

一种样本的清洗方法、装置、电子设备及存储介质

相关申请的交叉引用

本申请要求享有于 2024 年 3 月 22 日提交的名称为“一种样本的清洗方法、装置、电子设备及存储介质”的中国专利申请 CN202410332295.6 的优先权，该申请的全部内容通过引用并入本文中。

技术领域

本申请涉及人工智能技术领域，尤其涉及一种样本的清洗方法、装置、电子设备及存储介质。

背景技术

语料库中文本标注的质量对文本模型的结果有着重要影响，高质量的语料库数据可以帮助文本模型训练后得到更好的业务结果，因而在文本分类任务中通常要求语料库内文本的标注结果准确且唯一。但是，实际业务中由于语料库内的文本来源广、增速快、会存在标注不准确以及对相同的文本前后标注结果不一致的情况，这样的标注结果会对模型训练后的效果产生不良影响。因此，需要清洗语料库以便提高语料库的数据质量，进而提高模型训练后的业务效果。

在现有技术中通常采用以下这些方法清洗语料库：方法一、对标注完成后的文本全部重新进行质量检查。该方法可以发现标注错误的样本，但是相当于进行了第二遍标注，人力和时间投入成本较高。方法二、对标注完成后的文本进行分批，在每一批数据进行随机抽样地标注质量检查。该方法避免了对文本进行全量质检，但是当每批数据的合格率都较低时相当于全量重标，那么成本投入和方法一相当。方法三、假定认为低质量文本容易出现标注错误，采用用于识别文本质量的分类模型将语料库中低质量的文本进行剔除。该方法可以得到一个标注错误较少的语料库，但是无法

判别高质量且标注错误的文本是否标注错误；由于低质量的文本存在一定的业务价值，对其进行剔除不符合业务要求，且额外训练一个判断文本质量高低的分类模型也存在较高的成本。

发明内容

本申请提供了一种样本的清洗方法、装置、电子设备及存储介质，可以较快地定位到语义库中标注错误的样本，所确定的标注错误样本准确率较高，有效降低了时间和人力的成本投入。

第一方面，本申请提供了一种样本的清洗方法，该方法包括：

对语料库中样本进行语义聚类处理得到多个聚类类别的样本集；

确定每个聚类类别的样本集中包含的标签类别，基于所述标签类别将所述每个聚类类别的样本集划分为多个样本子集；

针对所述每个聚类类别的样本集，从所述多个样本子集中确定出存在标注错误的目标样本子集，从而实现对所述语料库的样本清洗任务。

进一步的，所述对语料库中样本进行语义聚类处理得到多个聚类类别的样本集，包括：确定多个聚类特征和每个聚类特征对应的权重；基于所述多个聚类特征和所述每个聚类特征对应的权重对所述样本进行语义聚类处理得到所述多个聚类类别的样本集；其中，所述聚类特征包括文本编辑距离、关键词语义相似度和文本语义相似度中的至少一者；所述样本集包括多个样本和每个样本对应的标注标签。

进一步的，针对当前聚类类别的样本集，所述从所述多个样本子集中确定出存在标注错误的目标样本子集，包括：确定每个样本子集的样本数量；若所述多个样本子集的样本数量都相同，则确定所述当前聚类类别的样本集中不存在标注错误的目标样本子集；若所述多个样本子集的样本数量不相同，则根据所述多个样本子集的样本数据确定出最多样本数量和最少样本数量；基于所述每个样本子集的样本数量、所述最多样本数量和所

述最少样本数量依次计算所述每个样本子集是否为标注错误的目标样本子集,从而得到所述当前聚类类别的样本集中存在标注错误的目标样本子集。

进一步的,所述基于所述每个样本子集的样本数量、所述最多样本数量和所述最少样本数量依次计算所述每个样本子集是否为标注错误的目标样本子集,包括:针对当前样本子集,基于所述当前样本子集的样本数量、所述最多样本数量和所述最少样本数量计算出所述当前样本子集的错误标注概率;若所述错误标注概率大于第一预设数值,则确定所述当前样本子集为标注错误的目标样本子集;遍历所述多个样本子集,得到所述每个样本子集的错误标注概率,从而得到所述每个样本子集是否为标注错误的目标样本子集。

进一步的,基于所述当前样本子集的样本数量、所述最多样本数量和所述最少样本数量通过如下公式计算出所述当前样本子集的错误标注概率:

$$S = 1 - \frac{m_i}{m1 - m2}, m1 \neq m2$$

式中,S表示错误标注概率,i表示多个样本子集中当前样本子集的索引号, m_i 表示当前样本子集的样本数量,m1表示所述最多样本数量,m2表示所述最少样本数量。

进一步的,所述从所述多个样本子集中确定出存在标注错误的目标样本子集,包括:确定每个样本子集的样本数量;若所述样本数量小于第二预设数值,则确定所述样本数量对应的样本子集为存在标注错误的目标样本子集。

进一步的,在所述对语料库中样本进行语义聚类处理得到多个聚类类别的样本集之前,还包括:基于预设格式对所述语料库中样本进行格式转换得到处理之后的样本。

第二方面，本申请提供了一种样本的清洗装置，该装置包括：

样本聚类模块，用于对语料库中样本进行语义聚类处理得到多个聚类类别的样本集；

样本划分模块，用于确定每个聚类类别的样本集中包含的标签类别，基于所述标签类别将所述每个聚类类别的样本集划分为多个样本子集；

样本清洗模块，用于针对所述每个聚类类别的样本集，从所述多个样本子集中确定出存在标注错误的目标样本子集，从而实现对所述语料库的样本清洗任务。

第三方面，本申请提供了一种电子设备，该电子设备包括：至少一个处理器；以及与所述至少一个处理器通信连接的存储器；其中，所述存储器存储有可被所述至少一个处理器执行的计算机程序，所述计算机程序被所述至少一个处理器执行，以使所述至少一个处理器能够执行本申请任意实施例所述的样本的清洗方法。

第四方面，本申请提供了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机指令，所述计算机指令用于使处理器执行时实现本申请任意实施例所述的样本的清洗方法。

第五方面，本申请实施例提供一种计算机程序产品，包括计算机程序指令，所述计算机程序指令被处理器读取并运行时，执行实现本申请任意实施例所述的样本的清洗方法。

为了解决背景技术中现有技术的缺陷，本申请实施例提供了一种样本的清洗方法，执行该方法能够带来以下有益效果：本申请先对语义库中样本聚类成多个样本集，再对每个样本集按照其所包含的标签类别划分为对应的样本子集，最后分析每个样本子集确定出每个样本子集是否存在标注错误。本申请可以较快地定位到语义库中标注错误的样本，所确定的标注错误样本准确率较高，有效降低了时间和人力的成本投入。

需要说明的是，上述计算机指令可以全部或者部分存储在计算机可读

存储介质上。其中，计算机可读存储介质可以与样本的清洗装置的处理器封装在一起，也可以与样本的清洗装置的处理器单独封装，本申请对此不做限定。

本申请中第二方面、第三方面以及第四方面的描述，可以参考第一方面的详细描述；并且，第二方面、第三方面以及第四方面的描述的有益效果，可以参考第一方面的有益效果分析，此处不再赘述。

应当理解，本部分所描述的内容并非旨在标识本申请的实施例的关键或重要特征，也不用于限制本申请的范围。本申请的其他特征将通过以下的说明书而变得容易理解。

可以理解的是，在使用本申请各实施例公开的技术方案之前，均应当依据相关法律法规通过恰当的方式对本申请所涉及个人信息类型、使用范围以及使用场景等告知用户并获得用户的授权。

附图说明

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图1为本申请实施例提供一种样本的清洗方法的第一流程示意图；

图2为本申请实施例提供一种样本的清洗方法的第二流程示意图；

图3为本申请实施例提供一种样本的清洗装置的结构示意图；

图4是用来实现本申请实施例的一种样本的清洗方法的电子设备的框图。

具体实施方式

为使本申请实施例的目的、技术方案和优点更加清楚，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描

述，显然，所描述的实施例仅是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都应当属于本申请保护的范围。

需要说明的是，本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”、“目标”以及“原始”等是用于区别类似的对象，而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换，以便这里描述的本申请的实施例能够实施除了在这里图示或描述之外的顺序。此外，术语“包括”“具有”以及他们的任何变形，意图在于覆盖不排他的包含，例如，包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元，而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

图 1 为本申请实施例提供的一种样本的清洗方法的第一流程示意图，本实施例可适用于语料库中样本进行清洗的情况。本实施例提供的一种样本的清洗方法可以由本申请实施例提供的样本的清洗装置来执行，该装置可以通过软件和/或硬件的方式实现，并集成在执行本方法的电子设备中。

参见图 1，本实施例的方法包括但不限于如下步骤：

S110、对语料库中样本进行语义聚类处理得到多个聚类类别的样本集。

其中，语料库为经过第一遍标注后的语料库，样本集包括多个样本和每个样本对应的标注标签。

进一步的，由于语料库中可能存在不同格式的样本，因而在对语料库中样本进行语义聚类处理得到多个聚类类别的样本集之前，还包括：基于预设格式对语料库中样本进行统一格式转换得到处理之后的样本，从而将语料库制作成待聚类的样本，样本以文本加标签的形式呈现。

具体的，对语料库中样本进行语义聚类处理得到多个聚类类别的样本集，包括：确定多个聚类特征和每个聚类特征对应的权重；基于多个聚类特征和每个聚类特征对应的权重对样本进行语义聚类处理得到多个聚类类

别的样本集。其中，聚类特征包括文本编辑距离、关键词语义相似度和文本语义相似度中的至少一者；样本集包括多个样本和每个样本对应的标注标签。根据本申请的一个优选实施例，聚类特征例如可以包括文本编辑距离、关键词语义相似度和文本语义相似度。

在内容安全业务场景的实施例中，可以将文本编辑距离的权重设置为 0.3，将关键词语义相似度的权重设置为 0.3，将文本语义相似度的权重设置为 0.4。本实施例的聚类特征和对应的权重可以根据业务场景不同自行调整。相比于直接使用文本语义进行聚类，该聚类方法的准确率更高，适合内容安全业务场景。

可选的，文本编辑距离和语义相似度得分可以采用例如 K 近邻算法 (k-NearestNeighbor, kNN) 或其他聚类方法对语料库中样本进行语义聚类，越准确的聚类方法可以进一步缩减为语义库中待质检样本的范围，减少质检成本投入。

示例性的，假设某语料库内有 1 万条样本，每条样本由文本和对应的标签组成，假设标签分别为“a, b, c, d, e, f”。对文本的文本编辑距离、关键词语义相似度和文本语义相似度分配不同的权重，通过语义聚类处理产生了 100 个聚类类别的样本集，标号分别为 0 至 99。

S120、确定每个聚类类别的样本集中包含的标签类别，基于标签类别将每个聚类类别的样本集划分为多个样本子集。

其中，样本子集的数量和样本集中所包含标签类别的数量是一样的，一个标签类别对应一个样本子集。

示例性地，假设统计得到标号为 0 的样本集中共有样本 500 条，标签类别有五个，分别为“a, b, c, d, e”。那么将这 500 条样本按照标签类别划分为五个样本子集。

S130、针对每个聚类类别的样本集，从多个样本子集中确定出存在标注错误的目标样本子集，从而实现对语料库的样本清洗任务。

在本申请实施例中，针对当前聚类类别的样本集，从当前聚类类别的样本集中多个样本子集里面确定出存在标注错误的目标样本子集；遍历完所有聚类类别的样本集，可以得到每个样本集中存在标注错误的目标样本子集，从而可以实现对语料库的样本清洗任务。其中，目标样本子集为语义库中待质检的样本。

具体的，从多个样本子集中确定出存在标注错误的目标样本子集，包括：确定每个样本子集的样本数量；若样本数量小于第二预设数值，则确定样本数量对应的样本子集为存在标注错误的目标样本子集。

在本申请实施例中，本领域通常认为具有相似语义的文本大概率会有相同的标签，因此可以将语料库中的样本进行聚类，进而统计每个聚类内每个标签的样本数量分布。

本领域通常认为在聚类内样本数量较少的标签，其对应的文本大概率是标注错误的文本，需要进行质量检查。第二预设数值可以根据实际应用需求进行设定。

示例性地，假设统计得到标号为 0 的样本集中共有样本 500 条，标签类别有五个，分别为“a, b, c, d, e”，其中，a 标签的样本子集的样本数量为 300 条，b 标签的样本子集的样本数量为 100 条，c 标签的样本子集的样本数量为 80 条，d 标签的样本子集的样本数量为 10 条，e 标签的样本子集的样本数量为 10 条。此时该聚类类别内，d 标签和 e 标签对应的样本大概率是标注错误的样本，应当对其进行质量检查。基于同样的方法，遍历完对聚类类别从 0 至 99 的样本集，可以筛选出大概率为标注错误的样本。

本实施例提供的技术方案，通过对语料库中样本进行语义聚类处理得到多个聚类类别的样本集；确定每个聚类类别的样本集中包含的标签类别，基于标签类别将每个聚类类别的样本集划分为多个样本子集；针对每个聚类类别的样本集，从多个样本子集中确定出存在标注错误的目标样本子集，从而实现对语料库的样本清洗任务。本申请先对语义库中样本聚类成多个

样本集，再对每个样本集按照其所包含的标签类别划分为对应的样本子集，最后分析每个样本子集确定出每个样本子集是否存在标注错误。本申请可以较快地定位到语义库中标注错误的样本，所确定的标注错误样本准确率较高，有效降低了时间和人力的成本投入。

下面进一步描述本申请实施例提供的样本的清洗方法，图 2 为本申请实施例提供的一种样本的清洗方法的第二流程示意图。本申请实施例是在上述各实施例的基础上进行优化，具体优化为：针对当前聚类类别的样本集，本实施例对从多个样本子集中确定出存在标注错误的目标样本子集的过程进行详细的解释说明。

参见图 2，本实施例的方法包括但不限于如下步骤：

S210、确定每个样本子集的样本数量。

在本申请实施例中，对语料库中样本进行语义聚类处理得到多个聚类类别的样本集，确定每个聚类类别的样本集中包含的标签类别，基于标签类别将每个聚类类别的样本集划分为多个样本子集。针对每个聚类类别的样本集，确定每个样本子集的样本数量，并确定每个样本子集的样本数量是否相同。

S220、若多个样本子集的样本数量都相同，则确定当前聚类类别的样本集中不存在标注错误的目标样本子集。

在本申请实施例中，当多个样本子集的样本数量相同时，表明对应的样本集的错误标注概率为 0，表示该样本集中无标注错误的样本。

S230、若多个样本子集的样本数量不相同，则根据多个样本子集的样本数据确定出最多样本数量和最少样本数量。

在本申请实施例中，当多个样本子集的样本数量不相同，确定出哪个标签类别的样本子集的样本数量最多，得到最多样本数量；确定出哪个标签类别的样本子集的样本数量最少，得到最少样本数量。

S240、针对当前样本子集，基于当前样本子集的样本数量、最多样本

数量和最少样本数量计算出当前样本子集的错误标注概率。

具体的，基于当前样本子集的样本数量、最多样本数量和最少样本数量通过如下公式计算出当前样本子集的错误标注概率：

$$S = 1 - \frac{m_i}{m1 - m2}, m1 \neq m2$$

式中，S 表示错误标注概率（相当于标签错误率得分），是各个聚类类别的样本集内标签数量的归一化分数，i 表示多个样本子集中当前样本子集的索引号， m_i 表示当前样本子集的样本数量，m1 表示所述最多样本数量，m2 表示所述最少样本数量。

某样本子集对应 S 的数值越大，表示该样本子集存在错误标注(即该样本子集为目标样本子集)的概率也越大，反之亦然。某样本子集的错误标注概率与该样本子集内的标签类别数量有关。

S250、若错误标注概率大于第一预设数值，则确定当前样本子集为标注错误的目标样本子集。

第一预设数值可以根据实际应用场景不同进行设置，根据本申请的一个优选实施例，在业务安全场景下，第一预设数值例如可以设置为 0.8。经实验证明，采用传统方法所确定的待质检样本范围占语义库样本总量的 50% 至 100%。在业务安全场景下，对错误标注概率大于 0.8 的样本子集进行质量检查，最终仅需检查一小部分样本(约语义库样本总量的 10%，较传统方法节约人力成本 50% 以上)，就可以最终令语料库达到与传统方法相同的样本准确率(一般准确率为 99.5% 以上)，清洗语料库的成本投入降低明显。

S260、遍历完多个样本子集，得到每个样本子集的错误标注概率，从而得到当前聚类类别的样本集中存在标注错误的目标样本子集。

在图 1 对应实施例中是采用某样本子集的样本数量小于第二预设数值时认为该样本子集存在标注错误。由于每个聚类类别的样本集内标签类别

的数量都不一样，较难对每个聚类类别的样本集统一设置第二预设数值的数值，那么还可以采用本实施例中的计算样本子集的错误标注概率的方法来判别该样本子集是否存在标注错误。

本实施例提供的技术方案，通过确定每个样本子集的样本数量；若多个样本子集的样本数量都相同，则确定当前聚类类别的样本集中不存在标注错误的目标样本子集；若多个样本子集的样本数量不相同，则根据多个样本子集的样本数据确定出最多样本数量和最少样本数量；针对当前样本子集，基于当前样本子集的样本数量、最多样本数量和最少样本数量计算出当前样本子集的错误标注概率；若错误标注概率大于第一预设数值，则确定当前样本子集为标注错误的目标样本子集；遍历完多个样本子集，得到每个样本子集的错误标注概率，从而得到当前聚类类别的样本集中存在标注错误的目标样本子集。本申请通过计算每个样本子集的错误标注概率的方法来判别每个样本子集是否存在标注错误，相比传统样本清洗方法可以减小目标样本子集(即待质检样本)的数量；本申请还可以较快地定位到语义库中标注错误的样本，所确定的标注错误样本准确率较高，有效降低了时间和人力的成本投入。

图3为本申请实施例提供的一种样本的清洗装置的结构示意图，如图3所示，该装置300可以包括：

样本聚类模块310，用于对语料库中样本进行语义聚类处理得到多个聚类类别的样本集；

样本划分模块320，用于确定每个聚类类别的样本集中包含的标签类别，基于所述标签类别将所述每个聚类类别的样本集划分为多个样本子集；

样本清洗模块330，用于针对所述每个聚类类别的样本集，从所述多个样本子集中确定出存在标注错误的目标样本子集，从而实现对所述语料库的样本清洗任务。

进一步的，上述样本聚类模块310，可以具体用于：确定多个聚类特征

和每个聚类特征对应的权重；基于所述多个聚类特征和所述每个聚类特征对应的权重对所述样本进行语义聚类处理得到所述多个聚类类别的样本集；其中，所述聚类特征包括文本编辑距离、关键词语义相似度和文本语义相似度中的至少一者；所述样本集包括多个样本和每个样本对应的标注标签。

进一步的，上述样本清洗模块 330，可以具体用于：针对当前聚类类别的样本集，确定每个样本子集的样本数量；若所述多个样本子集的样本数量都相同，则确定所述当前聚类类别的样本集中不存在标注错误的目标样本子集；若所述多个样本子集的样本数量不相同，则根据所述多个样本子集的样本数据确定出最多样本数量和最少样本数量；基于所述每个样本子集的样本数量、所述最多样本数量和所述最少样本数量依次计算所述每个样本子集是否为标注错误的目标样本子集，从而得到所述当前聚类类别的样本集中存在标注错误的目标样本子集。

进一步的，上述样本清洗模块 330，还可以具体用于：针对当前样本子集，基于所述当前样本子集的样本数量、所述最多样本数量和所述最少样本数量计算出所述当前样本子集的错误标注概率；若所述错误标注概率大于第一预设数值，则确定所述当前样本子集为标注错误的目标样本子集；遍历所述多个样本子集，得到所述每个样本子集的错误标注概率，从而得到所述每个样本子集是否为标注错误的目标样本子集。

进一步的，上述样本清洗模块 330，还可以具体用于：基于所述当前样本子集的样本数量、所述最多样本数量和所述最少样本数量通过如下公式计算出所述当前样本子集的错误标注概率：

$$S = 1 - \frac{m_i}{m1 - m2}, m1 \neq m2$$

式中，S 表示错误标注概率，i 表示多个样本子集中当前样本子集的索引号， m_i 表示当前样本子集的样本数量，m1 表示所述最多样本数量，m2

表示所述最少样本数量。

进一步的，上述样本清洗模块 330，可以具体用于：确定每个样本子集的样本数量；若所述样本数量小于第二预设数值，则确定所述样本数量对应的样本子集为存在标注错误的目标样本子集。

进一步的，上述样本的清洗装置，还可以包括：样本处理模块；

所述样本处理模块，用于在所述对语料库中样本进行语义聚类处理得到多个聚类类别的样本集之前，基于预设格式对所述语料库中样本进行格式转换得到处理之后的样本。

本实施例提供的样本的清洗装置可适用于上述任意实施例提供的样本的清洗方法，具备相应的功能和有益效果。

图 4 是用来实现本申请实施例的一种样本的清洗方法的电子设备的框图。电子设备 10 旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理器、蜂窝电话、智能电话、可穿戴设备(如头盔、眼镜、手表等)和其它类似的计算装置。本文所示的部件、它们的连接和关系以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本申请的实现。

如图 4 所示，电子设备 10 包括至少一个处理器 11，以及与至少一个处理器 11 通信连接的存储器，如只读存储器(ROM)12、随机访问存储器(RAM)13 等，其中，存储器存储有可被至少一个处理器执行的计算机程序，处理器 11 可以根据存储在只读存储器(ROM)12 中的计算机程序或者从存储单元 18 加载到随机访问存储器(RAM)13 中的计算机程序，来执行各种适当的动作和处理。在 RAM13 中，还可存储电子设备 10 操作所需的各种程序和数据。处理器 11、ROM12 以及 RAM13 通过总线 14 彼此相连。输入/输出(I/O)接口 15 也连接至总线 14。

电子设备 10 中的多个部件连接至 I/O 接口 15，包括：输入单元 16，例如键盘、鼠标等；输出单元 17，例如各种类型的显示器、扬声器等；存储单元 18，例如磁盘、光盘等；以及通信单元 19，例如网卡、调制解调器、无线通信收发机等。通信单元 19 允许电子设备 10 通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

处理器 11 可以是各种具有处理和计算能力的通用和/或专用处理组件。处理器 11 的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的处理器、数字信号处理器(DSP)以及任何适当的处理器、控制器、微控制器等。处理器 11 执行上文所描述的各个方法和处理，例如样本的清洗方法。

在一些实施例中，样本的清洗方法可被实现为计算机程序，其被有形地包含于计算机可读存储介质，例如存储单元 18。在一些实施例中，计算机程序的部分或者全部可以经由 ROM12 和/或通信单元 19 而被载入和/或安装到电子设备 10 上。当计算机程序加载到 RAM13 并由处理器 11 执行时，可以执行上文描述的样本的清洗方法的一个或多个步骤。备选地，在其他实施例中，处理器 11 可以通过其他任何适当的方式(例如，借助于固件)而被配置为执行样本的清洗方法。

本文中以上描述的系统和技术和各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、负载可编程逻辑设备(CPLD)、计算机硬件、固件、软件和/或它们的组合中实现。这些各种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置和该至少一个输出装置。

用于实施本申请的方法的计算机程序可以采用一个或多个编程语言的任何组合来编写。这些计算机程序可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器，使得计算机程序当由处理器执行时使流程图和/或框图中所规定的功能/操作被实施。计算机程序可以完全在机器上执行、部分地在机器上执行，作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

在本申请的上下文中，计算机可读存储介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的计算机程序。计算机可读存储介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的或半导体系统、装置或设备，或者上述内容的任何合适组合。备选地，计算机可读存储介质可以是机器可读信号介质。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM 或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备或上述内容的任何合适组合。

为了提供与用户的交互，可以在电子设备上实施此处描述的系统和技术，该电子设备具有：用于向用户显示信息的显示装置(例如，CRT(阴极射线管)或者 LCD(液晶显示器)监视器)；以及键盘和指向装置(例如，鼠标或者轨迹球)，用户可以通过该键盘和该指向装置来将输入提供给电子设备。其它种类的装置还可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈(例如，视觉反馈、听觉反馈或者触觉反馈)；并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如，作为数据服务器)或者包括中间件部件的计算系统(例如，应用服务器)或者

包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)或者包括这种后台部件、中间件部件或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)、区块链网络和互联网。

计算系统可以包括客户端和服务端。客户端和服务端一般远离彼此并且通常通过通信网络进行交互。客户端和服务端之间的关系是通过在各自计算机上运行的程序生成的,而这些程序之间存在着客户端-服务端的关系。服务端可以是云服务器,又称为云计算服务端或云主机,是云计算服务体系中的一项主机产品,以解决了传统物理主机与 VPS 服务中,存在的管理难度大,业务扩展性弱的缺陷。

注意,上述仅为本申请的较佳实施例及所运用技术原理。本领域技术人员会理解,本申请不限于这里所述的特定实施例,对本领域技术人员来说能够进行各种明显的变化、重新调整和替代而不会脱离本申请的保护范围。例如,本领域技术人员可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤;可以并行地执行、顺序地执行或者不同的次序执行本申请中记载的各步骤,只要能够实现本申请的技术方案所期望的结果,本文在此不进行限制。

上述具体实施方式并不构成对本申请保护范围的限制。本领域技术人员应该明白的是,根据设计要求和因素可以进行各种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等,均应包含在本申请保护范围之内。

权利要求书

1. 一种样本的清洗方法，其特征在于，所述方法包括：
对语料库中样本进行语义聚类处理得到多个聚类类别的样本集；
确定每个聚类类别的样本集中包含的标签类别，基于所述标签类别将
5 所述每个聚类类别的样本集划分为多个样本子集；
针对所述每个聚类类别的样本集，从所述多个样本子集中确定出存在
标注错误的目标样本子集，从而实现对所述语料库的样本清洗任务。
2. 根据权利要求1所述的样本的清洗方法，其特征在于，所述对语料
库中样本进行语义聚类处理得到多个聚类类别的样本集，包括：
10 确定多个聚类特征和每个聚类特征对应的权重；
基于所述多个聚类特征和所述每个聚类特征对应的权重对所述样本
进行语义聚类处理得到所述多个聚类类别的样本集；
其中，所述聚类特征包括文本编辑距离、关键词语义相似度和文本语
义相似度中的至少一者；所述样本集包括多个样本和每个样本对应的标注
15 标签。
3. 根据权利要求1所述的样本的清洗方法，其特征在于，针对当前聚
类类别的样本集，所述从所述多个样本子集中确定出存在标注错误的目标
样本子集，包括：
确定每个样本子集的样本数量；
20 若所述多个样本子集的样本数量都相同，则确定所述当前聚类类别的
样本集中不存在标注错误的目标样本子集；
若所述多个样本子集的样本数量不相同，则根据所述多个样本子集的
样本数据确定出最多样本数量和最少样本数量；
基于所述每个样本子集的样本数量、所述最多样本数量和所述最少样
25 本数量依次计算所述每个样本子集是否为标注错误的目标样本子集，从而

得到所述当前聚类类别的样本集中存在标注错误的目标样本子集。

4.根据权利要求 3 所述的样本的清洗方法，其特征在于，所述基于所述每个样本子集的样本数量、所述最多样本数量和所述最少样本数量依次计算所述每个样本子集是否为标注错误的目标样本子集，包括：

5 针对当前样本子集，基于所述当前样本子集的样本数量、所述最多样本数量和所述最少样本数量计算出所述当前样本子集的错误标注概率；

 若所述错误标注概率大于第一预设数值，则确定所述当前样本子集为标注错误的目标样本子集；

 遍历所述多个样本子集，得到所述每个样本子集的错误标注概率，从而得到所述每个样本子集是否为标注错误的目标样本子集。
10

5.根据权利要求 4 所述的样本的清洗方法，其特征在于，基于所述当前样本子集的样本数量、所述最多样本数量和所述最少样本数量通过如下公式计算出所述当前样本子集的错误标注概率：

$$S = 1 - \frac{m_i}{m1 - m2}, m1 \neq m2$$

15 式中，S 表示错误标注概率，i 表示多个样本子集中当前样本子集的索引号， m_i 表示当前样本子集的样本数量，m1 表示所述最多样本数量，m2 表示所述最少样本数量。

6.根据权利要求 1 所述的样本的清洗方法，其特征在于，所述从所述多个样本子集中确定出存在标注错误的目标样本子集，包括：

20 确定每个样本子集的样本数量；

 若所述样本数量小于第二预设数值，则确定所述样本数量对应的样本子集为存在标注错误的目标样本子集。

7.根据权利要求 1 所述的样本的清洗方法，其特征在于，在所述对语料库中样本进行语义聚类处理得到多个聚类类别的样本集之前，还包括：

25 基于预设格式对所述语料库中样本进行格式转换得到处理之后的样

本。

8.一种样本的清洗装置，其特征在于，所述装置包括：

样本聚类模块，用于对语料库中样本进行语义聚类处理得到多个聚类类别的样本集；

5 样本划分模块，用于确定每个聚类类别的样本集中包含的标签类别，基于所述标签类别将所述每个聚类类别的样本集划分为多个样本子集；

样本清洗模块，用于针对所述每个聚类类别的样本集，从所述多个样本子集中确定出存在标注错误的目标样本子集，从而实现对所述语料库的样本清洗任务。

10 9.一种电子设备，其特征在于，所述电子设备包括：至少一个处理器；以及与所述至少一个处理器通信连接的存储器；其中，所述存储器存储有可被所述至少一个处理器执行的计算机程序，所述计算机程序被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求1至7中任一所述的样本的清洗方法。

15 10.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有计算机指令，所述计算机指令用于使处理器执行时实现权利要求1至7中任一所述的样本的清洗方法。

11.一种计算机程序产品，其特征在于，包括计算机程序指令，所述计算机程序指令被处理器读取并运行时，执行如权利要求1至7任一所述的样本的清洗方法。
20