图书情报工作
第58卷 第18期 2014年9月

基于机器学习的自动文摘研究综述*

■ 曹洋 成颖 裴雷

[摘 要] 探讨基于机器学习的自动文摘研究中的特征选取、算法选择、模型训练、文摘提取和模型评测等主要过程;重点分析3种主要的机器学习算法:朴素贝叶斯、隐马尔科夫和条件随机场,阐释3种算法的基本思想,在对相关研究进行系统梳理的基础上,给出作者的思考;对3种机器学习算法在训练方法、协同训练与主动学习、类别平衡以及词汇分布等方面存在的共性问题进行深入讨论并提出未来的主要研究方向。

[关键词] 自动文摘 机器学习 NB HMM CRF
[分类号] G252.7
DOI:10.13266/j.issn.0252-3116.2014.18.018

1 引 言

信息时代的来临引发了文献的指数级增长,信息用户迅速由信息贫乏过渡到信息过载,传统手工处理文献的速度已经远远落后于文献增长的现实。为了应对这种状况,学术界尝试运用计算机技术实现对文献的自动处理,自动文摘即为其中之一。美国IBM公司的H. P. Luhn[1]于1958年发表的第一篇有关自动文摘的论文,拉开了该领域研究的序幕。近些年来,一系列国际测评会议的召开,例如文本理解会议(Document Understanding Conference,DUC)等公布了用于统一训练、测试的语料库,促进了该领域的研究。

自动文摘按照摘要的产生方式可以分为两类,即生成式和抽取式。生成式产生的文摘可以包括原文中没有出现的词和词组,一般基于实体信息、信息融合以及压缩技术等。由于生成式对自然语言处理技术要求非常高,目前还处于起步阶段,产生的文摘离实用化还有相当的距离[2]。抽取式通过选取原文中的句子形成文摘,通常依据预先定义的特征集合,对文档中的句子进行打分,得分高的输出为文摘句[3]。本文的研究主要聚焦于后者。

机器学习的出发点是设计和分析一些让计算机可以自动"学习"的算法,目前被广泛应用于自然语言处理、数据挖掘、搜索引擎等领域。机器学习可以分为监督学习、无监督学习和半监督学习。监督学习是指从给定的训练集中学习出模型,然后将其运用于新出现的数据以预测结果。监督学习的训练集包括输入与输出,即特征与目标,其中目标由人工标注。无监督学习与前者相比,主要区别在于训练集没有人工标注的结果。半监督学习介于前两者之间[4]。本文主要研究基于监督学习的自动文摘算法,同时会涉及后两者。

基于机器学习的自动文摘就是利用机器学习算法,依据给定的特征集,自动地从语料库中训练出模型,进而得到文摘。限于篇幅,本文主要研究3种经典的机器学习算法,即朴素贝叶斯(Naive Bayes,NB)、隐马尔科夫(Hidden Markov Model,HMM)和条件随机场模型(Conditional Random Fields,CRF)在自动文摘中的拓展与应用。

2 文献选取

2.1 文献选取原则

(1)选取基于机器学习的自动文摘研究文献,书评等非研究型文献排除在外。

(2)选取的研究对象为文本,不包括面向语音、视频以及图像等非文本型的相关研究。

(3)主要聚焦于基于机器学习的自动文摘算法研究,其他有关文摘评价等主题的文献不是本文研究重点。

2.2 文献选取过程

作者可以理解的语种为英文和中文,鉴于英文文

*本文系国家社会科学基金重大招标项目"面向学科领域的网络信息资源深度聚合与服务研究"(项目编号:12&ZD221)和国家自然科学基金项目"融合范式视角下的链接分析理论集成框架及其实证研究"(项目编号:71273125)研究成果之一。

[作者简介] 曹洋,南京大学信息管理学院硕士研究生;成颖,南京大学信息管理学院教授,博士,博士生导师,通讯作者,E-mail:chengy@nju.edu.cn;裴雷,南京大学信息管理学院副教授,博士。

收稿日期:2014-07-24 修回日期:2014-08-22 本文起止页码:122-130 本文责任编辑:刘远颖

献反映了该领域大部分高质量研究的现状，因此文献选择以英文为主，中文为辅，检索时间跨度为2014年3月10－15日。

对于英文文献，选择的数据库为：WOS、ACM、EI、IEEE、PQDD、Wiley、ScienceDirect和Google Scholar；检索词为：（"summari＊"or"abstract＊"）and（"machine learning"or"supervised method"）。对于具体的3种算法再分别构造检索式：①朴素贝叶斯：（"summari＊"or"abstract＊"）and（"Naive Bayes"）；②隐马尔科夫：（"summari＊"or"abstract＊"）and（"Hidden Markov Model"）；③条件随机场：（"summari＊"or"abstract＊"）and（"Conditional Random Fields"）。检索类型为：主题或者标题，有的数据库不支持主题检索。检索结果：去重后按照文献选取原则，共得到65篇文献。

对于中文文献，选择的数据库为中国知网和万方数据库；检索词为：（自动文摘 or 自动摘要 or 文摘自动化）and（机器学习 or 监督方法），同英文文献检索，分别对3种方法再构造检索式进行检索；检索类型为：主题；检索结果：根据文献选取原则，共得到8篇文献。

3 基于机器学习的自动文摘过程

基于机器学习的自动文摘过程可以概括为5个步骤[5-7]：特征选取、算法选择、模型训练、文摘提取和模型评测。

3.1 特征选取

特征选择对于自动文摘研究起着至关重要的作用。早先，H. P. Luhn[1]提出了一种基于高频词的特征选取方法，即对包含高频词的句子打分，得分高的输出为文摘句。P. B. Baxendale[8]提出将句子位置作为特征，然后对句子打分排序。H. P. Edmundson[9]结合了H. P. Luhn和P. B. Baxendale的做法，面向科技文献取得了很好的效果。除了词频与句子位置之外，句子长度、句子权重以及大写词等也都属于表层特征，后继研究中又提出了实体以及语言层面的特征[10]。其中，语义特征包括：修辞结构、句法特征以及语义概念[11]等；实体特征包括：命名实体、邻近度、实体词的联系[12-13]、相似度以及逻辑关系等。

3.2 算法选择

可以用于自动文摘的机器学习算法有：朴素贝叶斯、最大熵模型、条件随机场、隐马尔科夫模型、贝叶斯网络、支持向量机、遗传算法、回归模型等。由于篇幅的原因，本文第四部分将重点探讨朴素贝叶斯、隐马尔科夫和条件随机场模型。

3.3 模型训练

选择相应的语料库，一般将语料库分为训练部分和测试部分；利用选择好的算法对训练语料库进行训练，得到一个训练模型；然后用测试语料库测试模型的性能。

3.4 文摘提取

有监督的提取式方法将文摘任务看成句子层面的二元分类问题，文摘句是积极的一类，非文摘句是消极的一类。文摘句的提取过程就是利用机器学习分类器对句子进行打分，得分高的句子即文摘句。句子分类的过程可以分为两种方式：一种是有区别的方式（例如支持向量机算法），该方法是有效的，不过需要假设句子是独立的，分类时没有考虑句子间的联系；另一种是有联系的方式（例如隐马尔科夫模型），该方法考虑了位置间的联系，其不足之处是当特征集很大、特征之间不相互独立或者特征之间有重叠时，训练过程就会变得非常困难，所以该方法不能结合所有有用的特征。无监督的文摘提取方式则主要运用多种不同特征和句子间的关系进行，主要有修辞结构[14]、词汇链[15]、隐藏主题[11]以及基于图[16]的方法等。

3.5 模型评测

基于机器学习的自动文摘主要有F值和ROUGE两种评价方法。前者涉及准确率P、召回率R和F值：

$$P=\frac{a}{a+c}\quad R=\frac{a}{a+b}\quad F=\frac{2P\times R}{P+R}$$

其中a表示在文摘中且被标注为文摘的句子数；b表示在文摘中但没有被标注为文摘的句子数；c表示不在文摘中但被标注为文摘的句子数。

ROUGE（Recall Oriented Understudy for Gisting Evaluation）评价法由Lin Chin-yen等[17]在参考了K. Papineni等[18]的机器翻译自动评价方法后提出。该方法首先由多个专家分别生成人工文摘，构成标准文摘集；然后将系统生成的自动文摘与人工生成的标准文摘相对比，通过统计二者之间重叠的基本单元（n元语法、词序列或词对）的数目，来评价文摘的质量。通过与多专家人工文摘的对比，提高评价系统的稳定性和健壮性。ROUGE主要包括以下4种评价指标：

（1）ROUGE－N，基于n－gram的共现统计。

（2）ROUGE－L，基于最长公共子串。

（3）ROUGE－S，基于顺序词对统计。

（4）ROUGE－W，在ROUGE－L的基础上，考虑串的连续匹配。

4 面向机器学习的自动文摘算法

4.1 朴素贝叶斯

4.1.1 相关研究 贝叶斯分类算法是一种基于概率的分类算法,即通过对象的先验概率,利用贝叶斯公式计算其后验概率,也就是该对象属于某一类的概率,通常选择具有最大后验概率的类作为该对象的所属类。朴素贝叶斯[19]是一种经典的贝叶斯算法,该算法假设对象的特征相互独立。

1995 年,J. Kupiec 等[20]首次运用 NB 分类器对文本进行自动文摘。该研究的基本思想是:对每个句子 s 计算其成为文摘 S 的概率,文摘包括 k 个特征 $F_j(j=1\cdots k)$,基于贝叶斯定理得到公式(1):

$$P(s\in S|F_1,F_2\cdots F_K)=\frac{P(F_1,F_2\cdots F_K|s\in S)P(s\in S)}{P(F_1,F_2\cdots F_K)} \tag{1}$$

假设特征间相互独立,从而得到公式(2):

$$P(s\in S|F_1,F_2\cdots F_k)=\frac{\Pi_{j=1}^{k}P(F_j|s\in S)P(s\in S)}{\Pi_{j=1}^{k}P(F_j)} \tag{2}$$

其中,$P(s\in S)$ 是常数,$P(F_j|s\in S)$ 和 $P(F_j)$ 可以从训练语料库中估计,这样每个句子都有自己的概率得分,从高到低排列,高概率的句子输出为文摘句。

针对特征选取,J. Kupiec 吸收了 C. D. Paice[21]、H. P. Luhn[1]的做法,以及自己提出的另外 5 个特征:句长、固定短语、段落、主题词和大写字母词语。对于训练语料库,J. Kupiec 等选取了 Engineering Information 数据库的技术文献,其文摘由专业文摘人员编制,共包含了 21 个出版物的 188 篇面向科学技术类的文档。实验结果表明,使用段落、固定短语和句长的特征组合,文摘效果最好;当主题词和大写字母词语特征加入之后,效果会降低。当文摘长度为文档平均长度的 25% 时,84% 的文摘句与专业文摘员编写的一致。

1998 年,C. Aone 等[22]研制了基于 NB 的 DimSum 自动文摘系统。对于特征选取,相比较 J. Kupiec,C. Aone 引入了 TF * IDF 概念[23]和句子位置,并且把 J. Kupiec 的段落特征缩小到第一至四段,训练语料库为报纸集和 TREC - 5 文档集,实验表明,该系统具有较好的效果。

2005 年,B. Hachey 和 C. Grover[24]面向法律文本研制了运用了 NB 和最大熵模型相结合的自动文摘系统,特征方面引入了命名实体词;2008 年,A. Sharan 等[25]引入了名词特征;通过和微软 Word Summarizer 系统的对比,两种系统都具有更好的效果。

2010 年,M. Yousfi - Monod 等[26]也研制了面向法律文本的 ProdSum 系统。在运用 NB 训练前,该研究将从语料实例中抽取的句子分为 5 类:不在文摘中、引言、内容、推理和结论。语料库 2/3 的文档用于训练,1/3 用于测试。特征分为 3 类:表面特征、强调特征和内容特征。表面特征包括句子的位置、句长等;强调特征是指因为对文本采用了 HTML 源码分析,保存了一些突出的特征,例如,加粗、下划线、斜体、缩进和布尔值等;内容特征包括词项的 TF * IDF 得分和法律领域的专有词,例如,"apparently"、"dismissed"、"daughter"和"kill"等,计算专有词时考虑了词频。

2010 年,M. S. Pera 等[27]运用多项贝叶斯(Multinomial Naive Bayes,MNB)分类器进行文摘选取,利用词共现计算句子属于一个类别的概率。2013 年,J. E. Lee 等[28]针对新闻领域,利用 C4.5 和 NB 算法完成用户对文摘需求的建模,进而生成文摘。2013 年,A. Aries 等[29]提出一种无监督的方法,即首先对文档中的句子进行聚类形成不同的主题后,再利用 NB 算法生成文摘的研究思路。实验结果表明,这几种方法都获得了较好的文摘效果。

4.1.2 一些思考 朴素贝叶斯的优点是它发源于古典数学理论,有着坚实的数学基础以及较为稳定的分类效果;NB 模型所需估计的参数不多,对缺失数据不太敏感,算法也较为简单,与其他模型相比具有较小的误差率。其缺点是分类器假设特征间相互独立,把每个句子单独对待,忽略了句子间的联系,在特征个数比较多或特征间相关性较大时,选择 NB 模型生成文摘的效率较低;其次,NB 分类器是假设预知先验概率,预测出后验概率。不过,在实际应用中先验概率很难知道,通常的做法是取近似值,不过会给计算后验概率带来一定的负面影响[30]。因此,对规模比较小或者特征间关联性较弱的语料库,利用 NB 进行自动文摘的效果较好。如何将 NB 运用于规模较大或者特征间关联较强的语料库,将成为后继的研究重点。

4.2 隐马尔科夫模型

4.2.1 相关研究 在马尔科夫模型中,状态对于观察者来说是直接可见的,状态的转移概率便是全部的参数。隐马尔科夫模型则用来描述一个含有隐含未知参数的马尔科夫过程,其状态并不直接可见,不过受状态影响的某些变量则可见。每一个状态在可能输出的符号上都有一概率分布,因此输出符号的序列能够透露出状态序列的一些信息。20 世纪 80 年代,HMM 被应

用于语言识别[31]。90年代,逐步被运用于文字识别和生物医学等领域。

2001年,J. M. Conroy 和 D. P. O'Leary[32]将HMM运用于自动文摘。其文摘过程类似于HMM中的解码问题,即已知模型参数,利用观测序列寻找最可能的隐含状态序列。观测序列即句子位置、句子中词项的数量以及词项在文档中的概率等3个特征,隐含状态序列就是包括文摘状态和非文摘状态的一串序列,即待预测的状态。模型参数包括初始状态概率矩阵、隐含状态转移概率矩阵和观测状态转移概率矩阵。文献[32]构建的模型有$2s+1$个状态,其中包括s个文摘状态和$s+1$个非文摘状态,开始为非文摘状态,然后是文摘状态,依次间隔排列。隐含状态转移概率矩阵M可以从训练语料库中估计,矩阵M中的一个元素m_{ij}表示从状态i变为状态j的概率。研究中定义$b_i(O)=Pr(O|state\ i)$,其中O表示观测序列,即文档的3种特征,构成了观测状态转移概率矩阵。通过该模型,计算$\gamma_t(i)$,句子t符合状态i的概率,如果i为偶数,就代表句子t为第$i/2$个文摘句的概率;如果i为奇数,就代表是非文摘句的概率。若此,要判断一个句子是否为文摘句,只要对所有偶数i的$\gamma_t(i)$进行求和,依次计算好文档中的每个句子,求和最大的句子输出为文摘句。

文献[32]选取了TREC会议数据集的1 304篇文档作为训练语料库,与J. Kupiec等的训练语料库中文摘由专业文摘人员编制不同,其语料库的人工文摘部分都由同一人完成,该方法保证了文摘风格的一致性,不过也缺少了多样性。对于模型评价,采用F_1作为评价指标:

$$F_1=100\frac{2r}{k_h+k_m}$$

其中,k_h表示人工文摘的长度,k_m表示机器文摘的长度,r表示人工文摘和机器文摘共同含有的句子数量。实验结果表明,采用隐马尔科夫的模型比采用NB的DimSum系统有更好的效果。

2002年,J. D. Schlesinger等[33]在文献[32]的基础上,提出了基于HMM和logistic回归模型的自动文摘系统。在特征选取方面,J. D. Schlesinger吸收了文献[32]的前两个特征,对第三个特征进行了优化,将其定义为"虚假查询词"的数量,即在一篇文档中出现的概率大于在整个文档集中出现概率的词项。对于文摘的选取,通过设置阈值使得句子按照HMM的得分由高至低输出到文摘集合中,直到文摘集合中的词项大于100。

多项有关自动文摘的研究提出了先对文档进行聚类的技术思路,实验表明该思路生成的文摘具有更高的质量。2003年,P. Fung等[34]提出将聚类和HMM相结合的无监督多文档自动文摘方法,该研究运用K-means方法先对文本进行聚类,其优点是不需要标注好的训练语料。2006年,D. M. Dunlavy等[35]研制了QCS信息检索系统,该系统集成了查询、聚类和摘要等功能。对于给定的查询,QCS系统检索相关文档,将文档形成不同的主题聚类,并且对每个类生成摘要。具体做法是将潜在语义索引(LSI)用于检索、K-means用于文档聚类、隐马尔科夫用于文摘生成。D. M. Dunlavy在文献[32]构建的文摘模型基础上,将文摘分为两步:整理句子、选择句子生成文摘。其中整理句子的算法基于D. M. Dunlavy等2003年的工作[36],其主要作用是将那些不重要的句子剔除,从而提高了文摘的质量。

2007年,D. Zajic等[37]针对大型的多文档自动文摘系统,提出将句子先行压缩处理的自动文摘思路。句子压缩有两种方法:其一是"解析-整理"法,例如,Trimmer系统和延伸版本的Topiary系统都采用了该方法;其二是基于HMM的方法,被HMM Hege系统所采用。文摘的过程分为三步:选择文档中开头的五句话;利用不同的句子压缩方法对句子进行压缩,形成文摘候选集;基于给出的一系列特征的线性组合,选择句子形成文摘。

除了上述研究之外,HMM还被用于系统日志总结[38]。2011年,W. M. Darling等[39]运用潜在狄利克雷分布(Latent Dirichlet Allocation,LDA)和隐马尔科夫的混合模型构建文档概率分布,将主题与语义相结合,提高了文摘的质量。

2014年,刘江鸣等[40]基于A. Gruber等[41]提出的隐主题马尔科夫模型HTMM(Hidden Topic Markov Model)研制了一种基于HTMM的多特征自动文摘系统。HTMM是指句子间的主题关系符合马尔科夫性质,并且主题转移服从二项分布。刘江鸣等提出的模型消除了LDA主题模型的主题独立假设,使得文摘生成过程中充分利用文档的结构信息,并结合基于内容的多特征方法提高文摘质量。在DUC2007标准数据集上的实验证明了HTMM和文档特征的优越性,所实现的自动文摘系统ROUGE值有明显提高。

4.2.2 一些思考 HMM的优点是用联合分布表示特征集合,需要较少的独立性假设条件,可以解决NB句子之间缺少联系的问题;其次,利用连续统计方法对文档进行自动文摘,有坚实的数学基础;在实现层面,在处理不同自动文摘任务时具有较强的灵活性,且操作简单[42]。其缺点是序列数据须严格相互独立,才能保证推导的正确性,而事实上大多数序列数据不能表示为一系

列独立事件;HMM 有连续学习的能力,但是不能运用丰富的语言特征,而且难以运用句子间的关系特征;需要大量训练数据进行学习,主要通过累计概率的最大值来决定相应的状态。因此,HMM 适合于规模比较小的语料库或者特征间关系比较简单的语料库。如何运用丰富的语言特征将是 HMM 下一步研究的重点。

4.3 条件随机场

4.3.1 相关研究 条件随机场由 J. Lafferty 等[43]于 2001 年提出,结合了最大熵模型和 HMM 的特点,属于判别模型,也是一种无向图模型。CRF 模型是在给定观察序列的前提下,计算整个标注序列的概率。在研究文献中,CRF 与 HMM 常被一并提及,前者对于是否存在输入和输出概率分布假设没有后者要求得那么高。对于 HMM 中存在的两个假设——输出独立性假设和马尔可夫性假设,CRF 另辟蹊径,使用了一种概率图模型。CRF 具有表达长距离依赖性和交叠性特征的能力,能够较好地解决标注(分类)偏置等问题,而且所有特征可以进行全局归一化,能够得到全局最优解。

2006 年,M. Saravanan 等[44]提出了基于概率图模型的自动文摘思路,即给出观察序列 X 和对应的标注序列 Y,目标是在基于 CRF 给出的概率框架下,计算在 X 的条件下 Y 的概率。对应到自动文摘中,就是将文摘任务看成序列标注问题。每篇文档都由一系列句子构成,输出目标是一串 1 和 0 的序列,其中属于文摘的句子标签为 1,否则为 0。一个句子的标注依赖于其周围的句子。给出观察序列 $X=(x_1,\cdots,x_M)$ 和对应的标注序列 $Y=(y_1,\cdots,y_M)$,观察序列就是句子的特征表示,标注序列就是是否为文摘句,y_i 取 0 或 1。CRF 的目标是找到序列 Y,使公式(3)最大化:

$$P(Y|X,W)=\frac{1}{Z_X}\exp(W*F(X,Y)) \tag{3}$$

其中,Z_X 是归一化常数,确保概率和为 1;$F(X,Y)=\sum_{i=1}^{M}f(i,X,Y)$ 是维数为 T 的垂直向量,其中的垂直向量 $f=(f_1,f_2,\cdots,f_T)$ 代表的是 T 个特征向量,可以写成 $f_t(i,X,Y)\in R$, $t\in(1,\cdots,T)$, $i\in(1,\cdots,M)$。W 是特征函数权值。通过与 HMM、SVM 等 8 种方法的比较,采用 F_1 和 ROUGE-2 测评,基于 CRF 的自动文摘都取得了较好的效果。

2007 年,Shen Dou 等[45]改进了 M. Saravanan 的工作,研究结合了有监督和无监督学习的特征选择策略,将其分为基本特征和复杂特征两大类。其中,基本特征包括位置、长度、主题词、指示词、大写词和邻句相似度;复杂特征包括 LSA 得分和 HITS 得分,LSA 得分是指通过奇异值分解词-句子矩阵,以获得隐藏主题,从而将每个句子投影在对应的主题上,根据这种投影特征可以找到重要的句子;HITS 得分是将文档看成图,运用基于图的排序算法,例如 HITS 或者 PageRank,从而使每个句子可以获得反映重要程度的得分。

2009 年,吴晓峰和宗成庆[46]将 LDA 提取的主题作为特征加入 CRF 模型中进行训练,并分析了在不同主题下 LDA 对摘要结果的影响,获得了较高的文摘质量。2009 年,邓箴等[47]提出了基于 CRF,以统计为主、抽取关键词为辅的句子抽取方法,并通过篇章分析使文摘能涵盖文档的最大信息量,然后运用一组规则使文摘具有更好的可读性,实验表明,系统在处理主题为经济和新闻报道的文本时有较好的效果。2012 年,张龙凯[48]提出了基于 CRF 的句子抽取算法,即将文本看作是句子的序列,句子是序列中的一个点。如果句子出现在摘要中,则标注为“在”,否则,标注为“不在”。他采用 CRF 模型利用带标注的文本集合训练出一个序列标注模型。由于文摘通常都远远短于原始文本,文档中大多数句子都是非文摘句,据此引入了修正因子以平滑这一现象。对于特征选取,他引入了关联特征,即上下文特征,包括与标题的相似度、与其他句子的相似度以及与前后句的相似度等。

2013 年,N. K. Batcha 等[49]提出一种将 CRF 和非负矩阵分解(Non-negative matrix factorization ,NMF)相结合的自动文摘方法,利用 NMF 对文档词项矩阵进行降维。Wong Tak-Lam 等[50]针对拍卖网站的热门产品选取问题,设置了一个可以自动提取商品特征和自动对热门产品进行摘要的框架,并且利用 CRF 模型有效地将原始问题转化为单标记图问题。

4.3.2 一些思考 基于监督的自动文摘把文摘任务看成是二元分类问题,没有考虑句子间的联系。直观上,两个句子如果相邻且具有相类似的内容就不应该同时进入文摘句,基于监督的自动文摘难以处理这种情况。HMM 具有连续学习的能力,但是不能运用丰富的语言特征。无监督的自动文摘运用启发式的规则去选择最可能的句子生成文摘,不过该方法很难产生实用的文摘。基于 CRF 的自动文摘,结合了监督和无监督方法的优点,同时尽可能回避了它们的不足。

CRF 模型的优点是没有 NB 和 HMM 严格的独立性假设条件,因而可以容纳任意的上下文信息;能够较好地解决标注(分类)偏置等问题;所有特征可以进行全局归一化,对特征的融合能力比较强,能够求得全局的最优解;模型是在给定标注观察序列的条件下,计算

整个标注序列的联合概率分布，而不是在给定当前状态条件下，定义下一个状态的状态分布。CRF模型的缺点是特征的选择和优化是影响自动文摘质量的关键因素，特征选择的优劣直接决定了文摘系统性能的高低；模型需要训练的参数更多，训练代价大、复杂度高。对于CRF的下一步研究可以考虑增加全局信息[47]。

4.4 其他算法

除了以上3种算法外，很多学者利用其他算法进行自动文摘，并且取得了很好的效果。Lin Chen-Yew[51]提出的基于决策树算法进行自动文摘取得了很好的效果。M. Osborne[52]采用最大熵模型，实验结果表明，通过增加先验概率，该方法优于NB方法。Yeh Jen-Yuan等[53]采用遗传算法进行自动文摘。G. Murray等[54]将回归模型应用于会议记录的摘要。K. Kaikhah[55]利用神经网络对新闻进行自动摘要研究。M. Fuentes等[56]将支持向量机算法应用到自动文摘研究中。上述技术路线的研究也都使文摘质量得到提高，限于篇幅，不再展开。

5 讨 论

通过上述相关研究发现，如何获得高质量已标注的训练集、如何提升分类器的学习过程、如何选取更有效的特征等成为获得高质量文摘的关键。

5.1 训练方法

各模型标准的训练方法存在有待改进之处。例如，基于句子层面的准确度和文摘评价准则不一致，文摘评价准则一般是基于文摘层面的ROUGE得分；训练准则非最优。由于是二元分类问题，句子标签在训练过程中太严格。A. Aker等[57]提出了基于MERT的有区别训练方法，在训练过程中利用一种称为A*搜索的算法产生多个候选文摘，通过反复迭代特征的权重以达到ROUGE的最优值。Lin Shih-Hsiang等[58]提出基于排序的训练方法，这种方法运用排序SVM将句子成对进行训练，而非每个句子单独分类。Xie Shasha等[59]提出回归模型，即在训练过程中将句子的标签置为+1或-1，每个句子都贴上数字标签代表权重，对于文摘句赋予+1，否则赋予-1。结果表明，回归模型相较于传统的二元分类方法，文摘质量有显著提高。

5.2 协同训练与主动学习

抽取式文摘模型训练的过程中，需要已经标注好的训练集，如何获得高质量的训练数据非常重要。是否可以用未标注的数据来帮助机器学习自动产生文摘？对此，很多机器学习采用半监督的方法。其中，协同训练和主动学习已经应用于自动文摘。协同训练是指运用协同训练方法，利用无标注数据，特征集代表不同的观点，它们独立地贴上类标签，基于每个观点选择不同的句子，分类器在此基础上进行训练。Wong Kam-Fai等[60]利用协同训练完成了基于SVM和NB两种分类器的自动文摘系统。Xie Shasha等[59]在演讲摘要中也使用了协同训练。两位研究者将声学和词汇作为选择的特征，同时选择句子和文档作为选择单元。Zhang Jian Justin[61]将主动学习运用到讲座摘要中，抽取的摘要和教学课件中的句子越相似，则得分越高。

5.3 类别平衡

基于机器学习的分类中，文摘句所属的类别很小，二元分类导致了类别不平衡。不平衡数据会影响分类器训练的效果。可以通过对训练语料库进行抽样来解决该问题：向上抽样和向下抽样。向上抽样通过增加少数类的样本，即通过复制现有少数样本、插入内容形成综合样本等方式完成；向下抽样通过减少多数类的样本，即在现有多数类中随机抽取一部分。Xie Shasha等[59]提出了基于抽样的自动文摘，不同于上述两种经典的方法，即对于向上抽样，Xie Shasha采用基于余弦相似度或者ROUGE得分选择那些和文摘句相似度比较高的非文摘句；对于向下抽样，则选择和文摘句完全不同的句子。该方法同时解决了人工标注的错误问题。

5.4 词汇分布

在利用机器学习算法建模的研究中，词频特征以及分布规律都非常重要。不过，词汇集规模非常大，易引发数据稀疏问题。对此，从模型中发掘潜在的词汇分布规律成为一个可能的解决途径，现有研究中多采取降维的方法实现。W. B. Frakes等[62]采用TF*IDF方法，即将词项频度与词的稀疏度相结合，从而将任意长度的数据规模减少到词项级别，该方法虽有效但没有反映词项之间的相关关系。吴晓峰和宗成庆[46]提出的LDA方法包含分为三层的贝叶斯生成概率模型，把文本语料看作离散的数据，数据的每一个元素被看作由底层的有限个混杂在一起的主题产生，而每一个主题又被看作从一个更底层的主题概率模型中产生，该方法成功地将词项级别的规模降到了主题级别。

6 未来研究方向

综合上述分析，笔者认为以下几方面将是本领域未来可能的研究方向：

（1）选取特征，将成为自动文摘研究中的关键。从表层特征到语义特征，有学者提出基于Wikipedia的

特征选择[63-65],即利用 Wikipedia 实体信息进行排序。此外,对于不同领域,特征的选取不同,例如,对于新闻领域,句子的位置特征更为重要;对于网页内容,其入链与出链特征的重要性就显现了出来[66-68]。

(2)如何更好地进行无监督自动文摘或者半监督自动文摘研究将是以后的研究方向。有监督的机器学习需要人工对训练语料进行标注,对于大型语料库,工作量非常大。研究中可以考虑利用未标注的语料进行自动文摘,比如有学者提出基于聚类的自动文摘思想[69-71]。

(3)在对训练语料进行标注时,不同的专家会产生不同的观点,故如何更好地进行语料标注,从而有效地降低训练语料的噪声也将是亟待解决的问题。

(4)由于中文语言的特殊性,如何深入分析和理解中文文本的句法和语义特征,将是未来中文自动文摘研究的一个重要方向。中文的特殊性在于需要自动分词,且汉语的语言结构复杂、缺少词形变化等,这就加大了句法分析的复杂性。

(5)经典算法本身很难有太大的改进,但是可以将多种算法结合起来加以考虑,以取长补短。例如,M. A. Fattah[72]将最大熵模型、NB 和支持向量机结合在一起,Lei Yu 等[73]将隐马尔科夫和 CRF 结合在一起等都是可能的突破点。

(6)有关自动文摘的评价方法有很多,找到一种更为客观且有效的评价方法也将是下一步研究的方向。

参考文献:

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

[2] Mani I, Maybury M T. Advances in automatic text summarization [M]. Cambridge: MIT Press, 1999.

[3] Mani I, Bloedorn E. Machine learning of generic and user-focused summarization [C]//Proceedings of the Fifteenth National Conference on Artificial Intelligence. Reston VA: AAAI Press, 1998: 821-826.

[4] Mitchell T M. Machine learning[M]. Burr Ridge: McGraw Hill, 1997:45.

[5] 郭燕慧,钟义信,马志勇,等. 自动文摘综述[J]. 情报学报, 2002(2):582-591.

[6] Jones K S. Automatic summarizing: Factors and directions[C]// Advances in Automatic Text Summarization. Cambridge: MIT Press,1999:1-12.

[7] Hovy E, Marcu D. Automated text summarization[C]//The Oxford Handbook of Computational Linguistics. USA: Oxford University Press,2005:583-598.

[8] Baxendale P B. Machine-made index for technical literature: An experiment[J]. IBM Journal of Research and Development, 1958, 2(4): 354-361.

[9] Edmundson H P. New methods in automatic extracting[J]. Journal of the ACM (JACM), 1969, 16(2): 264-285.

[10] Ramezania M, Feizi-Derakhshi M. Automated text summarization: An overview[J]. Applied Artificial Intelligence: An International Journal,2014, 28(2):178-215.

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

[12] Boguraev B, Kennedy C. Salience-based content characterisation of text documents[C]//Advances in Automatic Text Summarization. Cambridge: MIT Press, 1999:99-110.

[13] Barzilay R. Lexical chains for summarization[D]. Beer-Sheva: Ben-Gurion University of the Negev, 1997.

[14] Marcu D. From discourse structures to text summaries[C]//Proceedings of the ACL. Madrid:ACL,1997: 82-88.

[15] Barzilay R, Elhadad M. Using lexical chains for text summarization [C]//Advances in Automatic Text Summarization. Cambridge: MIT Press, 1999:111-121.

[16] Mihalcea R. Language independent extractive summarization[C]// Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics. Ann Arbor:ACL, 2005: 49-52.

[17] Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries[C]//Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. Barcelona:ACL,2004: 74-81.

[18] Papineni K, Roukos S, Ward T, et al. BLEU:A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics. Priladelphia: ACL, 2002: 311-318.

[19] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers [J]. Machine Learning, 1997, 29(2-3): 131-163.

[20] Kupiec J, Pedersen J, Chen Francine. A trainable document summarizer[C]//Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Seattle: ACM, 1995: 68-73.

[21] Paice C D. Constructing literature abstracts by computer: Techniques and prospects[J]. Information Processing & Management, 1990, 26(1): 171-186.

[22] Aone C, Okurowski M E, Gorlinsky J. Trainable, scalable summarization using robust NLP and machine learning[C]//Proceedings of the 17th International Conference on Computational Linguistics-Volume 1. Montreal: Association for Computational Linguistics, 1998: 62-66.

[23] Salton G, McGill M J. Introduction to modern information retrieval [M]. New York: McDraw-Hill, 1983:20-25.

[24] Hachey B, Grover C. Sentence extraction for legal text sum-

marisation[C]//Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence. Edinburgh: Professional Book Center, 2005: 1686 - 1687.

[25] Sharan A, Imran H, Joshi M L. A trainable document summarizer using Bayesian classifier approach[C]//Emerging Trends in Engineering and Technology, 2008. ICETET' 08. First International Conference on. Nagpur: IEEE, 2008: 1206 - 1211.

[26] Yousfi-Monod M, Farzindar A, Lapalme G. Supervised machine learning for summarizing legal documents[C]//Advances in Artificial Intelligence. Berlin Heidelberg: Springer, 2010: 51 - 62.

[27] Pera M S, Ng Y K. A Naive Bayes classifier for Web document summaries created by using word similarity and significant factors [J]. International Journal on Artificial Intelligence Tools, 2010, 19(4): 465 - 486.

[28] Lee J E, Park H S, Kim K J, et al. Learning to predict the need of summarization on news articles[J]. Procedia Computer Science, 2013, 24: 274 - 279.

[29] Aries A, Oufaida H, Nouali O. Using clustering and a modified classification algorithm for automatic text summarization [C]// IS&T/SPIE Electronic Imaging. Burlingame: SPIE, 2013: 9 - 11.

[30] Caruana R, Niculescu - Mizil A. An empirical comparison of supervised learning algorithms[C]//Proceedings of the 23rd International Conference on Machine Learning. New York: ACM, 2006: 161 - 168.

[31] Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257 - 286.

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

[33] Schlesinger J D, Okurowski M E, Conroy J M, et al. Understanding machine performance in the context of human performance for multi - document summarization[C]//Proceedings of the Workshop on Automatic Summarization. Gaithersburg: NIST, 2002: 71 - 77.

[34] Fung P, Ngai G, Cheung C S. Combining optimal clustering and Hidden Markov models for extractive summarization[C]//Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering - Volume 12. Morristown: Association for Computational Linguistics, 2003: 21 - 28.

[35] Dunlavy D M, O' Leary D P, Conroy J M, et al. QCS: A system for querying, clustering and summarizing documents[J]. Information Processing & Management, 2007, 43(6): 1588 - 1605.

[36] Dunlavy D M, Conroy J M, Schlesinger J D, et al. Performance of a three - stage system for multi - document summarization[C]//Proceedings of the Document Understanding Conference. Gaithersburg: National Institution of Standards and Technology, 2003: 153 - 159.

[37] Zajic D, Dorr B J, Lin Jimmy, et al. Multi - candidate reduction: Sentence compression as a tool for document summarization tasks[J]. Information Processing & Management, 2007, 43(6): 1549 - 1570.

[38] Wang Peng, Wang Haixun, Liu Majin, et al. An algorithmic approach to event summarization[C]//Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2010: 183 - 194.

[39] Darling W M, Song Fei. Probabilistic document modeling for syntax removal in text summarization[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short papers - Volume 2. Portland: Association for Computational Linguistics, 2011: 642 - 647.

[40] 刘江鸣,徐金安,张玉洁. 基于隐主题马尔科夫模型的多特征自动文摘[J]. 北京大学学报(自然科学版),2014(1):187 - 193.

[41] Gruber A, Weiss Y, Rosen - Zvi M. Hidden topic Markov models [C]//International Conference on Artificial Intelligence and Statistics. San Juan: AISTATS, 2007: 163 - 170.

[42] Juang B H, Rabiner L R. Hidden Markov models for speech recognition[J]. Technometrics, 1991, 33(3): 251 - 272.

[43] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the Eighteenth International Conference on Machine Learning (ICML - 2001). Williamstown: Morgan Kaufmann, 2001: 282 - 289.

[44] Saravanan M, Ravindran B, Raman S. Improving legal document summarization using graphical models[C]//Legal Knowledge and Information System, JURIX 2006: The Nineteenth Annual Conference, Paris: IOS Press, 2006: 51 - 60.

[45] Shen Dou, Sun Jiantao, Li Hua, et al. Document summarization using conditional random fields[C]//Proceedings of the 20th International Joint Conference on Artificial Intelligence. Hyderabad: Morgan Kaufmann Publishers Inc., 2007: 2862 - 2867.

[46] 吴晓锋,宗成庆. 一种基于LDA的CRF自动文摘方法[J]. 中文信息学报,2009(6):39 - 45.

[47] 邓箴,包宏. 基于条件随机场的中文自动文摘系统[J]. 西安石油大学学报(自然科学版),2009(1): 96 - 99,102,114.

[48] 张龙凯,王厚峰. 文本摘要问题中的句子抽取方法研究[J]. 中文信息学报,2012(2):97 - 101.

[49] Batcha N K, Aziz N A, Shafie S I. CRF based feature extraction applied for supervised automatic text summarization[J]. Procedia Technology, 2013(11): 426 - 436.

[50] Wong Tak-Lam, Lam Wai. Learning to extract and summarize hot item features from multiple auction Web sites[J]. Knowledge and Information Systems, 2008, 14(2): 143 - 160.

[51] Lin Chin-Yew. Training a selection function for extraction[C]// Proceedings of the Eighth International Conference on Lnformation and Knowledge Management. Cambridge: ACM, 1999: 55 - 62.

[52] Osborne M. Using maximum entropy for sentence extraction[C]// Proceedings of the ACL - 02 Workshop on Automatic Summarization - Volume 4. Morristown: Association for Computational Lin-

guistics, 2002: 1 - 8.

[53] Yeh Jen-Yuan, Ke Hao-Ren, Yang Wei-Pang, et al. Text summarization using a trainable summarizer and latent semantic analysis[J]. Information Processing & Management, 2005, 41(1): 75 - 95.

[54] Murray G, Renals S, Carletta J, et al. Evaluating automatic summaries of meeting recordings[C]//Proceedings of ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization (MTSE). Ann Arbor: ACL, 2005: 33 - 40.

[55] Kaikhah K. Automatic text summarization with neural networks [C]// IEEE International Conference on Intelligent Systems. Varna: IEEE, 2004: 40 - 44.

[56] Fuentes M, Alfonseca E, Rodríguez H. Support vector machines for query - focused summarization trained and evaluated on pyramid data[C]//Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Stroudsburg: Association for Computational Linguistics, 2007: 57 - 60.

[57] Aker A, Cohn T, Gaizauskas R. Multi - document summarization using A * search and discriminative training[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2010: 482 - 491.

[58] Lin Shih-Hsiang, Chang Yumei, Liu Jiawen, et al. Leveraging evaluation metric - related training criteria for speech summarization[C]// Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on. Dallas: IEEE, 2010: 5314 - 5317.

[59] Xie Shasha, Liu Yang. Improving supervised learning for meeting summarization using sampling and regression[J]. Computer Speech & Language, 2010, 24(3): 495 - 514.

[60] Wong Kam-Fai, Wu Mingli, Li Wenjie. Extractive summarization using supervised and semi - supervised learning[C]//Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Manchester: Association for Computational Linguistics, 2008: 985 - 992.

[61] Zhang Jian Justin, Chan Ricky Ho Yin, Fung Pascode. Extractive speech summarization by active learning[C]//Automatic Speech Recognition & Understanding, 2009. ASRU 2009. Merano: IEEE Workshop on. Merano: IEEE, 2009: 392 - 397.

[62] Frakes W B, Baeza - Yates R. Information retrieval: Data structures and algorithms[M]. London: Prentice Hall, 1992: 100 - 133.

[63] Sankarasubramaniam Y, Ramanathan K, Ghosh S. Text summarization using Wikipedia[J]. Information Processing & Management, 2014, 50(3): 443 - 461.

[64] Ramanathan K, Sankarasubramaniam Y, Mathur N, et al. Document summarization using Wikipedia[C]//Proceedings of the First International Conference on Intelligent Human Computer Interaction. New Delhi: Springer India, 2009: 254 - 260.

[65] Xu Danyun, Cheng Gong, Qu Yunzhong. Preferences in Wikipedia abstracts: Empirical findings and implications for automatic entity summarization[J]. Information Processing & Management, 2014, 50(2): 284 - 296.

[66] 孙建军. 网络公共信息资源利用效率影响因素实证分析[J]. 图书情报工作,2012,56(10):35 - 40.

[67] 孙建军,屈良. 基于博客的链接分类体系设计[J]. 情报科学,2012(3):321 - 326,346.

[68] 孙建军,屈良. 加权入链数:对链接分析中绝对入链数的修正[J]. 情报科学,2012(2):161 - 165,172.

[69] Zhang Pei-ying, Li Cun-he. Automatic text summarization based on sentences clustering and extraction[C]//Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on. Dalian: IEEE, 2009: 167 - 170.

[70] ogly Alguliev R M, ogly Alyguliev R M. Automatic text documents summarization through sentences clustering[J]. Journal of Automation and Information Sciences, 2008, 40(9): 53 - 63.

[71] Amini M R, Usunier N, Gallinari P. Automatic text summarization based on word - clusters and ranking algorithms[C]//Advances in Information Retrieval. Berlin Heidelberg: Springer, 2005: 142 - 156.

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

[73] Lei Yu, Ren Fuji. A study on cross - language text summarization using supervised methods[C]//Natural Language Processing and Knowledge Engineering, 2009. NLP - KE 2009. International Conference on. Dalian: IEEE, 2009: 1 - 7.

A Review on Machine Learning Oriented Automatic Summarization

Cao Yang Cheng Ying Pei Lei

School of Information Management, Nanjing University, Nanjing 210093

[**Abstract**] This paper probes into the process of automatic summarization based on machine learning, including features selection, algorithm selection, model training, abstracts extraction, model evaluation. The Review focuses on three main machine learning algorithms: Naive Bayes, Hidden Markov Model and Conditional Random Fields, mainly elaborating the idea of these algorithms, summarizing related research, and giving reflections. Then it discusses the common problems with three machine learning algorithms, including training methods, collaborative training and active learning, category balance, terms distribution. In the end, future research directions are explored.

[**Keywords**] automatic summarization machine learning NB HMM CRF