文章编号：1005-8451（2018）10-0040-05

自然语言处理关键技术在智能铁路中的应用研究

薛　蕊，马小宁，李　平，杨连报
（中国铁道科学研究院集团有限公司　电子计算技术研究所，北京　100081）

摘　要：介绍自然语言处理发展历程和关键技术，结合智能运营、智能装备和智能建造3大领域，分析并总结自然语言处理相关技术在智能客服、安全管控、资产档案、智能维修、决策辅助和督查校验等方面的应用。通过对这些前沿应用的发展综述和探索发掘，论证自然语言处理相关技术方法可以成为铁路行业完成智能铁路转变的助力，并且随着自然语言处理领域自身的不断发展和突破，为铁路的智能化进程带来更显著的变革。

关键词：自然语言处理；智能铁路；自然语言理解；关键技术
中图分类号：U2：TP39　**文献标识码：**A

Nature language processing techniques and its applications in intelligent railway

XUE Rui, MA Xiaoning, LI Ping, YANG Lianbao
(Institute of Computing Technologies, China Academy of Railway Sciences Corporation Limited, Beijing　100081, China)

Abstract: This paper introduced the development process and key technologies of natural language processing. Combined with the three fields of intelligent operation, intelligent equipment and intelligent manufacturing, the application of natural language processing's related technologies in intelligent customer service, safety control, asset archives, intelligent maintenance, decision support and inspection and calibration was analyzed and summarized. By summarizing the development and exploration of these frontier applications, it was demonstrated that natural language processing's relevant technologies and methods could help the railway industry to implement the transformation of intelligent railway. With the continuous development and breakthroughs in the natural language processing field, it could bring more significant changes to the railway intellectualization process.

Keywords: nature language processing; intelligent railway; nature language understanding; key technology

自然语言处理涉及到人机交互的计算语言学和人工智能领域，它使得计算机和人类之间可以进行无缝交互，并且在机器学习的帮助下，使得计算机获得理解人类语言的能力。自然语言处理是一门融合语言学、计算机科学、数学于一体的科学[1]。已在多领域得到广泛应用，并通过智能信息服务产生应用价值[2-4]。在铁路行业内，虽然非结构化数据量十分庞大，但是自然语言处理的应用才刚刚起步，如Rosadini等人提出利用自然语言处理技术分析铁路信号制造商需求文档，从中探测铁路需求的缺陷[5]。未来通过自然语言处理相关技术可以对海量的文档进行有效管理，如存储和检索；对文档深入挖掘和分析，发现事件之间的内在联系和规律；与既有的技术手段相结合，促进和推动智能铁路的发展。以往的综述性研究多为总结某项技术在自然语言处理领域的发展和应用[6-8]。本文在概述自然语言处理发展历程和关键技术的基础上，将自然语言处理技术引入智能铁路，探索和分析自然语言处理在智能运营、智能装备和智能制造等方面的典型应用，展望自然语言处理在铁路行业的应用前景。

1　自然语言处理发展历程

作为计算机科学领域与人工智能领域中的一个重要方向，自然语言处理最早于上世纪50年代正式

收稿日期：2018-02-14
基金项目：中国铁道科学研究院重大课题（2017YJ005）。
作者简介：薛　蕊，研究实习员；马小宁，副研究员。

提出。最早的自然语言理解方面的研究工作是机器翻译，20 世纪 60 年代西方研究者对机器翻译做出了大量探索性的研究工作。然而，由于低估了自然语言的复杂性，和当时自然语言处理理论和技术的缺乏，自然语言处理领域的研究进展缓慢。直到 20 世纪 70 ～ 80 年代，机器学习相关算法的引入，为自然语言处理带来了革新。从此自然语言处理从基于规则的时代进入了广泛应用统计模型的时代，在这一阶段，很多自然语言处理任务得到了长足的发展。近年来，深度学习技术在各个方面取得瞩目的成果，通过应用深度学习相关技术方法，自然语言处理的多项任务取得了突破，比如语言建模，语义解析等。

2　自然语言处理关键技术

2.1　文本分类

文本分类是将文本划分至预设好类别中的过程。如果 D_i 是文档集合 D 中的一个文档，$\{C_1, C_2, C_3, \cdots, C_n\}$ 是类别集合，那么文档分类就是将其中一个类别 C_j 分配给文档 D_i 的过程。根据其特征，文档可以被标记为一个类别或者多个类别。如果文档仅属于一个类别，被称为“单个标签”，反之如果文档属于多个类别，则被称为“多个标签”。如果文档仅属于两个类别中仅有的一个，“单个标签”的文本分类问题可以进一步被理解为“二分类”问题 [9]。如图 1 所示，文档分类流程通常包括文档表征、特征选择或者特征变换、构建算法模型、训练算法模型、以及最终对算法模型的评价。

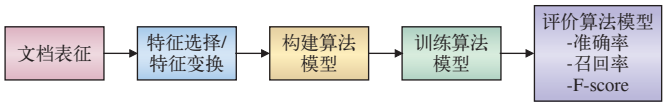

图1　文档分类过程图

2.2　命名实体识别

命名实体识别是对文本中的重要名词和指代词定位和分类的过程。例如，定位和识别新闻中的人名、地名和组织机构名称等重要的信息，用于进一步的语言处理和应用。命名实体识别作为自然语言处理中的重要任务，可被用于信息抽取、问答系统和机器翻译等领域中。例如，命名实体信息可以将专有名词定位为一个整体，从而辅助机器翻译系统进行整词翻译，以避免逐词翻译可能导致的翻译错误。

大部分命名实体识别系统包括人名、地名、组织机构名和定义更为宽泛的混合实体。这些类别主要用于与新闻相关的语料，在其他相关领域，命名实体模型需要用其相关语料和标注类别重新进行训练和测试 [10]。

2.3　自动摘要

自动文摘是对输入文本进行压缩和精炼，最终输出源文本中重要概念的过程 [11]。根据输入文档类型的差异（单个文档 / 多个文档）、目的的差异（泛化的 / 特定领域的 / 基于查询的）、输出文档类型的差异（抽取性的 / 概括性的），自动文摘系统可以被划分不同的类别 [12]。单个文档摘要是指对单个文档进行总结概括，同理多个文档摘要的数据源是多个文档，但是多个文档涉及的基本是同一个主题。泛化的自动文摘系统是指对所有的文本进行概括总结而不考虑其主题或者类别。特定领域的文摘系统则有着很强的专业或者领域的指向性，比如金融文章的摘要，生物制药文档的摘要等等。通常，该类型的摘要需要特定的专业知识以辅助句子的筛选过程。基于查询的摘要仅仅包含用户需要提取的信息，这些查询通常是自然语言问题或者是特定主题的关键词。抽取性文摘和概括性文摘的生成方式有所差异，抽取性文摘从文档中定位和抽取重要句子从而生成文摘，而概括性文摘是通过合并选定的文档，再将不重要的部分进行压缩生成最终的文摘。

2.4　知识图谱

知识图谱以实体和实体关系的形式对信息进行建模从而得到知识表征和它们的关联关系 [13]。知识图谱并非是一个全新的概念，而是基于在 2006 年提出的语义网概念，语义网强调使用本体模型来形式化表达数据中的隐含语义，由此产生了 RDF（resource description framework）模式 (RDF schema) 和万维网本体语言（OWL，Web ontology language）的形式化模型。基于以上研究，Google 于 2012 年 5 月 17 日正式提出了知识图谱 [14]。

三元组是知识图谱一种通用的表示方式，之前流传较广的是 RDF 的一种（主语、指向、宾语）三元组（SPO），其中，主语（subject）和宾语（object）均为实体，指向（predict）阐明了实体之间的关

系。该三元组可以用有向的图结构表示，如图 2 所示。知识图谱的三元组可表示为 $G=(E,R,S)$，其中，$E=\{e_1, e_2, \cdots, e_{|E|}\}$ 是知识库中的实体集合，共包含 $|E|$ 种不同实体；$R=\{r_1, r_2, \cdots, r_{|E|}\}$ 是知识库中的关系集合，共包含 $|R|$ 种不同关系；$S\subseteq E\bullet R\bullet E$ 代表知识库中的三元组集合。

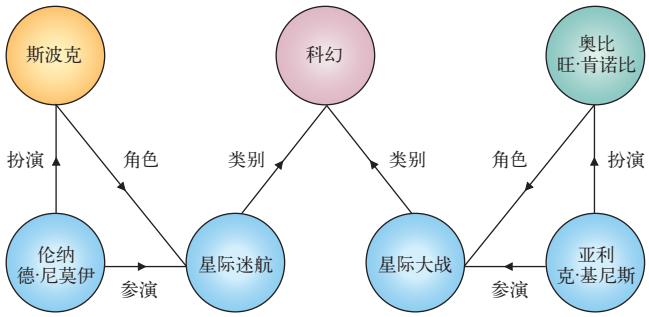

图2 （主语、指向、宾语）三元组示例

2.5 智能问答

智能问答旨在针对用户问题传递包含相应答案的精确信息。问答范式产生于 60 年代末，并在 70 年代初纳入自然语言理解的框架。根据问题的类型问答系统被分为两类，开放域问答系统与固定域问答系统。开放域系统主要基于网络，对专业领域没有限制，固定域系统对专业进行了限制，比如医药或天气预报等 [15]。

问答系统构建有诸多方案，如基于语言学的方法，基于统计模型的方法和基于模式匹配的方法。为了问答系统性能更优，往往采用混合的方法进行构建 [16]。近年来很多公司研发了语音助手，如苹果手机的 Siri，这类应用本质上是任务导向的智能问答系统，在之前的智能问答上集成了语音识别等技术，其流程如图 3 所示。

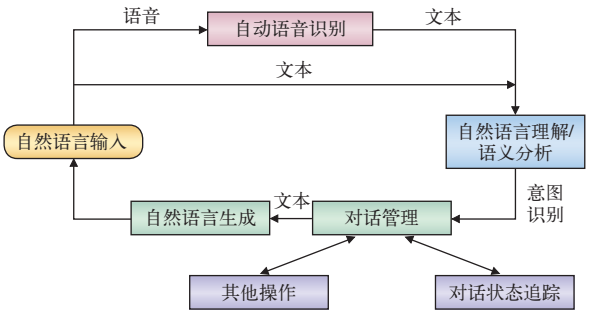

图3 任务导向的智能问答样例

3 自然语言处理在智能铁路中的应用场景

3.1 在智能运营中的应用

根据《中长期铁路网规划》，到 2020 年，全国铁路网规模达到 15 万 km，其中高速铁路 3 万 km，覆盖 80% 以上的大城市。随着全国铁路网规模不断扩大，铁路运营中的节能高效、安全管控等问题越发得到了人们的关注。自然语言处理的命名实体识别、知识图谱、智能问答等关键技术应用在铁路运营中，可以有效节约运营成本、改善乘客服务以及提高运营中的安全管控。

3.1.1 智能客服

智能客服是自然语言处理的一个重要的应用场景，其主要功能是与用户进行基本沟通，并自动回复用户有关产品或服务的问题，以达到降低企业客服运营成本、提升用户体验的目的。智能客服在电子商务、金融领域等已经得到了广泛的应用。在铁路运营中，智能客服可以在票务、车站等场景中给乘客提供优质高效的服务和良好的乘车体验。

3.1.2 安全管控

铁路运营中的安全涉及风险、隐患、事故故障等多个方面，对风险、隐患和事故故障的描述多以文本的方式存在，如风险库、隐患库、和事故故障报告等。通过文本分类和命名实体识别等技术手段，可以将非结构化的文本数据转化为结构化字段，便于存储、检索和统计分析。通过对事故故障进行关联分析和原因分析，可以挖掘事故故障之间的内在联系和事故故障的发生规律。结合风险和隐患方面的数据，解析风险、隐患和事故故障之间的相关关系和转化路径，有助于将事故故障扼杀在萌芽状态，提高铁路的行车安全。此外，结合相关结构化数据如设备数据、传感器数据等，可以对一些安全问题进行预测，促进设备检查维修从基于条件的维修向基于预测的维修转变，真正做到对安全问题的超前防范。

3.2 在智能装备中的应用

铁路行业拥有庞大且多元化的资产，如机车车辆、基础设施等，因此如何对这些资产进行有效的管理和优化的配置，一直是铁路行业关注的重点问

题。资产管理指的是通过一系列措施和方法降低资产的全生命周期的成本，同时获取资产使用的效益最大化。资产管理不仅仅局限于维修方面，而是从设计、制造、运维到淘汰更新的一个全生命周期管理。通过运用自然语言处理相关技术，可以有效提高资产管理的效率，推动资产管理向资产智能的转变。

3.2.1　资产档案

在资产管理中运用自然语言处理相关技术可以自动化地建立和管理资产档案，及时跟踪资产的状态变更，有助于简化资产管理的流程。同时通过对资产档案进行关联分析，可以将相关资源进行整合，合理高效地配置现有资源避免浪费。

3.2.2　智能维修

将自然语言处理应用于铁路的资产管理能够整合行业内有价值的信息、专家知识、安全条例、维修规定等相关规章制度，自动优化维修作业所需的车辆调度、工具设施、人力资源等。例如，香港铁路公司（MTR）利用人工智能进行工程师每周的工作派遣和调度。这一方式使得他们在维修制度内得以最大化使用资源，MTR 也因此在维修效率上提高了至少 50%，同时节省了时间和成本 [17]。

除此之外，结合基础设施等设备档案和设备监控数据，可以对设备维修、维修时间进行建模预测，有助于从按时维修和状态维修向预测维修进行转变，减少成本的同时提高效率 [18]。

3.3　在智能建造中的应用

随着建筑制造领域信息化的不断完善，建筑制造行业已经过渡到了数字化阶段，具有代表性的就是 BIM 系统的应用。作为强大的集成化系统，基于 BIM 的系统能够在工程设计、工程施工以及工程监察过程中高效地传递信息、进行资源的优化配置、以及通过一些监察手段提前发现施工问题以避免返工。在信息化和数字化之后，下一步则是智能化，智能化能够减少人力成本、进行资源配置和决策的最优化。建筑制造领域智能化的最大特点是，人工智能技术方法在行业中的广泛应用。人工智能技术的应用离不开数据的支持，而集成了大量数据和信息的 BIM 系统可以发挥重要的作用。将 BIM 系统与自然语言处理相结合，可以切实有效地处理建筑施工中的实际问题，促进建筑制造由数字化向智能化的转变。

3.3.1　决策辅助

建筑施工过程中需要进行大量的决策，例如，选择施工方法、承包方、施工材料等。自然语言处理中的知识图谱、智能问答以及推理等技术可以为相关业务人员提供决策依据，辅助业务人员在复杂场景下进行相关决策 [19]。

3.3.2　督查校验

建筑施工相关的标准和规程往往以非结构化文本的形式存在，例如施工质量验收规范。通过集成应用 BIM 技术和自然语言处理相关技术，可以按规范要求对 BIM 模型构件的尺寸及位置等进行自动检查，从而减轻有关人员的工作量。

4　结束语

本文在介绍自然语言处理发展历程、关键技术的基础上，结合智能铁路的发展，创新性地将自然语言处理技术全面引入铁路行业，阐述了自然语言处理在智能铁路中诸多可能的应用场景。在智能运营、智能装备和智能建造 3 大领域中，自然语言处理相关技术方法均可结合业务需要，在实际的场景中推动和促进铁路行业向智能化转变。

参考文献：

[1] Joseph, S.R., Hlomani, H., Letsholo, K., et al., Natural Language Processing: A Review[J]. International Journal of Research in Engineering and Applied Sciences, 2016, 6(3): 207-210.

[2] 徐　静，杨小平．基于 CRF 模型的网络新闻主题线索发掘研究 [J]. 中文信息学报，2017, 31（3）：94-100.

[3] 王　超，李　楠，李欣丽，等．倾向性分析用于金融市场波动率的研究 [J]. 中文信息学报，2009, 23（1）：95-98.

[4] 李丽双，郭元凯．基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别 [J]. 中文信息学报，2018，32（1）：116-122.

[5] Ferrari A, Gori G, Rosadini B, et al. Detecting requirements defects with NLP patterns: an industrial experience in the railway domain[J]. Empirical Software Engineering, 2018(1):1-50.

[6] 林奕欧，雷　航，李晓瑜，等．自然语言处理中的深度学习：方法及应用 [J]. 电子科技大学学报，2017, 46（6）：913-919.

（下转 P48）

[5] 李田科，于仕财，余春卫 . 导弹发射车综合诊断与健康管理系统 [J]. 兵工自动化，2012，31（4）：11-14.
[6] 齐渡谦，付毅飞 . 航天科工 PHM 系统正式搭载 C919[N]. 科技日报，2016-08-05（001）.
[7] 姚晓山，张卫东，周　平，等 . 基于油液监测的船舶柴油机故障预测与健康管理技术研究 [J]. 武汉理工大学学报（交通科学与工程版），2014，38（4）：874-877.
[8] 马　剑，吕　琛，陶来发，等 . 船舶主推进系统故障预测与健康管理设计 [J]. 南京航空航天大学学报，2011，43（7）：119-124.
[9] 何正友，程宏波 . 高速铁路牵引供电系统健康管理及故障预警体系研究 [J]. 电网技术，2012，36（10）：259-264.
[10] 蒋觉义，李　璠，曾照洋 . 故障预测与健康管理标准体系研究 [J]. 测控技术，2013，32（11）：1-5.
[11] 景　博，汤　巍，黄以锋，等 . 故障预测与健康管理系统相关标准综述 [J]. 电子测量与仪器学报，2014，28（12）：1301-1307.
[12] 崔　涛 . 基于 Web 的 TADS 实时监控功能设计 [J]. 铁路计算机应用，2011，20（7）：23-26.
[13] 毕汉岗 . 第二代红外线热轴判别准则算法的探讨 [J]. 铁路计算机应用，1995，4（4）：26-28.

责任编辑　陈　蓉

（上接 P43）

[7] 庞　亮，兰艳艳，徐　君，等 . 深度文本匹配综述 [J]. 计算机学报，2017，40（4）：985-1003.
[8] 李　芳，刘胜宇，刘　峥，等 . 生物医学语义关系抽取方法综述 [J]. 图书馆论坛，2017（6）：61-69.
[9] Jindal R, Malhotra R, Jain A. Techniques for text classification: Literature review and current trends[J]. 2015, 12(2): 1-28.
[10] Zitouni I. Natural Language Processing of Semitic Languages [M]. Springer Berlin Heidelberg, 2014: 221-245.
[11] Shetty A, Bajaj R. Auto Text Summarization with Categorization and Sentiment Analysis[J]. International Journal of Computer Applications, 2015, 130(7): 4053-4060.
[12] Yogan, Jaya Kumar, et al. A review on automatic text summarization approaches[J]. Journal of Computer Science, 2016, 12(4): 178-190.
[13] Nickel M, Murphy K, Tresp V, et al. A Review of Relational Machine Learning for Knowledge Graphs[J]. Proceedings of the IEEE, 2015, 104(1):11-33.
[14] 徐增林，盛泳潘，贺丽荣，等 . 知识图谱技术综述 [J]. 电子科技大学学报，2016, 45（4）：589-606.
[15] Andrenucci A, Sneiders E. Automated Question Answering: Review of the Main Approaches[C]//International Conference on Information Technology and Applications. IEEE Computer Society, 2005:514-519.
[16] Ajitkumar M, Khillare S.A., C Namrata. Question Answering System, Approaches and Techniques: A Review[J]. International Journal of Computer Applications. 2016, 141:34-39.
[17] Chun A H W, Suen T Y T. Engineering works scheduling for Hong Kong's rail network[C]//Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI Press, 2014: 2890-2897.
[18] Faiz RB, Edirisinghe EA. Decision making for predictive maintenance in asset information management[J]. Interdisciplinary Journal of Information, Knowledge, and Management. 2009, 4(1): 23-36.
[19] 马智亮，蔡诗瑶 . 基于BIM的建筑施工智能化[J]. 施工技术，2018, 47（6）：70-83.

责任编辑　陈　蓉