人工智能在方志编纂中的运用初探

——以百度人工智能在《浙江通志·大事记》编纂中的应用为例*

赵海良

提　要：大事记是新方志的重要组成部分，在新修《浙江通志》中就单列有《大事记》一卷。人工智能是当前计算机领域的一门新兴学科，在资料的自动获取以及文字处理方面有着独特的优势，可将大量人力从繁杂的基础工作中解脱出来。以百度人工智能在《浙江通志·大事记》编纂中的应用为例进行探讨，以期为相关技术在方志编纂中的应用提供启发和思路。

关键词：人工智能　地方志　大事记　编纂

人工智能（Artificial Intelligence）的概念最早于1956年被提出。随着大数据、高性能计算以及深度学习技术的快速发展，人工智能已经衍变成为用于模拟、延伸人的智能的新兴技术，主要研究领域包括机器人、语言识别、图像识别、自然语言处理和专家系统等。① 2014年7月，美联社与科技公司Automated Insights合作开发了Wordsmith人工智能写作平台，使其能够自动编写企业财报，该平台几秒钟便能生成一篇150—300单词的新闻快讯，可以大大提高写作效率。② 有学者认为，与传统的创作相比，人工智能写作有其独特优势，主要体现在文本采集精准化、文本加工高效化、文本定制个性化等方面。③ 国内目前对此方面的研究，大多集中于新闻写作且进展显著，已大规模运用于财经、体育比赛、地震预报、交通监控和社交网络等相关领域的新闻创作。④

目前，人工智能在方志编纂领域的应用研究尚属空白，大事记作为现代方志7种体裁之一，存在有别于其他方志体裁的体例特点和要求，这使得人工智能在大事记编纂中的应用变得可能。百度人工智能是由百度公司研发的人工智能服务平台，提供了语音技术、图像技术、自然语言技术等多项场景能力和解决方案，也是目前国内人工智能领域研究的先驱。笔者以参与《浙江通志·大事记》编纂为契机，尝试将百度人工智能相关技术运用到大事记编纂中，以提高编纂效率。

一　大事记的体例和特点

“志书对大事的记述，古已有之，但对一个地方从古至今或某个确定的年段的大事进行综合记述，专门集成一卷（编），却是现代志书的新创举”⑤，中国地方志指导小组印发《关于地方

---

*　此文撰写时，《浙江通志·大事记》尚未正式出版，文中所述均指《浙江通志·大事记》终审稿。

① 参见王燕鹏、韩涛、赵亚娟、陈芳、王思培：《人工智能领域关键技术挖掘分析》，《世界科技研究与发展》2019年第8期。

② 参见吕倩、任媛媛：《颠覆还是辅助？“新闻+人工智能”的实践与反思》，《青年记者》2018年第30期。

③ 参见李君婷：《人工智能写作发展前景探析》，《新闻研究导刊》2019年第13期。

④ 参见黄国春：《人工智能新闻写作的路径探析》，《出版广角》2019年第15期。

⑤ 李云章：《编纂地方志大事记之管见》，《福建史志》2018年第3期。

志编纂工作的规定》明确“地方志的体裁一般应包含述、记、志、传、图、表、录等”。其中“记”即为大事记，是一种按时间顺序客观记载特定行政区域、政府部门或事业单位在一定时期内发生的自然、政治、经济、文化、社会等方面大事要事的应用性文献。中国地方志指导小组印发的《地方志书质量规定》规定，大事记需达到“选录人事得当，重要事项不漏，时间、地点、人物（单位）、结果等要素齐备”等要求，但对大事记该采用何种体例，并未做限定。

从大事记编纂规律及编纂实践来看，其体例大致有 3 种：“（一）编年体。以时系事，一事一记。按照事件发生的时间，逐年、逐月、逐日的记叙。（二）纪事本末体。以事系文，着重于事件的始末，以事件为中心，按其时间次序做系统叙述。（三）编年体和纪事本末体相结合。即以编年体为主，对于某些特定的事件，在其开始时间或结束时间（如一项重点工程建设）做系统记叙。”①

就大事记的编纂过程而言，一般应遵循以下几个原则：一是一事一记原则。大事记所有条目必须做到一事一记，也就是发生的大事只能记录在大事记的一个条目中，一个条目中只能记录一件大事。绝不能出现几件事放在一个条目中，或一天内发生的数件大事记在一个条目中，或一件事记几次等情况。二是要素完整原则。大事记所要记述的内容包括时间、地点、单位、人员、事件等要素，也就是要记述在什么时间、什么地点、什么单位或人员发生了什么事、事件内容是什么、有什么影响等。三是客观真实原则。“大事记者，列其事之目而已，无所褒贬抑扬也”②，大记事所记载的事件都是真实发生过的，不能把没有发生的事凭想象写进大事记中，也不能违背所发生事件的历史原貌，随意进行包装、修改或创造，应尊重历史事实、尊重事件的原貌。四是简明精练原则。对每件大事的记述要简明扼要，用最简短的语言把事件过程和内容记述清楚。同时，要注意详略得当，重大事件、重要事件、首发事件、影响深远的事件适当详记，次要事件、经常性事件尽量简记，有些甚至可以不记。

二　人工智能技术在大事记编纂中的应用

传统的人工智能写作是计算机语言处理的结果，通过将数据输入计算机，再套用固定的算法将其重新排列组合并以特定格式呈现，目前在新闻创作领域广泛应用。③ 其大致可分为“获取和消化信息、分析数据和信息、选择新闻点套用模板优化、输出并发布”④ 4 个步骤。

大事记既有与新闻类似的一事一记、要素完整原则，又有其特殊的客观真实、简明精练原则。根据大事记的体例特点及编纂原则，人工智能在其编纂中应用的基本原理可概述为：通过计算机自动获取特定来源的资料，然后经语义分析、情感分析、文本摘要等技术进行优化精简，再套用固定算法将其重新排列组合，并以特定格式呈现。对于大事记编纂所涉及人工智能相关技术的具体实践应用，可阐述如下。

（一）自动化获取参考资料

人工智能写作，其本质是对已有信息的重新组合，相关技术的运用必须依靠大数据，没有数据的支撑，人工智能也是“巧妇难为无米之炊”。在大事记的编纂中，这个大数据即为各种参考

① 吕金祥、李海艳、谢奎江：《如何编写大事记》，《中国地方志》2011 年第 11 期。

② 吕祖谦：《大事记解题》，《四库全书》，台北“商务印书馆”影印文渊阁本，1986 年，第 324 册，第 124 页。

③ 新闻报道一般具有相对固定的 5W 原则，即何时（when）、何地（where）、何事（what）、为何（why）、何人（who）。

④ 王二龙、李明非：《“机器新闻写作”：历史、现状与应对策略》，《新闻战线》2019 年第 10 期。

资料。梁启超曾说："方志之著述，非如哲学家、文学家之可以闭户瞑目其理想而遂有创获也，其最主要之工作在调查事实，搜集资料。"① 就大事记而言，其所载大事，非编纂者凭空想象，均有资料来源以为佐证。在已出版的《浙江历史大事记》的编纂过程中，"仅主要参考文献就有历代典籍类 119 种，今人著作类 132 种，档案、文献资料类 92 种，报纸、杂志类 33 种，新编地方志 133 种，共计 509 种"②。

"报刊资料是历史活动的真实记录，所载内容也是极其丰富而又相当具体的，基本情况和线索还是能反映出来，所记时间、地点、人名等多较准确，不失为珍贵的历史资料。"③ 在《浙江通志·大事记》的编纂实践中，《浙江日报》是极为重要的参考资料。笔者对《浙江日报》电子版进行分析，发现可通过"http：//zjrb. zjol. com. cn/html/年 - 月/日/node_ 18. htm"的固定格式访问特定日期的报纸电子版。因此，笔者可以运用网络"爬虫技术"，自动从《浙江日报》电子版提取新闻报道。爬虫技术，又称网络蜘蛛和网络机器人，主要用于收集互联网上的各种资源。它是搜索引擎的重要组成部分，是一个可以自动提取互联网上特定页面内容的程序。④ 通过网络爬虫技术，可在极短时间内从网络上获取海量信息资源，并以此作为人工智能技术运用所需的大数据基础。通过网络爬虫获取的数据是未经筛选的数据，而大事记，顾名思义所记必为大事，一份报纸的新闻报道少则几十篇，多则上百篇，全部收录进大事记显然不现实亦不科学，因此需明确收录标准和筛选的原则。

目前，大事记的编纂尚无明确统一的收录标准。一般认为，"大事记是记本地发生的，对当地（乃至对外地区甚至全国）的国计民生和社会历史的发展有较大影响的事情"⑤，而对这一原则的把握，则完全依靠编纂者的主观判断。若要依赖人工智能技术进行分析筛选，需要制定相对客观的收录标准。

笔者试通过"百度指数"制定一个量化的标准，用于筛选大事。百度指数是以百度海量网民行为数据为基础的数据分析平台，其主要有两个特点：（1）数据来源于网民搜索行为；（2）关键词为数据基础。这两个特点使得百度指数被广泛运用于舆情分析、市场趋势研究等各领域。因此可根据相关新闻的百度指数变化，来确定其重要性，以此进行筛选。

以 2018 年 11 月 7 日《浙江日报》有关"第五届世界互联网大会开幕"的相关报道为例⑥，对于此篇报道，百度人工智能自动识别出其关键词为"世界互联网大会"（见图 1）。与此同时，在互联网上，以"世界互联网大会"为关键词的百度指数激增（见图 2），说明在此段时间内，"世界互联网大会"是社会所关注的重要事件，较为符合大事记的收录标准。通过以上方式，人工智能即可通过量化的方式，筛选出较为重要、达到大事记收录标准的新闻报道。

（二）清洗整理数据

得益于网络爬虫技术的发展，使得从特定网站抓取特定内容的信息数据已变得非常简便，人工智能技术在大事记编纂中的难点在于如何使这些信息的记叙符合地方志编纂规范。

方志与其他文体最大的区别是对词、句的运用，"善于驾驭词语，是编纂方志的重要基础"，

---

① 陈其泰、陆树庆、徐蜀：《梁启超论著选粹》，广东人民出版社，1996 年，第 975 页。

② 李志庭：《略谈〈浙江历史大事记〉的编纂特点》，《中国地方志》2011 年第 11 期。

③ 黄晓明：《资料收集与资料长编编写》，2017 年 10 月 24 日，http：//www. dag. ecnu. edu. cn/01/a7/c11140a131495/page. htm，2019 年 7 月 20 日。

④ 参见吴永聪：《浅谈 Python 爬虫技术的网页数据抓取与分析》，《计算机时代》2019 年 8 期。

⑤ 吕金祥、李海艳、谢奎江：《如何编写大事记》，《中国地方志》2011 年第 11 期。

⑥ 详见报道电子版 http：//zjrb. zjol. com. cn/html/2018 - 11/07/content_ 3177938. htm? div = - 1。

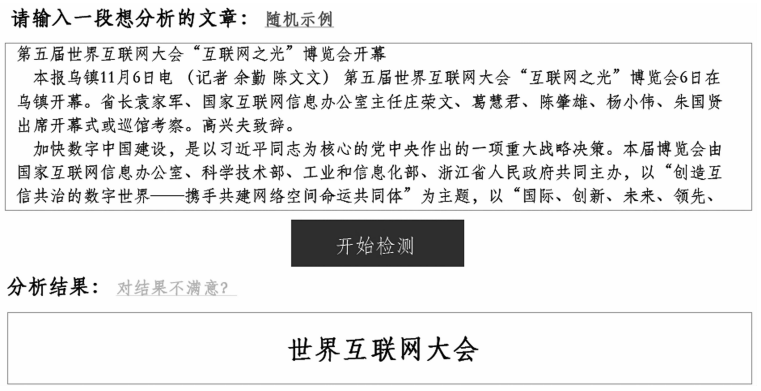

图 1

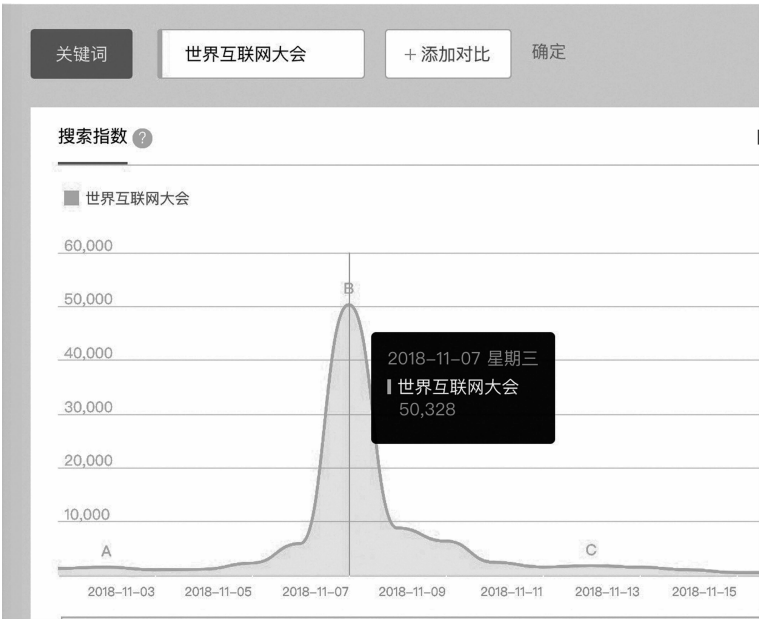

图 2

对其也有着“以少驭多”“去浮存实”“中性选词”等诸多要求。① 但新闻报道往往对事物的描述有所褒贬，为使之能符合志书的编纂要求，需对相关数据进行清洗整理。

资料的基础是文字，而文字也是人与人之间沟通和交流的桥梁，是传递信息的关键。在语言学中，通常将人类采用的语言文字统称为自然语言。自然语言处理技术是语言学与计算机学的交叉学科，其目标是通过人工智能技术，赋予计算机各种语言知识，使其能够接受人们采用自然语言给它输入的命令，理解人们所要表达的意思，实现从一种语言到另一种语言的翻译等功能。②

在自然语言的使用中，存在着主、谓、宾、定、状、补的文法，也存在着名、动、形、副等

① 参见毛东武：《方志词语运用种种》，《黑龙江史志》1996 年第 3 期。

② 参见陈开昌：《自然语言处理技术中的中文分词研究》，《信息与电脑》（理论版）2016 年第 19 期。

词性。自然语言处理领域，英语有其独特的优势，因其文法中，用空格来加以区分不同的词。而中文文法中，所有词都是连接在一起的，因此人工智能在大事记编纂中的第一步是要对待处理的文本进行分词。所谓分词是指以词作为基本单元，运用计算机自动地对中文文本进行词语的切分，将完整的一句话根据其语义分拣成一个词语项集。①

分词是中文信息处理的一项特有过程，此过程离不开中文语料库的参与。语料库是用来存放语言材料的数据库（或仓库），被视作语言研究与应用的重要基础资源。语料库构建的本质目标是给机器翻译、数据挖掘、知识关联、智能检索、语义标引等模块与功能的实现提供基础与支撑。② 语料库可分为通用语料库和专题语料库。顾名思义，所谓通用语料库是指未针对任何行业进行优化筛选的语料库，典型代表有国家现代汉语语料库及国家语委现代汉语语料库。而专题语料库相较于通用语料库，其针对性更强，所收录的语料，基本是某一行业所特有或特定的。

以百度人工智能的通用语料库为例，对于“中国社会科学院”一词，其识别为“机构名”（见图3），而“中国地方志指导小组办公室”一词，未能正确识别，而是将其进行了进一步的分词处理（见图4）。

图 3

图 4

① 参见李康康、龙华：《基于词的关联特征的中文分词方法》，《通信技术》2018 年第 10 期。

② 参见卫乃兴：《语料库语言学的方法论及相关理念》，《外语研究》2009 年第 5 期。

究其原因在于“中国地方志指导小组办公室”不在其通用语料库中。地方志中有众多的专门术语，制定方志专题语料库，是人工智能技术应用于方志编纂的关键之一。而百度人工智能所提供的“个性化词表”功能，“支持用户上传自定义专有名词词表，适用于特定领域内大量行业术语和专有名词以及互联网新兴词汇的识别”，可较好地解决该问题。

（三）文本资料的情感分析

清代方志学家章学诚认为“志乃史体，原属天下公物”①，认为修志应该“据事直书，善否自见”②。由此可见，述而不论、直陈其事，是志书有别于其他文体的重要特点之一。词有褒义、贬义及中性词之分，在志书编纂中，应尽可能选取中性词，以使记叙客观公正。而一般新闻报道均有惩恶扬善的倾向，褒贬之意较为明显，文字中也存在大量情感色彩强烈的用词及用语，这与大事记的编纂要求不符，因此需将这些词去除，这就需要运用人工智能的情感分析技术。

情感分析，也称情感倾向性分析，是指对文本的情感倾向进行研究的过程③，其目的是判定某个词语、句子或文章的情感倾向是正向、负向还是中立的，同时还可判断情感程度，即情感的强烈等级。

在新闻报道中，普遍存在着大量的情感词。所谓情感词是指可以表达主观感受、情感或者意见的词汇或短语。情感词典则是情感词组合而成的合集。构建中文情感词典的方式大致可以归纳为通过手工方式构建和通过语料库方式构建两种。④ 现在被广泛使用的英文情感词典有 WordNet、Senti Word Net 等，在中文领域，中国知网的中文情感词典则被学界广泛使用，其收录有表示程度的词语 219 个、表示正向态度的词语 836 个、表示反向态度的词语 1254 个。⑤

情感分析的主要原理是将文本分词处理后的结果，逐个在情感词典查找对应的情感值，随后分析整个语句的情感值。百度人工智能可“针对带有主观描述的中文文本，自动判断该文本的情感极性类别并给出相应的置信度”。其具体过程是通过人工智能的词法分析技术，进行分词与词性标注处理，去掉文本中的一些词语，如“的”“了”“么”等功能性、对表达文本主题并无影响的词语，随后通过情感词词典，查找并匹配相关词汇的情感值，进而对整个语句进行情感评分。

表 1　中国知网中文情感分析用词语集（部分）

正面情感	爱、赞赏、快乐、感同身受、好奇、喝彩、魂牵梦萦、嘉许
负面情感	哀伤、半信半疑、鄙视、不满意、不是滋味儿、后悔、大失所望
正面评价	不可或缺、部优、才高八斗、沉鱼落雁、催人奋进、动听、对劲儿
负面评价	超标、华而不实、荒凉、混浊、畸轻畸重、价高、空洞无物
程度级别	百分之百、倍加、备至、不得了、不堪、不折不扣、彻头彻尾
主　　张	觉得、看得出来、窥见、领教、听说、痛感、预感、自觉

① 章学诚著，叶瑛校注：《文史通义校注》，中华书局，1994 年，第 544 页。

② 章学诚：《答甄秀才论修志第一书》，《章学诚遗书》，文物出版社，1985 年，第 137 页。

③ 参见赵妍妍、秦兵、刘挺：《文本情感分析》，《软件学报》2010 年第 8 期。

④ 参见阳爱民、林江豪、周咏梅：《中文文本情感词典构建方法》，《计算机科学与探索》2013 年第 7 期。

⑤ 参见 http：//www. keenage. com/html/c_ bulletin_ 2007. htm。

以《浙江通志·大事记》为例，某条记载：

12月20日　省高级人民法院对国内第一宗农民告县政府的苍南县舥艚镇河道清障案作出终审判决。审理认为，原告包郑照侵占河道，违章建房，苍南县政府在多次教育、限期自行清障无效后，决定强行清障，是合法的。

以上记叙要素完整，语意清晰，初看之下并无不妥，而人工智能却给出了“情感偏负向”的分析结果（图5）。细究之下不难发现，“强行清障”一词贬义较为明显，在将其改为“进行清障”后，系统给出了“情感偏中性”的分析结果（图6）。

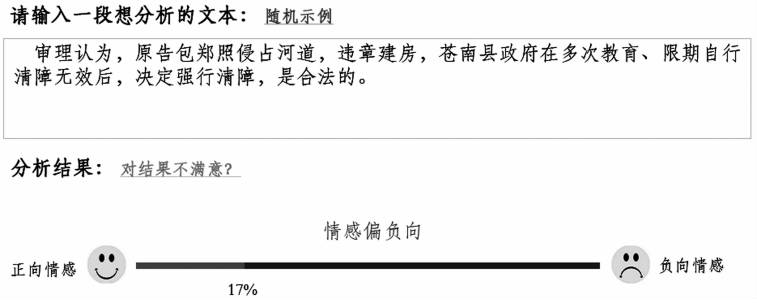

图5

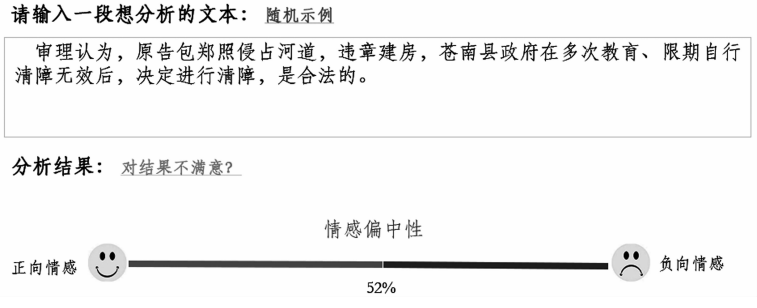

图6

（四）关键词的获取及信息摘要

通常情况下一篇文章是由几个对该篇文章主要内容具有概括功能的词或短语组成，称之为关键词。通过关键词可在保留资料原意的同时对文章内容进行压缩，也可以快速获取文章的大意并进行结构化的初步了解。

关键词的自动获取技术是文本信息挖掘领域的一项关键技术。TF-IDF（term frequency-inverse document frequency）算法是一种评估文本中一个词或短语重要程度的统计方法，也是常用的关键词提取算法。其基本思想为词或短语的重要性随该词或短语在文本中出现的次数增加而增加，但随着其在语料库中出现的次数增加而减少。该算法可简单理解为“若某一词或短语，在文章中大量出现，但在日常中却并不常见，则该词或短语是文章关键词的概率大”。例如“的、是、了”等词，在任何文章中都会大量出现，却非任何文章的关键词。

以百度人工智能获取“浙江首个世界遗产江郎山申遗成功”新闻报道的关键词为例①，400 余字的报道，其识别出核心关键词是“江郎山”和“世界遗产”（见图 7）。

图 7

从资料中提取出能够表达文章主题的关键词，并将关键词整合排列成远小于原始文章篇幅的一小段文本即是文本摘要。② 与关键词相比，摘要是对文章的高度概括，表明了文章的主要内容，它比标题、关键词等更具有代表性，同时增加了连贯性、可读性的要求。而大事记的编纂也有同样的要求：“以较少的文字载录尽可能多的信息，文字凝炼集中，只要把事件最主要的特征实质叙述出来即可，不必过详过细，文体风格要明快，文辞简洁流畅，切忌套话、空话、长话，每条记事文字多少，视其具体内容而定，要以精炼简短而又说明问题为准。”③

资料的自动摘要是人工智能中受到广泛关注的领域，以往各种方法重点考虑的是文章包含的信息量，忽视了摘要本身的语句连贯性，生成的摘要信息可读性不强。

百度人工智能提供的新闻摘要功能，可“结合传统语义特征和深度学习模型，充分考虑段落分布和篇章结构，准确计算新闻语句的重要性，对新闻内容进行全面的语义理解与分析”，同时也可“根据需求灵活控制摘要长度，自动抽取关键信息，形成摘要结果”④。

以《浙江通志·大事记》有关“浙江吉利汽车收购瑞典沃尔沃汽车”的相关记载为例，其资料来源于《浙江日报》2010 年 8 月 3 日头版⑤，百度人工智能将相关新闻报道的全文自动生成了以下摘要（见图 8）：“8 月 2 日中国浙江吉利控股集团有限公司 2 日在伦敦宣布，已经完成对美国福特汽车公司旗下沃尔沃轿车公司的全部股权收购。同时，吉利控股集团宣布，德国大众汽车公司北美区前首席执行官斯特凡·雅各比自 8 月 16 日起将正式就任沃尔沃总裁兼首席执行官，并将加入沃尔沃轿车公司董事会。吉利控股集团上月 15 日宣布，李书福将担任沃尔沃董事长，沃尔沃前总裁兼首席执行官汉斯—奥洛夫·奥尔松将担任副董事长。”

此段摘要，在用词规范、言语精练等方面虽不及人工撰写，尚需斟酌，但已然基本达到语义

① 详见报道电子版 http：//zjrb. zjol. com. cn/html/2010 - 08/03/content_ 479151. htm? div = - 1。

② 参见刘子平：《基于主题句语义融合的多文档摘要算法研究》，重庆大学计算机系硕士学位论文，2016 年，第 27 页。

③ 付华、张方芳：《浅谈大事记的写法》，《东疆学刊》1998 年第 15 卷第 4 期。

④ 褚晓敏、朱巧明、周国栋：《自然语言处理中的篇章主次关系研究》，《计算机学报》2017 年第 4 期。

⑤ 详见报道电子版 http：//zjrb. zjol. com. cn/html/2010 - 08/03/content_ 479154. htm? div = - 1。

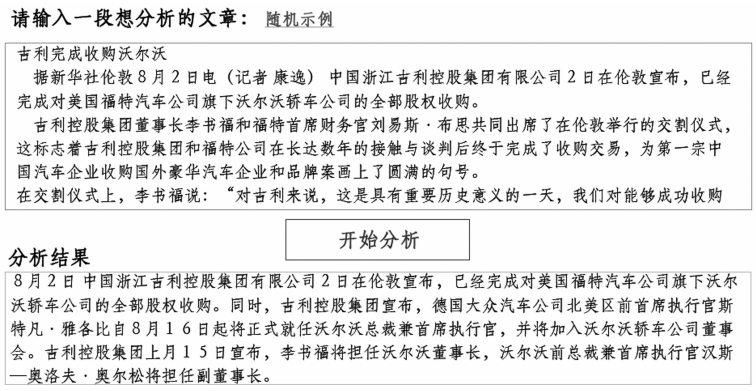

图 8

清晰、客观真实的编纂要求。

综上所述，笔者认为，人工智能技术在大事记中的编纂运用具体可分为以下几个过程：

第一，寻找资料来源，通过网络爬虫技术获取数据，并根据资料的重要程度，自动初步筛选出拟收录的资料；

第二，清洗整理数据，通过词性分析、情感分析等技术，结合大事记的编纂规范，去除资料中的无效信息；

第三，将经过分析处理的资料进一步精简，通过提取关键词、信息摘要等技术，将有价值的信息通过排列、组合形成较为符合大事记规范的文档。

在笔者的实践中，运用网络爬虫技术及百度人工智能所提供的涉及前文所述相关功能的API①，实现了批量从已出版的《浙江日报》电子版抓取新闻报道，并经过情感分析、信息摘要等人工智能处理，自动生成指定长度的摘要的功能。

结　语

诚然，人工智能技术的飞速发展，对当前社会生活的各领域都产生了深远的影响，通过算法、机器学习，信息的获取将会变得越来越智能，所得到的信息也将更翔实、准确和可靠，也许在某种程度上能达到客观意义上的拟人效果。但地方志编纂离不开优秀的方志工作者，更需要方志工作者的主观意识及人文精神。在人工智能的冲击下，广大方志工作者应善于利用新技术，采用新方法，把大量人力从繁杂的基础工作中解脱出来，以此提高编纂效率。

（作者单位：浙江省人民政府地方志办公室）

本文责编：周　全

① API（Application Programming Interface，应用程序接口）是一些预先定义的函数，或指软件系统不同组成部分衔接的约定。目的是提供应用程序与开发人员基于某软件或硬件得以访问一组例程的能力，而又无需访问原码，或理解内部工作机制的细节。