基于文本挖掘的ISO标准术语自动识别与标准术语知识图谱构建研究

方思怡

（上海市质量和标准化研究院）

摘　要：ISO标准术语蕴含特定的领域知识，是ISO标准文本数据的重要组成。在标准数字化转型下，ISO术语自动识别技术面临迫切的发展需求。本研究通过深入分析ISO标准术语的编写要求，总结了ISO标准术语核心要素的文本特性，基于此采用基于规则的文本挖掘方法构建了ISO标准术语自动识别模型及结构化和可视化加工路径，在ISO 26262标准上完成验证与应用，生成ISO 26262的标准术语知识图谱。本研究的技术路径能够为ISO标准实体抽取和相关标准数字化平台的构建提供一定的参考。

关键词：ISO，国际标准，术语自动识别，标准数字化，文本挖掘

DOI编码：10.3969/j.issn.1674-5698.2024.08.012

Research on Automatic Recognition of ISO Standard Terminology and Construction of Standard Terminology Knowledge Graph Based on Text Mining

FANG Si-yi

（Shanghai Institute of Quality and Standardization）

Abstract: ISO standard terminology contains specific domain knowledge and is an important component of ISO standard text data. In the context of the digital transformation of standards, ISO terminology automatic recognition technology is facing urgent development needs. This study conducted an in-depth analysis of the requirements for writing ISO standard terminology and summarized the text characteristics of the core elements of ISO standard terminology. Based on this, a rule-based text mining method was used to construct an automatic recognition model for ISO standard terminology and a structured and visualization processing path. The model was validated and applied on the ISO 26262 series of standards. The study can provide some reference for the extraction of ISO standard entities and the construction of related standard digital platforms.

Keywords: ISO, international standard, terminology automatic recognition, standard digitization, text mining

0　引言

术语（Term）是蕴含特定学科领域核心概念的专用名词，与特定学科的领域知识密切相关[1,2]。术语识别（Terminology recognition，TR）是指从语料中抽取具有领域代表性的词汇或短语的过程，被视为信息抽取和命名实体识别（Naming entity recognition，NER）领域的重要分支[3]。近年来术语

基金项目：本文受上海市质量和标准化研究院院立项目“国际标准核心要素标注方法研究”（项目编号：YRY202406）资助。

作者简介：方思怡，硕士研究生，工程师，研究方向为标准数字化、标准数据挖掘、标准知识图谱。

自动识别（Terminology automatic recognition,TAR）逐渐引起各界研究者的关注。标准是领域技术情报的重要来源，标准术语也是领域技术信息的核心载体，具有较强的专业性与系统性。在标准文本中，ISO国际标准是推进国际贸易与合作的重要准绳，其地位和影响力不言而喻。在标准数字化转型下，ISO术语自动识别将为标准语料库、标准知识图谱、标准智能检索、标准自动标引、标准智能翻译、标准本体、相关产业画像和知识体系构建等标准知识服务奠定重要的数据基础[4, 5]。

1 标准术语自动识别的研究现状

1.1 术语自动识别的相关研究进展

纵观已有的研究，术语识别技术历经多个发展阶段，迄今为止已形成了基于专家人工、基于规则与统计、基于传统机器学习以及基于深度学习的识别方法。

受限于技术水平，早期的术语识别研究多通过专家人工模式进行，该方法能确保术语抽取的质量，但人力和时间成本较高，可推广性不强。随着计算机技术的发展，术语自动识别逐渐取代了专家人工识别，成为各领域术语识别的主流发展方向。术语自动识别的具体技术取决于其所针对的文本语料特性。本研究系统梳理了术语自动识别研究的技术方法，表1概括了不同术语自动识别方法的原理、特点及案例。

1.2 标准术语自动识别的现状与发展趋势

当前国内外标准数字化转型正处于起步阶段。2021年发布的《国家标准化发展纲要》明确指出要加快标准的数字化、网络化和智能化转型，由此对标准数字化文本的知识自动抽取与加工技术提出了全新的要求[6]。与专利、科技论文等领域相比，标准术语自动识别研究尚存在大量的提升空间。作为标准的基本要素之一，标准术语是标准文本技术信息的重要组成，也是标准知识自动抽取的对象之一。近来涉及标准实体识别的国内外研究大多针对标准起草单位、标准提出单位、标准指标和标准规范性引用文件[7-9]，尚未对国际和国内外标准进行术语自动识别的深入探索。

尽管标准术语自动识别尚存在大量研究空白，在标准数字化转型的驱动下，随着标准知识服务对细粒度和深层次的需求日益增加，国际和国内外标准的术语自动识别方法将成为大势所趋。作为国际标准的重要品种，ISO标准术语自动识别技术也存在迫切的发展和应用需求。

2 ISO标准术语自动识别的研究方法

2.1 研究思路

本研究以上海市质量和标准化研究院“标准文献发行系统”中现有的ISO文本为数据来源，结合ISO文本编写的相关要求[10]和对ISO文本结构特性的深入分析，形成相应的研究思路。经过分析可知，当前ISO国际标准的载体为PDF格式的数字化文本，通常以英语语种为主，可能存在多语种的情

表1 不同术语自动识别方法的特点

术语自动识别方法	原理	特点	举例
基于规则与统计的方法	总结术语编写的词性规则或文本结构规则，根据规则在文本语料中匹配获得术语数据集	术语识别精度高，灵活性较低	正则表达式、C-value、NC-Value
基于传统机器学习的方法	通常将术语识别转化为二分类问题或序列标注问题，采用特定的模型学习语料的浅层语言学或深层文本特征	对特征工程的要求较高，可移植性不高，灵活性强于基于规则和统计的方法	决策树、最大熵、条件随机场（CRF）、支持向量机（SVM）、隐马尔科夫（HMM）
基于深度学习的方法	在序列标注的视角下采用特定的神经网络模型学习语料的文本特征，多采用混合神经网络模型	无需繁琐的特征工程或规则设定，采用端到端的形式，可移植性较强	以BiLSTM-CRF为基准衍生而出的混合模型，包括CNN-BiLSTM-CRF、Bert-BiLSTM-CRF等

况，ISO术语条目的编写也遵循较为明确的规则。综上所述，本研究选取基于规则的文本挖掘技术作为ISO标准术语的自动识别方法，由此制定相应的技术路径。

2.2 研究流程

本研究基于研究思路，制定了如图1所示的研究流程框架。

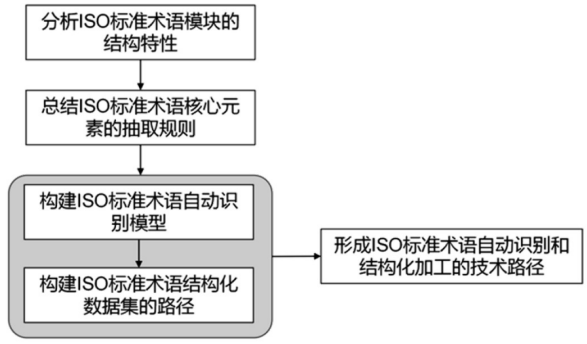

图1 ISO标准术语自动识别及结构化加工的研究流程

2.2.1 分析ISO标准术语模块的结构特性

研究流程的第一步为分析ISO标准术语模块的结构特性，以ISO的编写指南为依据，结合现行ISO文本术语数据的实际情况，明确ISO术语自动识别的研究范畴，并概括研究范畴内ISO术语模块核心元素的文本结构特性。

经过系统分析可知，ISO标准术语模块可以囊括的元素有术语介绍（Introductory wording）、术语编号（Term number）、首选术语（Preferred term）、首选术语的缩略语、可接受的术语同义词（Accepted term）、弃用术语（Deprecated term）、术语领域（Term domain）、术语定义（Term definition）、术语示例（Term example）、术语条目注释（Note to entry）以及术语来源（Term source）。本研究坚持以应用到导向，重点关注ISO术语中与技术密切相关的信息，故将研究范畴界定为术语编号、首选术语、首选术语缩略语、可接受的术语同义词、弃用术语、术语定义、术语示例、术语条目注释和术语来源。结合现有的ISO文本数据，将上述核心要素的定义、结构特性和示例概括见表2。

2.2.2 总结ISO标准术语模块中核心元素的抽取规则

在此基础上开展研究流程的第二步，也即根据文本结构特性，从文本表述形式、在术语条目中的所在位置等几个方面总结ISO标准术语模块中核心元素的抽取规则。以术语元素中的术语编号为例，其抽取规则可以概括为两点，其一是通常位于术语模块的第一个位置，其二是由阿拉伯数字和间隔点构成，且间隔点位于两个阿拉伯数字之间。

2.2.3 构建ISO标准术语自动识别模型

研究流程的第三步为针对ISO标准术语模块的各个核心元素，采用Python构建基于规则的ISO标准术语自动识别模型，本研究首先将ISO标准术语模块中各核心元素的抽取规则转化为伪代码，进而通过Python编写程序，形成适用于ISO术语各核心元素的自动识别算法，再将上述算法整合成为统一的算法模块，完成ISO术语自动识别模型的构建。

2.2.4 构建ISO标准术语结构化数据集的路径

研究流程的第四步为采用Python构建ISO标准术语结构化数据集的实现路径，主要包括设定结构化数据集的数据表达框架和结构化数据集的转化算法。该步骤旨在将ISO标准术语自动识别模型中获得的ISO术语各核心要素的抽取结果自动转化为结构化数据集的形式，为后续的ISO标准数据深度挖掘和加工奠定一定的技术基础。

2.2.5 形成ISO标准术语自动识别和结构化加工的技术路径

在完成上述步骤后，采用Python将ISO标准术语自动识别模型和结构化数据集的技术路径相结合，完成上述两者的代码模块的顺利链接，形成ISO标准术语自动识别和结构化加工的完整技术路径。

2.3 模型设计

ISO标准术语自动识别模型是实现ISO标准术语自动抽取的关键所在。

本研究采用基于规则的文本挖掘方法，通过Python编写了提取ISO标准术语条目模块与核心元素的抽取算法以及结构化和可视化加工的算法，形成了基于规则的ISO标准术语自动识别模型，模型框架详如图2所示。

由图2可知，该模型主要由ISO标准术语条目

表2 ISO标准术语核心要素的概况

标准术语核心要素		定义	结构特性	示例
术语编号		术语在所属ISO标准文本中的唯一标识符（Unique identifier）	通常用粗体书写，根据结构顺序采用数字编号而非字母编号，位于首选术语之前，具有跨语种的可识别性	3.1，3.1.1
首选术语	首选术语名称	是用来表示特定概念的主要术语的名称，其使用将贯穿于该ISO文本	通常用粗体小写表示，矢量沉积除外，位于术语编号之后，当存在多个首要术语名称时需罗列呈现	measuring distance
	首选术语的缩略语	是主要术语名称的缩写形式，其使用也可贯穿于该ISO文本	通常用粗体大写表示，位于所对应的首选术语名称之后	MD
术语同义词		其他可接受的术语表述方式	通常用小写非粗体表示，位于首选术语之后	measurement distant
弃用术语		已不再适合用来表述该特定概念的术语	通常用小写非粗体表示，表述形式为“DEPRECATED:#”，其中#是弃用术语的具体字符，如有术语同义词则位于术语同义词之后，如无则位于首选术语之后	DEPRECATED: measure distance
术语领域		表示术语所属的具体学科领域	通常用小写非粗体表示，外设尖角括号，表述形式为“<#>”，其中#是术语领域的具体字符，如存在则位于术语名称相关元素之后并位于术语定义之前的同一行中	<magnetic flux density>
术语定义		用来描述术语概念的单个短语	通常用小写非粗体书写，不能以冠词（The或a）开头，不能以句号结尾，每个术语只能对应一条术语定义，允许出现公式或图等非语言的定义表述形式，如存在术语领域则位于术语领域之后的同一行中	shortest distance from the surface of appliance to the closest point of the sensor surface
术语示例		提供解释术语概念的信息	通常用非粗体书写，表述形式为“EXAMPLE #”，其中#为术语示例的具体字符，#的首字母需大写，用句号结尾，如有多个术语示例，则需逐一编号，单一示例无需编号，采用陈述事实的语气，不得包含“shall”“should”和“may”，位于术语定义之后	EXAMPLE Verification review types can be technical review (3.127), walk–through (3.182) or inspection (3.82).
术语条目注释		提供补充术语数据的额外信息，包括使用单位、首选术语的选取原因、术语条目注释的编号等	通常用非粗体书写，表述形式为“Note”+数字+“to entry: #”，其中#为术语条目注释的具体字符，#的首字母需大写，用句号结尾，术语条目注释必须编号，位于术语定义之后	Note 1 to entry: Individual requirements on verification reviews are given in specific clauses of individual parts of the ISO 26262 series of standards.
术语来源		如该术语条目来自另一标准文本，提供术语条目的具体来源信息，主要包括标准文本的编号、出版年份并简要说明条目是否修改情况	通常用非粗体书写，外设方括号，表述形式为“[SOURCE: #]”，其中#为术语来源的具体字符，位于术语条目的最后一部分	[SOURCE:IEC 62233:2005, 3.2.6,modified – The abbreviated term,admitted and deprecated terms,and domain have been added.]

模块的抽取算法、ISO标准术语核心元素的抽取算法、ISO标准术语结构化数据库的构建算法以及ISO标准术语知识图谱的构建算法组成，其中ISO标准术语条目模块的抽取算法、ISO标准术语核心元素的抽取算法旨在实现ISO标准术语核心元素的自动抽取，ISO标准术语结构化数据库的构建算法旨在完成自动抽取结果的结构化加工，形成可适用于标准数字化平台的ISO文本结构化数据集，而ISO标准术语知识图谱的构建算法的目的在于对ISO标准术语的自动识别结果进行可视化展现并形成可供深度挖掘的数据集，为标准智能决策奠定数据基础。

3 ISO标准术语自动识别的研究结果

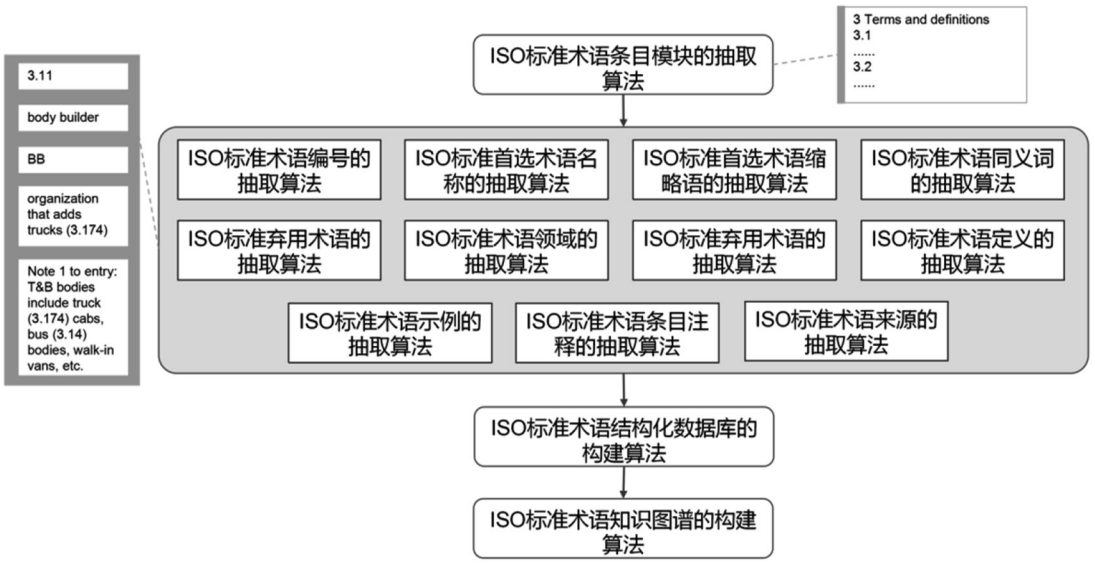

图2 基于规则的ISO标准术语自动识别模型框架

3.1 ISO标准术语结构化数据库的文献计量学分析

本研究在完成ISO标准术语自动识别及结构化和可视化加工的技术路径后，在特定的ISO标准文本上开展实证研究。本研究选取ISO 26262作为ISO术语自动识别技术的应用对象。ISO 26262针对汽车安全相关的电子电气系统，与汽车功能安全密切相关，其所涉及的汽车芯片也是近年来集成电路产业的热点方向之一。

经文本切词后统计可知，ISO 26262共计10篇标准文本，含有194.42万个字符与66.89万个单词。在上述文本中应用ISO标准术语自动化识别及结构化和可视化加工的模型，最终抽取获得的ISO标准术语条目的字符数为6.2万，单词数为2.1万；所涉及的ISO标准术语核心元素有术语编号、首要术语名称、首要术语缩略语、术语同义词、术语定义、术语条目注释和术语示例，其标准术语核心元素的数量分布情况如图3所示。

为了掌握ISO 26262标准术语数据的词频分布概貌，通过构建ISO标准术语条目干扰词库，剔除ISO术语条目常见的无关词，并采用Python描绘词云图，所得结果如图4所示。

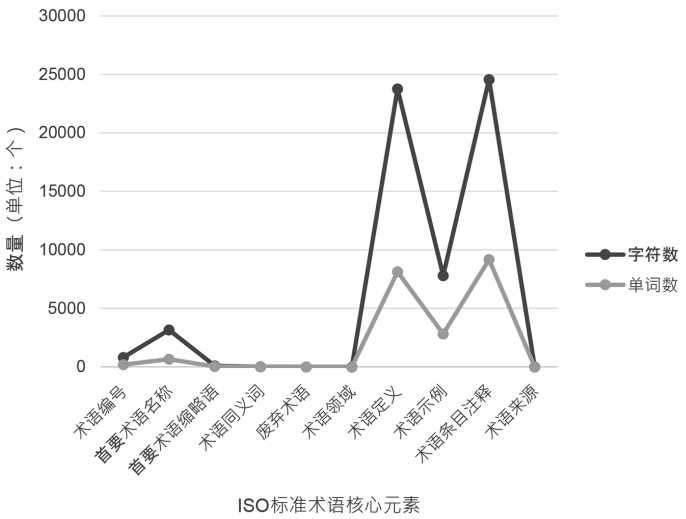

图3 ISO 26262系列标准的术语核心元素数量分布

图4 ISO 26262系列标准的术语条目词云图

3.2 ISO标准术语知识图谱

本研究基于ISO标准术语编写的文本特性，初步设计了ISO标准术语知识图谱的模式层，由此明确了标准术语的实体和关系类型。ISO标准术语知识图谱的模式层框架主要以ISO标准术语核心元素为实体类型，以核心元素的名称指向形式也即“ISO标准术语核心要素+是”的英文表述形式为关系类型。

采用Python编写了ISO标准术语知识图谱的可视化路径，在Neo4j平台中实现相关应用，图谱的示例和界面截图分别如图5与图6所示。该图谱共含有547个不同的标准实体和7种不同的标准关系类型。

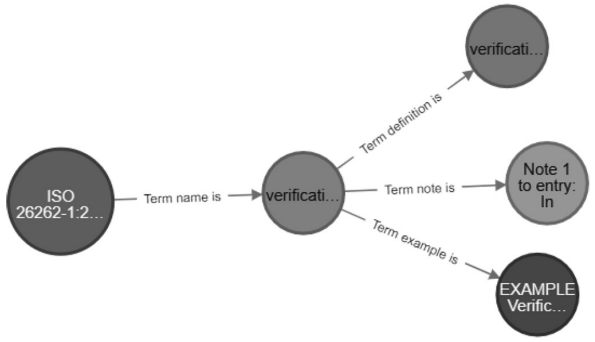

图5 ISO 26262系列术语标准知识图谱的示例

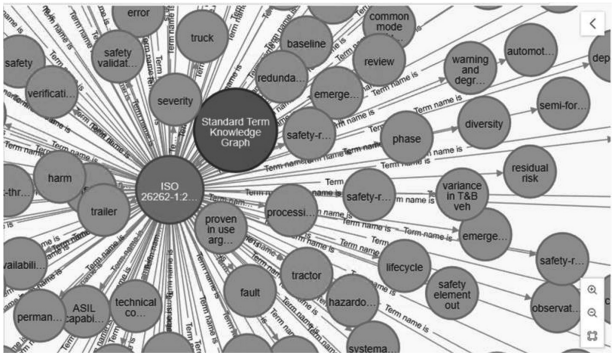

图6 ISO 26262系列的标准术语知识图谱界面截图

4 总结与展望

4.1 总结

本研究针对ISO标准的文本特性，构建了适用于ISO的术语自动识别模型及结构化和可视化加工路径，在ISO 26262标准上完成验证与应用，形成了ISO 26262的标准术语知识图谱。

4.2 展望

本研究为ISO标准的标准术语自动识别提供了一定的技术参考，在后续工作中将继续深化ISO标准实体抽取模型的研究，将其应用在标准数字化平台中，以期能够实现细粒度和深层次的ISO标准知识抽取与自动加工，推动标准数字化转型下标准知识服务的发展。

参考文献

[1] 陈翀,高欣妍,黄红. 基于BLSTM-CRF的自举式术语识别方法研究[J]. 情报工程, 2023,9(05):97-111.

[2] 冯鸽英. 基于深度学习的领域术语抽取方法研究[D]. 西安: 西安电子科技大学, 2022.DOI:10.27389/d.cnki.gxadu. 2022.000564.

[3] 阮光册,钟静涵,张祎,笛. 基于深度学习的术语识别研究综述[J/OL].数据分析与知识发现,2024,8(4):1-16[2024-03-05]. http://kns.cnki.net/kcms/detail/10.1478.G2.20230918.1824.002.html.

[4] 孙甜,陈海涛,吕学强,等.新能源专利文本术语抽取研究[J]. 小型微型计算机系统,2022,43(05):950-956.DOI:10.20009/j.cnki.21-1106/TP.2020-1022.

[5] 杨双龙,吕学强,李卓,等. 中文专利文献术语自动识别研究[J]. 中文信息学报, 2016,30(03):111-117+124.

[6] 马小雯,孙红军,刘彦林,等.标准知识数字化表达通用模型与自动抽取技术研究[J]. 标准科学, 2024(01):83-87.

[7] 赵伟,张览,望俊成. 金融领域标准文献知识图谱的构建与实现[J]. 情报工程, 2022,8(06):103-113.

[8] 程名. 基于BiLSTM+CRF的渔业标准术语识别研究[D]. 大连: 大连海洋大学, 2020.DOI:10.27821/d.cnki.gdllhy.2020.000002.

[9] Irlan Grangel-González.A Knowledge Graph Based Integration Approach for Industry 4.0[D]. 2019.

[10] ISO/IEC Directives, Part 2 Principles and rules for the structure and drafting of ISO and IEC documents (Eighth Edition)[S].