, POURIYEH S, ASSEFI M, et al. A brief survey of text mining: classification, clustering and extraction techniques[J]. arXiv preprint arXiv:1707.02919, 2017.
- [48] KUPIEC J, PEDERSEN J, CHEN F. A trainable document summarizer[C]//The 18th annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1995: 68-73.
- [49] CONROY J M, O'LEARY D P. Text summarization via hidden markov models[C]//The 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2001: 406-407.
- [50] LIN C Y. Training a selection function for extraction[C]// The Eighth International Conference on Information and Knowledge Management. ACM, 1999: 55-62.
- [51] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.
- [52] MRKŠIĆ N, VULIĆ I, ŠEAGHDHA D Ó, et al. Semantic specialisation of distributional word vector spaces using monolingual and cross-lingual constraints[J]. arXiv preprint arXiv:1706.00374, 2017.
- [53] XIONG Z, SHEN Q, WANG Y, et al. Paragraph vector representation based on word to vector and CNN learning[J]. CMC: Computers, Materials & Continua, 2018, 55(2): 213-227.
- [54] WANG X, ZHANG H, LIU Y. Sentence vector model based on implicit word vector expression[J]. IEEE Access, 2018, 6: 17455-17463.
- [55] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.
- [56] NALLAPATI R, XIANG B, ZHOU B. Sequence-to-sequence rnns for text summarization[J]. arXiv preprint arXiv: 1602.06023v1, 2016.
- [57] RUSH A M, CHOPRA S, WESTON J. A neural attention model for abstractive sentence summarization[J]. arXiv preprint arXiv: 1509.00685, 2015.
- [58] CHOPRA S, AULI M, RUSH A M. Abstractive sentence summarization with attentive recurrent neural networks[C]//The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 93-98.
- [59] NALLAPATI R, ZHOU B, GULCEHRE C, et al. Abstractive text summarization using sequence-to-sequence RNNs and beyond[J]. arXiv preprint arXiv:1602.06023v5, 2016.

- [60] CAO Z, LI W, LI S, et al. Attsum: Joint learning of focusing and summarization with neural attention[J]. arXiv preprint arXiv:1604.00125, 2016.
- [61] SEE A, LIU P J, Manning C D. Get to the point: summarization with pointer-generator networks[J]. arXiv preprint arXiv:1704.04368, 2017.
- [62] ABADI M, BARHAM P, CHEN J, et al. TensorFlow: a system for large-scale machine learning[C]//OSDI. 2016: 265-283.
- [63] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Advances in Neural Information Processing Systems. 2014: 3104-3112.
- [64] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning[J]. arXiv preprint arXiv:1705.03122, 2017.
- [65] LIU L, LU Y, YANG M, et al. Generative adversarial network for abstractive text summarization[J]. arXiv preprint arXiv:1711.09357, 2017.
- [66] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. 2014: 2672-2680.
- [67] TAN J, WAN X, XIAO J. Abstractive document summarization with a graph-based attentional neural model[C]//The 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1171-1181.
- [68] MIHALCEA R, TARAU P. Textrank: bringing order into text[C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004.
- [69] ERKAN G, RADEV D R. Lexrank: graph-based lexical centrality as salience in text summarization[J]. Journal of Artificial Intelligence Research, 2004, 22: 457-479.

[作者简介]



明拓思宇 (1994-), 男, 湖南长沙人, 国家数字交换系统工程技术研究中心硕士生, 主要研究方向为文本摘要。



陈鸿昶 (1964-), 男, 河南郑州人, 国家数字交换系统工程技术研究中心教授、博士生导师, 主要研究方向为电信网信息安全。