企业文档自动生成系统的设计与实现

柴丽萍　杜一玮　李题印　刘宗泽　张　屏
国网浙江省电力有限公司　杭州　310000

摘　要：文档信息资源是企业知识资源的重要组成部分，无纸化办公的盛行要求企业配备文档的自动生成与智能服务系统。企业文档的自动生成是在机器学习、自然语言处理等新技术背景下提升企业办公、协作效率的重要方式，能够促进人力资源的合理分配，推进企业智慧化办公、智慧化管理。

关键词：企业文档；自动生成；系统设计；智能服务

Design and implementation of enterprise document automatic generation system

Liping Chai, Yiwei Du, Tiyin Li，Zongze Liu ,Ping Zhang
Zhejiang Electric Power Corporation Hangzhou 310000

Abstract: Document information resources are an important part of enterprise knowledge resources. The prevalence of paperless offices requires enterprises to equip with automatic document generation and intelligent service systems. The automatic generation of enterprise documents is an important way to improve the efficiency of enterprise offices and collaboration under the background of new technologies such as machine learning and natural language processing. It can promote the rational allocation of human resources and the intelligent office and management of enterprises.

Keywords: Enterprise document; Automatic generation; System design; Intelligent service

引言

随着知识经济时代的到来，知识成为第一生产要素，如何开发利用企业知识资源为企业重大决策提供知识支持，已经成为企业全面提升管理能力的关键要素之一。文档信息资源是企业知识资源的重要组成部分，企业对文档智能化管理的需求也正进一步提升。企业文档的自动生成与智能服务系统基于先进的信息技术和管理理念，能够提高企业文档知识资源的利用效率，实现企业智能化管理，将是文档管理的高级形式。而诸如自然语言处理技术与自然语言处理增强算法等先进技术手段的不断应用，也为这一管理方式革新提供了必要条件。

自然语言处理（Natural Language Processing, 简称NLP）是目前人工智能领域的主要研究课题之一[1]。自然语言处理作为人工智能的一个子领域，与机器学习和深度学习算法具有天然联系。机器学习能够依据大量的结构化数据训练，自动适应人的行文与沟通习惯，进而实现机器翻译、文本生成等功能。这其中，文本生成这一功能受到人们的广泛关注，人们希望通过机器学习理解语义，最终模仿人类编辑语言文本。目前，文本生成主要有提取和生成两种途径，其中提取技术已相对完善且被广泛应用，而生成技术还有较大发展空间。在利用机器处理文本的过程中，NLP能够构建和分析文本的内容框架，通过不断学习训练和迭代升级，最终使机器具有与人类相当甚至超越人类的阅读能力和处理速度。在这一过程中，企业的历史文档会促使机器了解文字组合方式以及语境下的实际意义，进而提供灵活多变可定制的个性化解决方案。

因此，结合自然语言处理、机器学习、大数据分析等新晋技术，设计与实现公文文档自动生成系统[2]，对历史数据和文档进行有效的开发与利用，挖掘其潜在的知识价值，有利于提高相关工作人员的工作效率，进一步增强文档质量与管理质量[3]。基于此，本文结合企业文档自动生成全流程的多方位、多角度需求与技术领域的拓展，构建了企业自动文档生成系统，通过企业文档的自动生成减少企业文档管理过程中的人力工作，提高企业办公效率，实现企业智慧化办公、智慧化管理。

一、研究述评

企业文档自动生成系统功能实现的核心是技术应用，因此有必要对相关技术的发展现状作有效梳理。在诸多人工智能技术之中，自然语言处理技术为系统实现提供了主要保障，自然语言生成是其中一个极为重要实践应用方向。其主要研究内容多偏向于将结构化数据转换为自然人可读可理解的自然语言。在此基础上，文档自动生成系统其作用可以理解为二进制符号语言与自然语言之间的的翻译器[4]。

文本生成技术是指根据操作者给定的信息和数据自

动生成相关文本，例如生成天气预报文本、体育新闻、财经报道、医疗报告等。文本生成技术具有极强的应用前景，目前该领域的研究已经取得了较大进展，业界已经构建出面向不同领域和应用的多个成熟的系统。文本生成技术在天气预报领域应用最为成功，业界研制了多个系统对天气预报数据进行总结，生成天气预报文本。例如，FoG 系统能够从用户操作过的数据中生成双语天气预报文本。业界面向其他领域也研制了多个文本生成系统，例如针对空气质量的文本生成系统，针对财经数据的文本生成系统，面向医疗诊断数据的文本生成系统等。其中医疗诊断数据的文本生成系统 BT-45 能够为新生儿重症监护病房（NICU）的监控数据生成文本摘要，帮助医生进行决策。由于文本生成技术的巨大应用价值，制造业界成立了多家从事文本生成的企业，能够为多个行业基于行业数据生成行业报告或新闻报道，从而节省大量的人力[5]。比较知名的公司有 ARRIA、AI、NarrativeScience 等。如 AI（Automated Insights）是一家美国人工智能公司，目前能为包括金融、个人健身、商业智能、网站分析等在内的多个领域内的数据生成文本报告，其核心技术为 WordSmithNLG 引擎。目前，AI 公司已经为美联社等多家单位生成数亿篇新闻报道，造成了巨大的影响力。NarrativeScience 则是根据美国西北大学的研究项目 StatsMonkey 发展而来，其核心技术为 QuillNLG 引擎。在标准化与规范化研究上，Dale[6] 和 Reiter[7] 也相继提出了自然语言生成的典型阶段与关键步骤。这对自然语言处理的相关细节提供了较为具体的理论依据。

上述技术的成熟与发展，不仅推动了企业数智化发展的进程，也顺应了无纸化办公的行业趋势。当前，部分企业已经认识到企业生产经营过程中历史数据与文档的重要价值，并采取了一系列措施对其中蕴含的知识进行开发与管理。除此之外，某些特定部门的工作人员专门从事文字撰写工作，这些文档具有较高的规范性和格式，从以往的文档中提取有价值的段落以完成文稿的撰写，对于提高公文质量和写作速度也具有重要意义。综上所述，基于自然语言处理技术设计文档自动生成系统，能够较好地实现历史文档的组织管理，进一步完善机器对文本内容的提取与生成，有效辅助相关业务的辅助编辑，能够有效提高工作效率。

二、企业自动文档生成系统架构设计

系统平台主要负责实现基于企业文档大数据的领域知识图谱构建、基于深度学习的供电企业文档自动化创作以及企业文档自动生成系统平台部署三项主要内容。具体如下：

（1）前端模块，实现本地文档上传、内部系统集成、第三方接入集成、浏览器以及接入网关功能；

（2）业务中台，实现文档知识积累、文档知识利用、文档知识分享及文档知识创新功能；

（3）基础中台，提供系统功能基础支撑，实现加密与共享、权限控制、快速索引、协同编辑、用户管理、版本控制、模板库及文档预览功能；

（4）基础技术支撑，与业务中台及基础中台连接，提供系统级别基础技术支撑，包括监控、日志、存储、消息服务及任务调度功能；

（5）知识库模块，提供灵活便捷的多级分类目录及全面的多维全权限控制，快速构建知识库，方便所有员工持续高效的将文档知识积累到知识库中；

（6）文本生成模块，包括：获得文本生成模型和调用文本生成模型两个阶段。第一阶段包括数据预处理、深度学习算法模型构建、训练深度学习模型和获得文本生成模型模块。第二阶段包括：接受用户输入的文本、提取用户输入文本的特征信息、调用文本生成模型和生成与用户输入文本的特征信息相匹配的文本模块。第一阶段采用深度学习算法模型，使得训练过程更加自动化，免去了过多的人工干预，训练过程采用一系列训练策略，使得文本生成模型生成的文本可读性更强。第二阶段，对用户输入信息进行分类，识别用户意图，跟据用户的意图生成出用户想要的文本。

（7）企业文档知识平台应用模块，包括知识文档检索、知识信息门户和知识地图功能。

三、企业自动文档生成系统应用实践

依据上文中定义的系统架构，根据系统各个业务模块的设计流程，将相关操作、算法及各类函数封装为特定的功能模块，最终实现企业文档的自动智能化生成。系统功能主要包括加密存储与共享、权限控制、快速索引、协同编辑、用户管理、版本控制、模版库和文档预览等。

信息理解的本质是理解信息来源的过程，并将其按照企业文档信息库的标准与要求进行归纳和存储。信息理解功能的信息流方向为：形态学 – 句法 – 语义 – 习惯表达 – 形式语言 – 文本。所有的分析都会将这条自动文字翻译作为前提，语法和语义分析作为核心，规则和背景作为基础。基于文本单元的系统理解模型可分为分词层、句子层、段落层和文本层 4 个层次。信息理解功能按照既定的标准以及人工操作对信息源进行分析理解。文字处理是分析理解的基础，句子处理是关键，段落处理是整合手段，文本处理是终极目标。

信息分析功能可以查找知识数据库和关键字字典对信息源进行分类处理，并使用语法分析输入文本中的句子之间的关系，这些句子按语法和表达习惯进行分类，并形成一种与数据树类似的数据结构。语义分析能够识别在上一阶段形成的数据结构，并理解其蕴含的内在含义，形成句法结构和信息之间的映射和转换，进而在更高层次上理解文本含义。由于在分析中包含人工干预的成分，在对企业历史文档进行分析学习后，能够形成本企业特定的文档规范化结构和语句表达，弥补机器分析

的不足，有效提高信息分析和理解的精确度。输出是在理解和分析的基础上的一种知识表示，本文所描述的规则库主要包括企业文档格式和文本规范规则库、语法规则库、语义规则库、习语规则库、推理学习规则库。上述规则库为存储在机器中的文本模版提供了特定规则。知识库管理子系统起到引导用户有效操作知识库的作用，包括查看、修改、删除、添加词汇和语用知识。

文档生成功能包括文本生成、主题词池汇集、文档标签自动生成和反馈评估等。系统可以依据用户提供的类型和关键词，并结合企业和用户的背景、行业、特点等，生成相应的模版和内容。在这一过程中，系统能够生成具有普适性的业务词汇表，用户通过导入词汇表、建立模板后输入模板名直接即可使用。系统可以根据前期预设的算法和规则，根据用户提供的信息自动生成摘要。该系统通过对企业前期积累的流程文档知识和工作场景文档知识进行分析处理，对企业关键业务信息进行筛选，形成文档摘要。同时，系统还能够对文档模版实现可视化跟踪管理，用户可以通过预览功能随时查看文档进度，并对文档模版数据库进行人工干预。同时系统也为人工输入和选择信息数据创造了机会，能够在较短时间内生成符合企业业务状况并与用户需求信息相对应的企业文档模板内容。用户提供必要信息并完成人工干预后，系统即将会从企业数据库文档中随机提取出来最可能符合提交条件中的前后三个企业文档元素显示信息给用户。

用户管理和权限管理则由相应的管理员为工作人员分配账号与权限，为协同文档编辑奠定基础。同时，为了真正能同时满足企业文档编写者对于个性化文件的在线编辑和使用的要求以及保存企业业务记录文件的特定工作需要，协同在线文档编辑平台对个性化文档编辑用户提供出了多文件并发编辑与管理的功能。包括但也不限于基本的文档协作编写的管理和功能、文档协作的创建、保存、重命名、打开文档的近期和编辑、查看文档信息、高级设置、关键字搜索和检索功能等。不同级别权限下的每个文档的协作的作者均可以根据该作者的自己所负责的文档与写作的权限去独立完成其自己所相应的类别中的文档写作的内容。

四、结语

随着企业数智化转型的推进以及无纸化办公的盛行，大量的生产运营数据和文档被积累下来，对这些历史数据和文档进行开发和利用，能够就能为今后工作的开展提供极大便利。企业文档的自动生成是文本生成技术应用的新领域，其形成过程中的多维、动态、多方参与、文档类型广泛等特点对系统的架构与生成算法模型的学习、训练与修正提供了一定的挑战。本文结合企业文档自动生成全流程的多方位、多角度需求与技术领域的拓展，较为合理构建了企业自动文档生成系统。通过企业文档的自动生成减少企业文档管理过程中的人力工作，提高企业办公效率，能够使工作人员将主要精力放在更需要创造力的工作岗位上，推进企业智慧化办公、智慧化管理。此次设计的系统经过一定的实践检验，但受限于客观条件，在不同的场景中可能还存在一些不足之处，因此未来研究将着重着眼于拓宽系统应用场景，在逻辑架构与技术结构上展开持续的研究和优化。

参考文献：

[1] Dalpiaz F, Ferrari A, Franch X, et al. Natural Language Processing for Requirements Engineering: The Best Is Yet to Come[J].IEEE Software,2018,35(5):115-119.

[2] 邵欣欣，张明会，高梓峻. 文档生成技术研究与应用 [J]. 软件工程 ,2018(1):15-17.

[3] 史姣丽，黄传河，何凯，等. 支持多用户协同编辑的云存储访问控制方法 [J]. 计算机研究与发展 ,2017,54(7):1603-1616.

[4] 徐东风，彭红星，廖俊杰. 基于 Java 的文档格式检查技术的研究及其应用 [J]. 计算机工程与设计 ,2010,31(19):4309-4311.

[5] Mariani M, Baggio R, Fuchs M,et al.Business intelligence and big data in hospitality and tourism:a systematic literature review[J].International Journal of Contemporary Hospitality Management,2018,30(12):3514-3554.

[6] Greff K, Srivastava R K, Koutnik J, et al. LSTM: A Search Space Odyssey[J].IEEE Transactions on Neural Networks & Learning Systems,2017,28(10):2222-2232.

[7] 柳林青，余瀚，费宁，等. 一种基于 TextRank 的单文本关键字提取算法 [J]. 计算机应用研究 ,2018,35(3):705-710.