35.240.99
L70

CES XXX

团 体 标 准

T/CES XXX-XXXX

移动端智能交互训练语料基本要求与规范

Basic requirements and specifications of mobile terminal intelligent interactive training corpus

（征求意见稿）

XXXX-XX-XX 发布　　XXXX-XX-XX 实施

中国电工技术学会发布

目 次

前　言

本文件按照GB/T1.1—2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》给出的规则起草。

本文件由四川中电启明星信息技术有限公司提出。

本文件由中国电工技术学会标准工作委员会能源智慧化工作组归口。

本文件起草单位：国网信息通信产业集团有限公司、四川中电启明星信息技术有限公司、国网重庆市电力公司、国网重庆市电力公司电力科学研究院、重庆大学。

本文件主要起草人：李强、宋卫平、王红蕾、赵峰、周孔均、钟加勇、倪平波、田鹏、李欢欢、徐小云、刘礼、崔秋实、张强、李立、李军、高攀、高胜杰。

本文件为首次发布。

1 范围

本标准规定了移动端智能交互训练时自然语言处理样本(对话机器人交互文本意图识别样本)的基本要求、标注要求和标注流程。

本标准适用于各单位进行移动端智能交互训练模型开发时样本标注和样本入库的统一管理，包括样本的质量管控、样本标注的技术要求和流程管控。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 1.1-2009 标准化工作导则　第 1 部分:标准的结构和编写

GB/T 5271.28—2001 信息技术 词汇 第 28 部分;人工智能 基本概念与专家系统

ZYF 001-2018 语料库通用技术规范

T/CESA 1040—2019 信息技术 人工智能 面向机器学习的数据标注规程

Q/GDW 1560.1—2014 输电线路图像/视频监控装置技术规范 第 1 部分：图像监控装置

Q/GDW 1906—2013 输变电一次设备缺陷分类标准

3 术语和定义

下列术语和定义适用于本文件。

3.1

人工智能 artificial intelligence

一门交叉学科，通常视为计算机科学的分支，研究表现出与人类智能（如推理和学习）相关的各种功能的模型和系统。

3.2

自然语言处理 natural language process

是计算机科学领域与人工智能领域中的一个重要方向，能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。

3.3

样本数据 sample data: [Q/GDW 12118.1—2021,定义 3.5]

其具备的特征能够反映总体数据情况的一部分个体数据。

3.4

标注 corpus annotation: [Q/GDW 1906—2013,定义 3.5]

采用人工或计算机自动方式对样本的属性或特征进行描述，可用于实用的目的，如客户服务或资讯获取等。

3.5

标签 label: [T/CESA 1040—2019 定义 3.2]

标识数据的特征、类别和属性等内容，可用于建立数据及深度学习训练要求所定义的机器可读数据编码间的联系。

3.6

智能交互 intelligent interaction

智能交互一般指智能语音交互。 智能语音交互是基于语音输入的新一代交互模式，通过说话就可以得到反馈结果。

3.7

语料 corpus

即语言材料，是语言学研究的内容，也是构成语料库的基本单元。

3.8

语料库 corpora

语料库指经科学取样和加工的大规模电子文本库，其中存放的是在语言的实际使用中真实出现过的语言材料。

3.9

意图 intent

用户表达的句子希望达到某种目的打算。

3.10

槽位 slot

在用户表达意图的句子中，用来准确表达该意图的关键信息的标识。

3.11

标注工具 annotation tool: [T/CESA 1040—2019 定义 3.5]

标注人员执行标注任务生成标注结果的过程中使用的工具和软件。标注工具按照自动化程度分手动、半自动和自动三种。

3.12

半自动化标注 semi-automatic annotation

基于少量人工标注、机器预标注来训练标注模型，用于批量标注样本数据的半人工智能标注方法。

3.13

特色语种 special language

汉语普通话外的其他语种。

4 缩略语

下列缩略语适用于本文件。

BIOES：BIOES 标注模式 (B-begin, I-inside, O-outside, E-end, S-single) 属于序列标注模式之一，其中 B-begin 表示标注元素的开头，I-inside 表示标注元素的中间或结尾，O-outside 表示不属于待标注内容，E-end 表示标注元素的结尾，S-single 表示单个字符且本身就是一个标注元素。

BIO：BIO 标注模式 (B-begin, I-inside, O-outside) 属于序列标注模式之一，其中 B-begin 表示标注元素的开头，I-inside 表示标注元素的中间或结尾，O-outside 表示不属于待标注内

容。

JSON：JavaScript 对象表示法(JavaScript Object Notation)，是一种轻量级的文本数据交换格式

5 总则

本文件共分为样本基本要求、样本标注要求、样本标注流程三部分，具体内容组织框架见图 1:

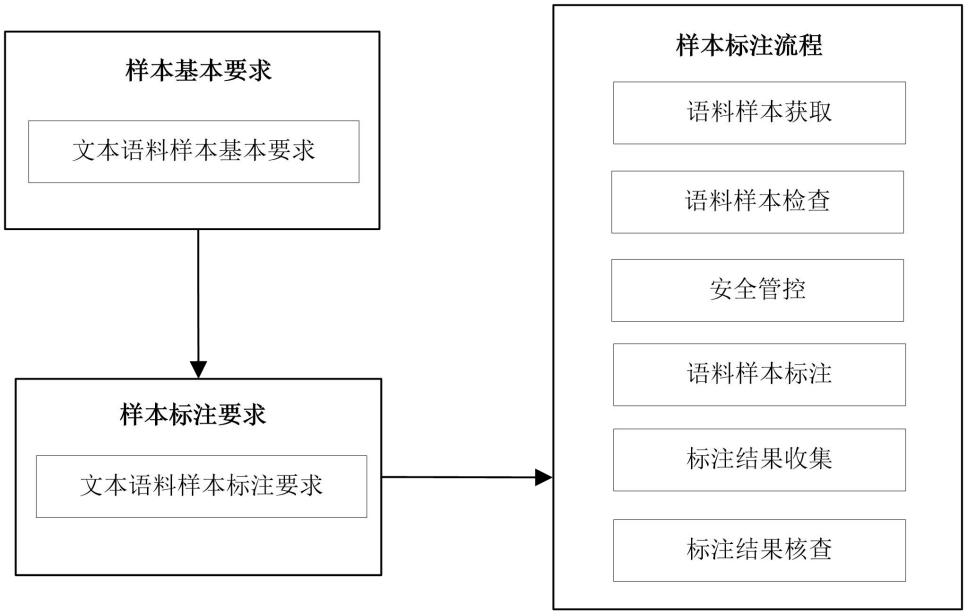

图 1 组织框架

6 文本类样本基本要求

6.1 文本文件存储格式要求

文本样本数据应采用 txt、csv、Json、xlsx、xls、xml 存储格式。

6.2 文本文件命名要求

文本文件名称应由四个部分组成：

a) 第一部分为项目命名词或文本来源；

b) 第二部分为当前文本文件的专业信息；

c) 第三部分为原始源文件生成时的日期，日期格式：YYYY-MM-DD；

d) 第四部分为文件唯一性编号，从 1 开始计数；

e) 这四部分用下划线连接，且文件名称长度和扩展名在内最大长度不超过 100 个字符(包含中英文字符和特殊字符)；

f) 文件命名举例：××项目_酒店预订语料_2022-06-08_1。

6.3 文本类样本质量要求

文本类样本质量应满足下述要求：

a) 应支持计算机正常读取，文本内容无乱码；

b) 内容应满足相关业务需求；

c) 应使用 UTF-8 编码。

6.4 文本样本描述文件

每批次文本样本集应有一个描述文件，应满足下述要求：

a) 存储格式应为 txt 格式；

b) 命名应由三个部分组成：

1) 项目命名词或样本来源；

2) 本文件创建的日期，日期格式：YYYY-MM-DD；

3) 文件唯一性编号，从 1 开始计数；

4) 文件名的各部分用下划线连接，文件命名示例：××项目_2022-06-08_1。

c) 文档内容应描述本样本集的基本信息，应包括样本所属项目、样本来源、创建日期、联系人、样本标注信息、标注格式、样本用途等信息。

7 文本类样本标注要求

7.1 基本要求

应满足标注对象范围、标注方式、标注文件命名要求。具体要求包括：

a) 文本语料样本标注应支持意图、槽位等信息的标注；

b) 文本语料样本标注应支持序列标注、指针标注等多种标注方式；

c) 标注可通过线上标注(样本+标注平台)和线下标注(线下小工具和线下文本 txt、csv、Json)实现；

d) 序列标注应采用 B、I、E、O、S 标签列表，宜采用 BIO、BIOES 标签方案进行标注；

f) 已完成标注的文本文件应按照规定的命名格式命名。

7.2 意图标注要求

样本意图标注应满足下述要求：

a)样本标注前应确定意图类别数和意图类别名称；

b)样本意图类型的确定需要结合具体的应用场景和待标注样本数据；

c)若一条样本可标注为多个意图类别时，应根据应用场景将该样本标注为一个可能性最大的意图类别，必要时可由多位标注人员共同确定待标注样本的意图类别；

d)每条语料样本都应标注出其意图类别，若一条语料意图不属于已定义的意图中的任何一类则可将该语料删除，或者新增一个意图类别以将语料样本中不属于已定义意图类别的语料样本均归类于该意图类别；

e)标注时应做到准确标注意图类别；

f)应用场景如：新增日程、查询日程

7.3 槽位标注要求

样本槽位标注应满足下述要求：

a) 样本标注前应定义槽位的类别数和类别名称；

b) 一条待标注样本中有可能存在多个槽位，应标出所有的槽位信息；

c) 一条样本中可能不存在槽位信息，应允许槽位信息为空；

d) 样本中的槽位信息可能存在重叠，对存在重叠的槽位信息是否标注以及怎样标注需要根据具体情况确定；

e) 标注槽位信息要准确、全面；

f) 应用场景如：“定一个早上九点在北京评审的日程”，需要标注的槽位信息为：“早上九点”、“北京”、“评审”，标注结果：“定一个[早上九点](TIME)在[北京](address)[评审](Theme)的日程”。

7.4 标注文件命名与存储要求

标注文件应由两部分组成，第一部分与对应标注文本命名一致，第二部分为“-bz”，应保存为 txt 等满足应用需求的格式，具体如：××项目_酒店预订语料_2022-06-08_1-bz。

8 样本标注流程

8.1 总体要求

样本标注应包含语料样本获取、语料样本检查、安全管控、标注工具选择、语料样本标注、标注结果收集和标注结果核查等环节，具体如图所示：

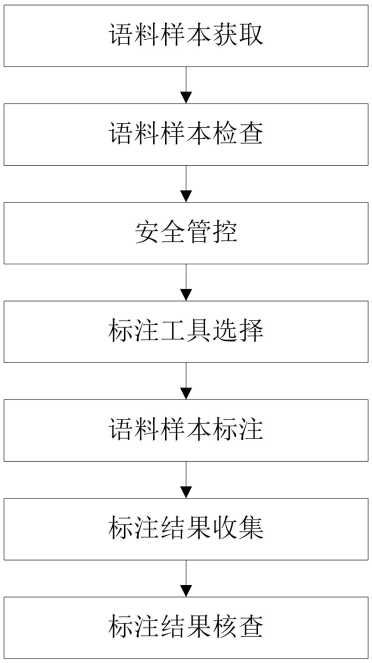

图 2 样本标注流程

8.2 样本获取

根据应用场景搜集整理相关语料样本数据，并按照第 6 章内容样本文件进行样本文件命名、创建样本描述文件等操作。

8.3 样本检查

在样本标注前应按照本文第 6 章要求对待标注样本进行检查，应根据业务需求和样本的数量采用全量检查或抽样检查，方式如下：

a) 全量检查应对指定范围内的所有样本进行逐条检查；

b) 抽样检查可采用随机抽样或分层抽样，方式如下：

1) 随机抽样，即：针对不同业务类型的数据样本采用随机抽样进行检查；

2) 分层抽样，即：针对同一业务类型的样本数据，根据样本类型不同采取分层抽样的方式进行检查。

8.4 安全管控

应满足对标注环境及标注人员的安全管控要求。具体要求包括：

a) 标注过程应在内网环境下的指定机器中进行，机器应开启防火墙，安装杀毒软件，并禁用 USB 接口功能。机器中的所有数据文件需定期做好数据备份，不得擅自拷贝、传输，防止数据丢失或泄露；

b) 标注人员应经过标注工作培训获得相关单位资格认证并签署样本标注保密协议后才可上岗操作。

8.5 标注工具选择

应使用标注格式通用、易操作的标注工具进行标注。

8.6 语料样本标注

8.6.1 基本要求

应根据业务需求和标注任务难易度选择人工标注或半自动化标注。

8.6.2 人工标注

人工标注任务应按照试标注、批量标注顺序执行，具体要求如下：

a) 试标注：

1) 从标注任务的待标注样本中抽取试标注样本。可采用随机抽样或分层抽样方法抽取样本，抽取比例不宜低于待标注样本总量的 1%；

2) 标注人员对抽取样本进行标注；

3) 标注项目负责人对标注结果进行确认；

4) 标注人员重复执行标注错误的标注任务，直至标注项目负责人确认无误。

b) 标注人员批量执行标注任务。

8.6.3.半自动化标注

半自动标注任务应按照样本构建、模型构建、模型批量标注顺序执行，具体要求如下：

a) 样本构建：

1) 从待标注样本中抽取训练样本和测试样本。可采用随机抽样或分层抽样方法抽取样本，训练样本与测试样本占样本总量的比例均不宜低于 1%，训练样本与测试样本的比例宜为 7:3，练样本与测试样本应无交集；

2) 标注人员通过人工方式标注训练样本和测试样本；

3) 标注项目负责人对标注结果进行确认；

4) 标注人员重复执行标注错误的标注任务，直至标注项目负责人确认无误。

b) 模型构建：

1) 使用标注后的训练样本建立标注模型；

2) 使用标注后的测试样本测试标注模型。可采用召回率、精确率指标评估模型性能；

c) 使用标注模型批量执行标注任务，并通过人工对模型标注的样本进行检查、修改和完善。

8.7 标注结果收集

标注结果收集应满足以下具体要求：

a) 样本标注结果应由统一的人员进行回收和存放，防止文件外泄；

b) 标注结果收集人员应对样本标注结果的相关信息（包括任务名称、任务类型、任务开始时间、任务结束时间、任务描述进行核对）进行检查，防止文件遗漏；

c) 标注结果收集人员，宜按照不同应用场景的标注结果对标注样本进行安全保存。

8.8 标注结果检查

在样本标注结果收集后应按照本文第 7 章要求对标注结果进行检查，应根据业务需求和样本标注的数量采用全量检查或抽样检查，方式如下：

a) 全量检查应对指定范围内的所有样本进行逐条检查；

b) 抽样检查可采用随机抽样或分层抽样，方式如下：

1) 随机抽样，即：针对不同业务类型的数据样本采用随机抽样进行检查；

2) 分层抽样，即：针对同一业务类型的样本数据，根据样本类型不同采取分层抽样的方式进行检查。