(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119378564 A
(43)申请公布日 2025.01.28

(21)申请号 202411949511.8

(22)申请日 2024.12.27

(71)申请人 数据堂(北京)科技股份有限公司
地址 100080 北京市海淀区宝盛南路1号院11号楼1层101-01

(72)发明人 齐红威 王大亮 丰强泽 栗全峰 高禹 郑继龙

(74)专利代理机构 郑州欧凯专利代理事务所(普通合伙) 41166
专利代理师 毛志强

(51)Int.Cl.
G06F 40/30(2020.01)
G06F 40/126(2020.01)
G06F 40/216(2020.01)

权利要求书3页 说明书12页 附图6页

(54)发明名称

大模型数据智能标注方法及系统

(57)摘要

本申请涉及数据标注技术领域,其具体地公开了一种大模型数据智能标注方法及系统,其采用基于深度学习的自然语言处理技术对未标注文本数据集中的各个未标注文本数据进行置信度评估,选择最小置信度对应的文本数据作为代表样本数据,并对所述代表样本数据进行语料扩充,进而,通过对所述代表样本数据和语料扩充后的代表样本数据进行语义特征提取和补偿式交互融合,以充分利用两者之间的共有信息和独特信息,从而实现对所述代表样本数据的全面语义理解和智能标注。通过这种方式,可以显著提高数据标注的效率和准确性,同时大幅度减少人工干预的需求,降低标注成本。

CN 119378564 A

1.一种大模型数据智能标注方法,其特征在于,包括:

获取未标注文本数据集;

对所述未标注文本数据集中的各个未标注文本数据进行置信度评估以确定代表样本数据;

对所述代表样本数据进行语料扩充以得到代表样本数据扩充描述;

分别对所述代表样本数据和所述代表样本数据扩充描述进行语义编码以得到代表样本数据语义编码特征向量和代表样本数据扩充描述语义编码特征向量;

对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行补偿式交互融合以得到代表样本数据语义交互补偿融合表示向量;

基于所述代表样本数据语义交互补偿融合表示向量,生成标注结果;

其中,对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行补偿式交互融合,包括:提取所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量之间的共性特征以得到代表样本数据特征间共性特征表示向量;以所述代表样本数据特征间共性特征表示向量为掩码,引导所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行特征互补性增强交互融合以得到所述代表样本数据语义交互补偿融合表示向量。

2.根据权利要求1所述的大模型数据智能标注方法,其特征在于,对所述未标注文本数据集中的各个未标注文本数据进行置信度评估以确定代表样本数据,包括:

对所述未标注文本数据集中的各个未标注文本数据进行语义嵌入编码以得到未标注数据语义嵌入编码向量的集合;

将所述未标注数据语义嵌入编码向量的集合中的各个未标注数据语义嵌入编码向量输入基于分类器的置信度评估器以得到未标注样本置信度的集合;

选择所述未标注样本置信度的集合中最小置信度对应的未标注文本数据作为所述代表样本数据。

3.根据权利要求2所述的大模型数据智能标注方法,其特征在于,对所述代表样本数据进行语料扩充以得到代表样本数据扩充描述,包括:

将所述代表样本数据输入基于大语言模型的语料扩充模块以得到所述代表样本数据扩充描述。

4.根据权利要求3所述的大模型数据智能标注方法,其特征在于,提取所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量之间的共性特征以得到代表样本数据特征间共性特征表示向量,包括:

将所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量输入共性特征提取网络以得到所述代表样本数据特征间共性特征表示向量,其中,所述共性特征提取网络对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行点加融合后,使用基于tanh函数的神经网络层对其进行共性特征提取以得到所述代表样本数据特征间共性特征表示向量。

5.根据权利要求4所述的大模型数据智能标注方法,其特征在于,以所述代表样本数据特征间共性特征表示向量为掩码,引导所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行特征互补性增强交互融合以得到所述代表样本数

据语义交互补偿融合表示向量,包括:

基于所述代表样本数据语义编码特征向量、所述代表样本数据扩充描述语义编码特征向量相对于所述代表样本数据特征间共性特征表示向量的特定特征,对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行特征补偿调制以得到代表样本数据语义特征补偿向量和代表样本数据扩充描述语义特征补偿向量;

将所述代表样本数据语义特征补偿向量、所述代表样本数据扩充描述语义特征补偿向量和所述代表样本数据特征间共性特征表示向量进行级联以得到所述代表样本数据特征间共性特征表示向量。

6.根据权利要求5所述的大模型数据智能标注方法,其特征在于,基于所述代表样本数据语义编码特征向量、所述代表样本数据扩充描述语义编码特征向量相对于所述代表样本数据特征间共性特征表示向量的特定特征,对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行特征补偿调制以得到代表样本数据语义特征补偿向量和代表样本数据扩充描述语义特征补偿向量,包括:

将所述代表样本数据语义编码特征向量、所述代表样本数据扩充描述语义编码特征向量和所述代表样本数据特征间共性特征表示向量输入基于Sigmoid函数的向量概率化单元以得到概率化代表样本数据语义编码特征向量、概率化代表样本数据扩充描述语义编码特征向量和概率化代表样本数据特征间共性特征表示向量;

计算所述概率化代表样本数据语义编码特征向量相对于所述概率化代表样本数据特征间共性特征表示向量的特定特征以得到代表样本数据特定特征补偿表示向量;

计算所述概率化代表样本数据扩充描述语义编码特征向量相对于所述概率化代表样本数据特征间共性特征表示向量的特定特征以得到代表样本数据扩充描述特定特征补偿表示向量;

将所述代表样本数据特定特征补偿表示向量和所述代表样本数据语义编码特征向量输入细粒度补偿模块以得到所述代表样本数据语义特征补偿向量;

将所述代表样本数据扩充描述特定特征补偿表示向量和所述代表样本数据扩充描述语义编码特征向量输入所述细粒度补偿模块以得到所述代表样本数据扩充描述语义特征补偿向量。

7.根据权利要求6所述的大模型数据智能标注方法,其特征在于,计算所述概率化代表样本数据语义编码特征向量相对于所述概率化代表样本数据特征间共性特征表示向量的特定特征以得到代表样本数据特定特征补偿表示向量,包括:

计算所述概率化代表样本数据语义编码特征向量和所述概率化代表样本数据特征间共性特征表示向量之间的点除向量,并计算所述点除向量中各个特征值的绝对值的以二为底的对数值以得到代表样本数据特定特征向量;

计算所述代表样本数据特定特征向量与所述概率化代表样本数据语义编码特征向量之间的点乘向量,并计算以e为底,以所述点乘向量的各个特征值为指数的指数函数值以得到指数化代表样本数据特定特征表示向量;

将所述指数化代表样本数据特定特征表示向量输入softmax函数进行归一化处理以得到所述代表样本数据特定特征补偿表示向量。

8.根据权利要求7所述的大模型数据智能标注方法,其特征在于,将所述代表样本数据

特定特征补偿表示向量和所述代表样本数据语义编码特征向量输入细粒度补偿模块以得到所述代表样本数据语义特征补偿向量,包括:

计算所述代表样本数据特定特征补偿表示向量和所述代表样本数据语义编码特征向量之间的哈达玛积以得到所述代表样本数据语义特征补偿向量。

9.根据权利要求8所述的大模型数据智能标注方法,其特征在于,基于所述代表样本数据语义交互补偿融合表示向量,生成标注结果,包括:

将所述代表样本数据语义交互补偿融合表示向量输入基于分类器的智能标注模块以得到所述标注结果。

10.一种大模型数据智能标注系统,其特征在于,包括:

数据集获取模块,用于获取未标注文本数据集;

置信度评估模块,用于对所述未标注文本数据集中的各个未标注文本数据进行置信度评估以确定代表样本数据;

语料扩充模块,用于对所述代表样本数据进行语料扩充以得到代表样本数据扩充描述;

语义编码模块,用于分别对所述代表样本数据和所述代表样本数据扩充描述进行语义编码以得到代表样本数据语义编码特征向量和代表样本数据扩充描述语义编码特征向量;

特征交互融合模块,用于对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行补偿式交互融合以得到代表样本数据语义交互补偿融合表示向量;

标注结果生成模块,用于基于所述代表样本数据语义交互补偿融合表示向量,生成标注结果。

大模型数据智能标注方法及系统

技术领域

[0001] 本申请涉及数据标注技术领域,且更为具体地,涉及一种大模型数据智能标注方法及系统。

背景技术

[0002] 随着互联网和信息技术的快速发展,数据量呈爆炸性增长。在大数据时代背景下,数据的质量和价值越来越受到重视,而高质量的数据标注是机器学习尤其是深度学习领域不可或缺的一环。具体而言,数据标注是指为原始数据添加标签的过程,以便于用于训练监督学习算法,使其能够从已标注的数据中学习并应用于新的未见过的数据上。

[0003] 然而,传统的人工标注方式虽然能够确保标注质量,但面对规模庞大的数据集时,其效率低下且成本高昂,难以满足实际应用的需求。而大多数的自动化标注方法主要依赖于预定义的规则或者简单的统计模型,缺乏对数据集的深入语义理解,从而导致标注结果的准确性和可靠性不足。

[0004] 基于此,期待一种优化的大模型数据智能标注方法及系统。

发明内容

[0005] 为了解决上述技术问题,提出了本申请。本申请的实施例提供了一种大模型数据智能标注方法及系统,其采用基于深度学习的自然语言处理技术对未标注文本数据集中的各个未标注文本数据进行置信度评估,选择最小置信度对应的文本数据作为代表样本数据,并对所述代表样本数据进行语料扩充,进而,通过对所述代表样本数据和语料扩充后的代表样本数据进行语义特征提取和补偿式交互融合,以充分利用两者之间的共有信息和独特信息,从而实现对所述代表样本数据的全面语义理解和智能标注。通过这种方式,可以显著提高数据标注的效率和准确性,同时大幅度减少人工干预的需求,降低标注成本。

[0006] 根据本申请的一个方面,提供了一种大模型数据智能标注方法,其包括:

获取未标注文本数据集;

对所述未标注文本数据集中的各个未标注文本数据进行置信度评估以确定代表样本数据;

对所述代表样本数据进行语料扩充以得到代表样本数据扩充描述;

分别对所述代表样本数据和所述代表样本数据扩充描述进行语义编码以得到代表样本数据语义编码特征向量和代表样本数据扩充描述语义编码特征向量;

对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行补偿式交互融合以得到代表样本数据语义交互补偿融合表示向量;

基于所述代表样本数据语义交互补偿融合表示向量,生成标注结果。

[0007] 在上述大模型数据智能标注方法中,对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行补偿式交互融合,包括:提取所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量之间的共性特

征以得到代表样本数据特征间共性特征表示向量;以所述代表样本数据特征间共性特征表示向量为掩码,引导所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行特征互补性增强交互融合以得到所述代表样本数据语义交互补偿融合表示向量。

[0008] 在上述大模型数据智能标注方法中,对所述未标注文本数据集中的各个未标注文本数据进行置信度评估以确定代表样本数据,包括:对所述未标注文本数据集中的各个未标注文本数据进行语义嵌入编码以得到未标注数据语义嵌入编码向量的集合;将所述未标注数据语义嵌入编码向量的集合中的各个未标注数据语义嵌入编码向量输入基于分类器的置信度评估器以得到未标注样本置信度的集合;选择所述未标注样本置信度的集合中最小置信度对应的未标注文本数据作为所述代表样本数据。

[0009] 在上述大模型数据智能标注方法中,对所述代表样本数据进行语料扩充以得到代表样本数据扩充描述,包括:将所述代表样本数据输入基于大语言模型的语料扩充模块以得到所述代表样本数据扩充描述。

[0010] 在上述大模型数据智能标注方法中,提取所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量之间的共性特征以得到代表样本数据特征间共性特征表示向量,包括:将所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量输入共性特征提取网络以得到所述代表样本数据特征间共性特征表示向量,其中,所述共性特征提取网络对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行点加融合后,使用基于tanh函数的神经网络层对其进行共性特征提取以得到所述代表样本数据特征间共性特征表示向量。

[0011] 在上述大模型数据智能标注方法中,以所述代表样本数据特征间共性特征表示向量为掩码,引导所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行特征互补性增强交互融合以得到所述代表样本数据语义交互补偿融合表示向量,包括:基于所述代表样本数据语义编码特征向量、所述代表样本数据扩充描述语义编码特征向量相对于所述代表样本数据特征间共性特征表示向量的特定特征,对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行特征补偿调制以得到代表样本数据语义特征补偿向量和代表样本数据扩充描述语义特征补偿向量;将所述代表样本数据语义特征补偿向量、所述代表样本数据扩充描述语义特征补偿向量和所述代表样本数据特征间共性特征表示向量进行级联以得到所述代表样本数据特征间共性特征表示向量。

[0012] 在上述大模型数据智能标注方法中,基于所述代表样本数据语义编码特征向量、所述代表样本数据扩充描述语义编码特征向量相对于所述代表样本数据特征间共性特征表示向量的特定特征,对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行特征补偿调制以得到代表样本数据语义特征补偿向量和代表样本数据扩充描述语义特征补偿向量,包括:将所述代表样本数据语义编码特征向量、所述代表样本数据扩充描述语义编码特征向量和所述代表样本数据特征间共性特征表示向量输入基于Sigmoid函数的向量概率化单元以得到概率化代表样本数据语义编码特征向量、概率化代表样本数据扩充描述语义编码特征向量和概率化代表样本数据特征间共性特征表示向量;计算所述概率化代表样本数据语义编码特征向量相对于所述概率化代表样本数据

特征间共性特征表示向量的特定特征以得到代表样本数据特定特征补偿表示向量;计算所述概率化代表样本数据扩充描述语义编码特征向量相对于所述概率化代表样本数据特征间共性特征表示向量的特定特征以得到代表样本数据扩充描述特定特征补偿表示向量;将所述代表样本数据特定特征补偿表示向量和所述代表样本数据语义编码特征向量输入细粒度补偿模块以得到所述代表样本数据语义特征补偿向量;将所述代表样本数据扩充描述特定特征补偿表示向量和所述代表样本数据扩充描述语义编码特征向量输入所述细粒度补偿模块以得到所述代表样本数据扩充描述语义特征补偿向量。

[0013] 在上述大模型数据智能标注方法中,计算所述概率化代表样本数据语义编码特征向量相对于所述概率化代表样本数据特征间共性特征表示向量的特定特征以得到代表样本数据特定特征补偿表示向量,包括:计算所述概率化代表样本数据语义编码特征向量和所述概率化代表样本数据特征间共性特征表示向量之间的点除向量,并计算所述点除向量中各个特征值的绝对值的以二为底的对数值以得到代表样本数据特定特征向量;计算所述代表样本数据特定特征向量与所述概率化代表样本数据语义编码特征向量之间的点乘向量,并计算以e为底,以所述点乘向量的各个特征值为指数的指数函数值以得到指数化代表样本数据特定特征表示向量;将所述指数化代表样本数据特定特征表示向量输入softmax函数进行归一化处理以得到所述代表样本数据特定特征补偿表示向量。

[0014] 在上述大模型数据智能标注方法中,将所述代表样本数据特定特征补偿表示向量和所述代表样本数据语义编码特征向量输入细粒度补偿模块以得到所述代表样本数据语义特征补偿向量,包括:计算所述代表样本数据特定特征补偿表示向量和所述代表样本数据语义编码特征向量之间的哈达玛积以得到所述代表样本数据语义特征补偿向量。

[0015] 在上述大模型数据智能标注方法中,基于所述代表样本数据语义交互补偿融合表示向量,生成标注结果,包括:将所述代表样本数据语义交互补偿融合表示向量输入基于分类器的智能标注模块以得到所述标注结果。

[0016] 根据本申请的另一方面,提供了一种大模型数据智能标注系统,其包括:

数据集获取模块,用于获取未标注文本数据集;

置信度评估模块,用于对所述未标注文本数据集中的各个未标注文本数据进行置信度评估以确定代表样本数据;

语料扩充模块,用于对所述代表样本数据进行语料扩充以得到代表样本数据扩充描述;

语义编码模块,用于分别对所述代表样本数据和所述代表样本数据扩充描述进行语义编码以得到代表样本数据语义编码特征向量和代表样本数据扩充描述语义编码特征向量;

特征交互融合模块,用于对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行补偿式交互融合以得到代表样本数据语义交互补偿融合表示向量;

标注结果生成模块,用于基于所述代表样本数据语义交互补偿融合表示向量,生成标注结果。

[0017] 与现有技术相比,本申请提供的大模型数据智能标注方法及系统,其采用基于深度学习的自然语言处理技术对未标注文本数据集中的各个未标注文本数据进行置信度评

估,选择最小置信度对应的文本数据作为代表样本数据,并对所述代表样本数据进行语料扩充,进而,通过对所述代表样本数据和语料扩充后的代表样本数据进行语义特征提取和补偿式交互融合,以充分利用两者之间的共有信息和独特信息,从而实现对所述代表样本数据的全面语义理解和智能标注。通过这种方式,可以显著提高数据标注的效率和准确性,同时大幅度减少人工干预的需求,降低标注成本。

附图说明

[0018] 通过结合附图对本申请实施例进行更详细的描述,本申请的上述以及其他目的、特征和优势将变得更加明显。附图用来提供对本申请实施例的进一步理解,并且构成说明书的一部分,与本申请实施例一起用于解释本申请,并不构成对本申请的限制。在附图中,相同的参考标号通常代表相同部件或步骤。

[0019] 图1为根据本申请实施例的大模型数据智能标注方法的流程图。

[0020] 图2为根据本申请实施例的大模型数据智能标注方法的数据流动示意图。

[0021] 图3为根据本申请实施例的大模型数据智能标注方法的子步骤S2的流程图。

[0022] 图4为根据本申请实施例的大模型数据智能标注方法的子步骤S5的流程图。

[0023] 图5为根据本申请实施例的大模型数据智能标注方法的子步骤S52的流程图。

[0024] 图6为根据本申请实施例的大模型数据智能标注方法的子步骤S521的流程图。

[0025] 图7为根据本申请实施例的大模型数据智能标注系统的框图。

具体实施方式

[0026] 如本申请和权利要求书中所示,除非上下文明确提示例外情形,“一”、“一个”、“一种”和/或“该”等词并非特指单数,也可包括复数。一般说来,术语“包括”与“包含”仅提示包括已明确标识的步骤和元素,而这些步骤和元素不构成一个排它性的罗列,方法或者设备也可能包含其他的步骤或元素。

[0027] 虽然本申请对根据本申请的实施例的系统中的某些模块做出了各种引用,然而,任何数量的不同模块可以被使用并运行在用户终端和/或服务器上。所述模块仅是说明性的,并且所述系统和方法的不同方面可以使用不同模块。

[0028] 本申请中使用了流程图用来说明根据本申请的实施例的系统所执行的操作。应当理解的是,前面或下面操作不一定按照顺序来精确地执行。相反,根据需要,可以按照倒序或同时处理各种步骤。同时,也可以将其他操作添加到这些过程中,或从这些过程移除某一步或数步操作。

[0029] 下面,将参考附图详细地描述根据本申请的示例实施例。显然,所描述的实施例仅仅是本申请的一部分实施例,而不是本申请的全部实施例,应理解,本申请不受这里描述的示例实施例的限制。

[0030] 针对上述背景技术中所述的技术问题,本申请提出了一种优化的大模型数据智能标注方法,其采用基于深度学习的自然语言处理技术对未标注文本数据集中的各个未标注文本数据进行置信度评估,选择最小置信度对应的文本数据作为代表样本数据,并对所述代表样本数据进行语料扩充,进而,通过对所述代表样本数据和语料扩充后的代表样本数据进行语义特征提取和补偿式交互融合,以充分利用两者之间的共有信息和独特信息,从

而实现对所述代表样本数据的全面语义理解和智能标注。通过这种方式,可以显著提高数据标注的效率和准确性,同时大幅度减少人工干预的需求,降低标注成本。

[0031] 图1为根据本申请实施例的大模型数据智能标注方法的流程图。图2为根据本申请实施例的大模型数据智能标注方法的数据流动示意图。如图1和图2所示,所述大模型数据智能标注方法,包括步骤:S1,获取未标注文本数据集;S2,对所述未标注文本数据集中的各个未标注文本数据进行置信度评估以确定代表样本数据;S3,对所述代表样本数据进行语料扩充以得到代表样本数据扩充描述;S4,分别对所述代表样本数据和所述代表样本数据扩充描述进行语义编码以得到代表样本数据语义编码特征向量和代表样本数据扩充描述语义编码特征向量;S5,对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行补偿式交互融合以得到代表样本数据语义交互补偿融合表示向量;S6,基于所述代表样本数据语义交互补偿融合表示向量,生成标注结果。

[0032] 在上述大模型数据智能标注方法中,所述步骤S1,获取未标注文本数据集。在本申请的实施例中,所述未标注文本数据集可以为任意类型的文本数据集,例如新闻文章、社交媒体帖子、学术论文、产品评论等。此外,本申请中,所有获取数据的动作都是在遵照所在地国家相应数据保护法规政策的前提下,并获得由相应装置所有者给予授权的前提下进行的。

[0033] 在上述大模型数据智能标注方法中,所述步骤S2,对所述未标注文本数据集中的各个未标注文本数据进行置信度评估以确定代表样本数据。其中,图3为根据本申请实施例的大模型数据智能标注方法的子步骤S2的流程图。如图3所示,所述步骤S2,包括步骤:S21,对所述未标注文本数据集中的各个未标注文本数据进行语义嵌入编码以得到未标注数据语义嵌入编码向量的集合;S22,将所述未标注数据语义嵌入编码向量的集合中的各个未标注数据语义嵌入编码向量输入基于分类器的置信度评估器以得到未标注样本置信度的集合;S23,选择所述未标注样本置信度的集合中最小置信度对应的未标注文本数据作为所述代表样本数据。

[0034] 具体地,所述步骤S21,对所述未标注文本数据集中的各个未标注文本数据进行语义嵌入编码以得到未标注数据语义嵌入编码向量的集合。应可以理解,为了实现对未标注文本数据的语义理解和统一表示,以便于进行后续的数据标注分析,本申请进一步对所述未标注文本数据集中的各个未标注文本数据进行语义嵌入编码,以将文本数据映射到高维语义特征空间,从而更好地理解各个未标注文本数据的内在含义,并将其转换为语义特征空间中的连续向量表示。通过这种方式,可以实现对未标注文本数据的上下文含义理解,而非仅限于表层的词汇统计分析,从而有效提高了数据标注的准确性。在本申请的实施例中,采用word2vec模型来实现对所述未标注文本数据集的语义嵌入编码。

[0035] 具体地,所述步骤S22,将所述未标注数据语义嵌入编码向量的集合中的各个未标注数据语义嵌入编码向量输入基于分类器的置信度评估器以得到未标注样本置信度的集合。应可以理解,考虑到在大规模未标注文本数据集中,对所有的文本数据进行标注不仅效率低下,而且成本较高,因此本申请提出了一种基于置信度评估的代表样本选择方法。具体地,本申请通过基于分类器的置信度评估器来计算各个未标注数据语义嵌入编码向量的置信度得分,以揭示各个未标注文本数据的语义信息丰富度和数据标注的必要性,以便于针对性的选择具有代表性的样本进行标注。

[0036] 具体地,所述步骤S23,选择所述未标注样本置信度的集合中最小置信度对应的未标注文本数据作为所述代表样本数据。应可以理解,置信度得分较低的文本数据表明其语义较为模糊,通常位于不同类别的边界处,具有较高的不确定性,模型难以做出明确的分类决定,通过选择所述未标注样本置信度的集合中最小置信度对应的未标注文本数据作为代表样本数据,可以在有限的标注资源下最大化标注的效果,提高标注的效率和质量,同时避免对模型已经较为确定的样本进行不必要的标注,从而节省标注成本和时间。

[0037] 在上述大模型数据智能标注方法中,所述步骤S3,对所述代表样本数据进行语料扩充以得到代表样本数据扩充描述。在本申请的一个具体示例中,所述步骤S3,进一步包括:将所述代表样本数据输入基于大语言模型的语料扩充模块以得到所述代表样本数据扩充描述。应可以理解,由于所述代表样本数据可能包含的信息量较少,难以直接进行分类标注。因此,为了进一步丰富其内容,本申请采用了基于大语言模型的语料扩充模块对所述代表样本数据进行语料扩充,以生成与所述代表样本数据语义相关且包含更多上下文信息和细节描述的代表样本数据扩充描述,从而增加样本数据的语义信息量和多样性。通过这种方式,可以为代表样本数据的数据标注提供额外的上下文信息,增强模型对代表样本数据的语义理解,从而提高标注的准确性。在本申请的实施例中,可以使用诸如GPT(Generative Pre-trained Transformer)模型等先进的大语言模型来执行语料扩充任务。

[0038] 在上述大模型数据智能标注方法中,所述步骤S4,分别对所述代表样本数据和所述代表样本数据扩充描述进行语义编码以得到代表样本数据语义编码特征向量和代表样本数据扩充描述语义编码特征向量。应可以理解,为了提取出所述代表样本数据和所述代表样本数据扩充描述的语义特征,本申请进一步采用预训练语言模型分别对两者进行语义编码,以捕捉到文本数据中的深层上下文语义信息,将两者转换为连续的向量表示,从而得到代表样本数据语义编码特征向量和代表样本数据扩充描述语义编码特征向量。在本申请的实施例中,使用BERT模型对所述代表样本数据和所述代表样本数据扩充描述分别进行语义编码。本领域普通技术人员应知晓,BERT模型基于双向Transformer架构设计,能够同时考虑所述代表样本数据和所述代表样本数据扩充描述中各个词汇单元的左侧和右侧上下文信息,从而更有效地理解文本结构和语义信息,为后续的分类标注任务提供更为准确的语义表示。

[0039] 在上述大模型数据智能标注方法中,所述步骤S5,对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行补偿式交互融合以得到代表样本数据语义交互补偿融合表示向量。应可以理解,考虑到所述代表样本数据扩充描述虽然与所述代表样本数据具有语义相关性,但两者可能在某些方面存在信息的差异。因此,为了在特征融合阶段更好地利用两者之间的互补信息,增强对代表样本数据的全面理解,本申请提出了一种补偿式交互融合方法,其通过挖掘所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量之间的共性特征,并以共性特征为掩码对两者进行特征补偿式交互融合,从而能够充分利用所述代表样本数据及其扩充描述之间的共有信息和独特信息,同时减少两者之间的冗余,以实现对代表样本数据的全面语义理解。

[0040] 图4为根据本申请实施例的大模型数据智能标注方法的子步骤S5的流程图。如图4所示,所述步骤S5,包括步骤:S51,提取所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量之间的共性特征以得到代表样本数据特征间共性特征

表示向量;S52,以所述代表样本数据特征间共性特征表示向量为掩码,引导所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行特征互补性增强交互融合以得到所述代表样本数据语义交互补偿融合表示向量。

[0041] 具体地,所述步骤S51,进一步包括:将所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量输入共性特征提取网络以得到所述代表样本数据特征间共性特征表示向量,其中,所述共性特征提取网络对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行点加融合后,使用基于tanh函数的神经网络层对其进行共性特征提取以得到所述代表样本数据特征间共性特征表示向量。也就是,将所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量输入共性特征提取网络,使用神经网络架构来学习两者之间的共同模式,挖掘出两者之间的内在联系,以生成代表样本数据特征间共性特征表示向量,从而为后续的特征互补性增强交互融合提供一个参考基准。

[0042] 图5为根据本申请实施例的大模型数据智能标注方法的子步骤S52的流程图。如图5所示,所述步骤S52,包括步骤:S521,基于所述代表样本数据语义编码特征向量、所述代表样本数据扩充描述语义编码特征向量相对于所述代表样本数据特征间共性特征表示向量的特定特征,对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行特征补偿调制以得到代表样本数据语义特征补偿向量和代表样本数据扩充描述语义特征补偿向量;S522,将所述代表样本数据语义特征补偿向量、所述代表样本数据扩充描述语义特征补偿向量和所述代表样本数据特征间共性特征表示向量进行级联以得到所述代表样本数据特征间共性特征表示向量。

[0043] 图6为根据本申请实施例的大模型数据智能标注方法的子步骤S521的流程图。如图6所示,所述步骤S521,包括步骤:S5211,将所述代表样本数据语义编码特征向量、所述代表样本数据扩充描述语义编码特征向量和所述代表样本数据特征间共性特征表示向量输入基于Sigmoid函数的向量概率化单元以得到概率化代表样本数据语义编码特征向量、概率化代表样本数据扩充描述语义编码特征向量和概率化代表样本数据特征间共性特征表示向量;S5212,计算所述概率化代表样本数据语义编码特征向量相对于所述概率化代表样本数据特征间共性特征表示向量的特定特征以得到代表样本数据特定特征补偿表示向量;S5213,计算所述概率化代表样本数据扩充描述语义编码特征向量相对于所述概率化代表样本数据特征间共性特征表示向量的特定特征以得到代表样本数据扩充描述特定特征补偿表示向量;S5214,将所述代表样本数据特定特征补偿表示向量和所述代表样本数据语义编码特征向量输入细粒度补偿模块以得到所述代表样本数据语义特征补偿向量;S5215,将所述代表样本数据扩充描述特定特征补偿表示向量和所述代表样本数据扩充描述语义编码特征向量输入所述细粒度补偿模块以得到所述代表样本数据扩充描述语义特征补偿向量。

[0044] 在本申请的一个具体示例中,计算所述概率化代表样本数据语义编码特征向量相对于所述概率化代表样本数据特征间共性特征表示向量的特定特征以得到代表样本数据特定特征补偿表示向量,包括:计算所述概率化代表样本数据语义编码特征向量和所述概率化代表样本数据特征间共性特征表示向量之间的点除向量,并计算所述点除向量中各个特征值的绝对值的以二为底的对数值以得到代表样本数据特定特征向量;计算所述代表样

本数据特定特征向量与所述概率化代表样本数据语义编码特征向量之间的点乘向量,并计算以e为底,以所述点乘向量的各个特征值为指数的指数函数值以得到指数化代表样本数据特定特征表示向量;将所述指数化代表样本数据特定特征表示向量输入softmax函数进行归一化处理以得到所述代表样本数据特定特征补偿表示向量。

[0045] 在本申请的一个具体示例中,将所述代表样本数据特定特征补偿表示向量和所述代表样本数据语义编码特征向量输入细粒度补偿模块以得到所述代表样本数据语义特征补偿向量,包括:计算所述代表样本数据特定特征补偿表示向量和所述代表样本数据语义编码特征向量之间的哈达玛积以得到所述代表样本数据语义特征补偿向量。

[0046] 也就是,首先,将所述代表样本数据语义编码特征向量、所述代表样本数据扩充描述语义编码特征向量以及生成的代表样本数据特征间共性特征表示向量输入基于Sigmoid函数的向量概率化单元进行归一化处理,以确保特征向量的值域在0到1之间,从而更好地衡量特征的重要程度。然后,分别计算概率化代表样本数据语义编码特征向量、概率化代表样本数据扩充描述语义编码特征向量相对于概率化代表样本数据特征间共性特征表示向量的特定特征,通过学习原始特征和共性特征之间的特征差异,以挖掘出在共性特征之外,所述代表样本数据和所述代表样本数据扩充描述独有的语义信息,从而生成代表样本数据特定特征补偿表示向量和代表样本数据扩充描述特定特征补偿表示向量。随后,以所述代表样本数据特定特征补偿表示向量和所述代表样本数据扩充描述特定特征补偿表示向量作为权重,通过按位置点乘的方式对原始的代表样本数据语义编码特征向量和代表样本数据扩充描述语义编码特征向量进行补偿强化,以确保在融合过程中,尽可能多地保留代表样本数据和代表样本数据扩充描述的独特信息。

[0047] 更为具体地,所述步骤S522,将所述代表样本数据语义特征补偿向量、所述代表样本数据扩充描述语义特征补偿向量和所述代表样本数据特征间共性特征表示向量进行级联以得到所述代表样本数据特征间共性特征表示向量。也就是,将经过补偿强化的代表样本数据语义编码特征向量和代表样本数据扩充描述语义编码特征向量与所述代表样本数据特征间共性特征表示向量进行融合,以综合考虑所述代表样本数据和所述代表样本数据扩充描述的共性信息以及各自特征源中的独特信息,生成代表样本数据语义交互补偿融合表示向量,从而实现对代表样本数据语义信息的全面捕捉。

[0048] 相应地,所述步骤S5,包括:以如下特征补偿交互公式对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行处理以得到代表样本数据语义交互补偿融合表示向量,其中,所述特征补偿交互公式为:

$$V_j = tanh(W_{VT}(V_1 \oplus V_2) + b_{VT})$$

[0049] $$V_a = sigmoid(V_1)$$

[0050] $$V_b = sigmoid(V_2)$$

[0051] $$V_c = sigmoid(V_j)$$

[0052] $$v_a{}' = softmax\left[exp\left(V_{a_i} log\left|\frac{V_{a_i}}{V_{c_i}}\right|\right)\right]$$

[0053] $$v_b{}' = softmax\left[exp\left(V_{b_i} log\left|\frac{V_{b_i}}{V_{c_i}}\right|\right)\right]$$

[0054] $$v_{1a} = V_1 \odot v_a{}'$$

[0055] $$v_{2b} = V_2 \odot v_b{}'$$

[0056] $$V = concat[V_J; v_{1a}; v_{2b}]$$

[0057] 其中，V_1表示所述代表样本数据语义编码特征向量，V_2表示所述代表样本数据扩充描述语义编码特征向量，V_J表示代表样本数据特征间共性特征表示向量，W_{VT}和b_{VT}分别表示权重参数矩阵和偏置项，$tanh$表示双曲正切函数，$sigmoid$表示sigmoid函数，$\oplus$表示按位置点加，V_a表示概率化代表样本数据语义编码特征向量，V_b表示概率化代表样本数据扩充描述语义编码特征向量，V_c表示概率化代表样本数据特征间共性特征表示向量，V_{a_i}、V_{b_i}和V_{c_i}分别表示所述概率化代表样本数据语义编码特征向量的第i个特征值、所述概率化代表样本数据扩充描述语义编码特征向量的第i个特征值和所述概率化代表样本数据特征间共性特征表示向量的第i个特征值，$log(\cdot)$表示以2为底的对数函数，$exp(\cdot)$表示自然指数函数，$softmax$表示归一化指数函数，$v_a{}'$和$v_b{}'$分别表示代表样本数据特定特征补偿表示向量和代表样本数据扩充描述特定特征补偿表示向量，$\odot$表示点乘，v_{1a}和v_{2b}分别表示代表样本数据语义特征补偿向量和代表样本数据扩充描述语义特征补偿向量，$concat[\cdot;\cdot;\cdot]$表示级联操作，$V$表示所述代表样本数据语义交互补偿融合表示向量。

[0058] 在上述大模型数据智能标注方法中，所述步骤S6，基于所述代表样本数据语义交互补偿融合表示向量，生成标注结果。在本申请的一个具体示例中，所述步骤S6，进一步包括：将所述代表样本数据语义交互补偿融合表示向量输入基于分类器的智能标注模块以得到所述标注结果。具体而言，分类器通过对所述代表样本数据语义交互补偿融合表示向量进行多层感知处理，来学习其中所蕴含的代表样本数据的深层次语义信息，并结合其在训练过程中学习到的特征分类映射规则，将所述代表样本数据语义交互补偿融合表示向量映射到相应的类别标签上，从而实现对所述代表样本数据的准确语义标注。

[0059] 这里，在所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量分别表示代表样本数据的文本语义编码特征和代表样本数据扩充描述的文本语义编码特征的情况下，在进行以共性特征为掩码的特征间补偿式交互时，所述代表样本数据语义交互补偿融合表示向量的共性特征掩码交互补偿表示也会相对于源模态数据各自的文本内容分布场域选择性状态空间而具有交互特征实例判决化缺失，从而影响其通过基于分类器的智能标注模块得到的标注结果的准确性。

[0060] 优选地，将所述代表样本数据语义交互补偿融合表示向量输入基于分类器的智能

标注模块以得到标注结果包括：

计算所述代表样本数据语义交互补偿融合表示向量的每对特征值之间的距离，例如L2距离，并取所述距离的平方根以获得代表样本数据语义交互补偿融合距离表示矩阵，即

$$D_{i,j} = \sqrt{d(v_i, v_j)}, v_i, v_j \in V$$

[0061] 其中，V表示所述代表样本数据语义交互补偿融合表示向量，v_i和v_j表示所述代表样本数据语义交互补偿融合表示向量的第i和第j特征值，$d(v_i, v_j)$表示所述代表样本数据语义交互补偿融合表示向量的第i和第j特征值之间的距离，$D_{i,j}$表示代表样本数据语义交互补偿融合距离表示矩阵中(i,j)位置的特征值；

获得作为行向量的所述代表样本数据语义交互补偿融合表示向量的代表样本数据语义交互补偿融合自关联矩阵，即$D' = V^T \otimes V$，其中，V表示所述代表样本数据语义交互补偿融合表示向量，T表示向量的转置，$\otimes$表示矩阵乘法，D'表示代表样本数据语义交互补偿融合自关联矩阵；

将所述代表样本数据语义交互补偿融合表示向量与所述代表样本数据语义交互补偿融合距离表示矩阵进行矩阵相乘以获得代表样本数据语义交互补偿融合一级映射向量，即$V' = V \otimes D$，其中，V表示所述代表样本数据语义交互补偿融合表示向量，D表示代表样本数据语义交互补偿融合距离表示矩阵，$\otimes$表示矩阵乘法，V'表示代表样本数据语义交互补偿融合一级映射向量；

将所述代表样本数据语义交互补偿融合一级映射向量与所述代表样本数据语义交互补偿融合距离表示矩阵和所述代表样本数据语义交互补偿融合自关联矩阵的矩阵乘积进行矩阵相乘以获得代表样本数据语义交互补偿融合多级映射向量$V'' = V' \otimes (D \otimes D')$，其中，$V'$表示代表样本数据语义交互补偿融合一级映射向量，$D$表示代表样本数据语义交互补偿融合距离表示矩阵，$\otimes$表示矩阵乘法，$D'$表示代表样本数据语义交互补偿融合自关联矩阵，$V''$表示代表样本数据语义交互补偿融合多级映射向量；

将所述代表样本数据语义交互补偿融合多级映射向量与由所述代表样本数据语义交互补偿融合自关联矩阵的本征值组成的代表样本数据语义交互补偿融合关联本征向量进行点加以获得优化的代表样本数据语义交互补偿融合表示向量，其中所述本征值不足的情况下插值或补零；

将所述优化的代表样本数据语义交互补偿融合表示向量输入基于分类器的智能标注模块以得到标注结果。

[0062] 相应地，通过基于所述代表样本数据语义交互补偿融合表示向量的相似性距离表

示矩阵的线性目标映射表示,来对所述代表样本数据语义交互补偿融合表示向量的自关联的完全相似性实例化进行基于多级分布层次的二次目标映射表示,并通过关联化融合内核偏置来补偿关联不匹配负影响因子,以提升所述代表样本数据语义交互补偿融合表示向量在相似性限制下的特征值实例判决化程度,即特征值作为实例对于分类回归判决的显著性程度,提升所述代表样本数据语义交互补偿融合表示向量通过基于分类器的智能标注模块得到的标注结果的准确性。

[0063] 综上所述,基于本申请实施例的大模型数据智能标注方法被阐明,其采用基于深度学习的自然语言处理技术对未标注文本数据集中的各个未标注文本数据进行置信度评估,选择最小置信度对应的文本数据作为代表样本数据,并对所述代表样本数据进行语料扩充,进而,通过对所述代表样本数据和语料扩充后的代表样本数据进行语义特征提取和补偿式交互融合,以充分利用两者之间的共有信息和独特信息,从而实现对所述代表样本数据的全面语义理解和智能标注。通过这种方式,可以显著提高数据标注的效率和准确性,同时大幅度减少人工干预的需求,降低标注成本。

[0064] 进一步地,还提供一种大模型数据智能标注系统。

[0065] 图7为根据本申请实施例的大模型数据智能标注系统的框图。如图7所示,根据本申请实施例的大模型数据智能标注系统100,包括:数据集获取模块110,用于获取未标注文本数据集;置信度评估模块120,用于对所述未标注文本数据集中的各个未标注文本数据进行置信度评估以确定代表样本数据;语料扩充模块130,用于对所述代表样本数据进行语料扩充以得到代表样本数据扩充描述;语义编码模块140,用于分别对所述代表样本数据和所述代表样本数据扩充描述进行语义编码以得到代表样本数据语义编码特征向量和代表样本数据扩充描述语义编码特征向量;特征交互融合模块150,用于对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行补偿式交互融合以得到代表样本数据语义交互补偿融合表示向量;标注结果生成模块160,用于基于所述代表样本数据语义交互补偿融合表示向量,生成标注结果。

[0066] 这里,本领域技术人员可以理解,上述大模型数据智能标注系统中的各个模块的具体操作已经在上面参考图1到图6的大模型数据智能标注方法的描述中得到了详细介绍,并因此,将省略其重复描述。

[0067] 以上结合具体实施例描述了本发明的基本原理,但是,需要指出的是,在本发明中提及的优点、优势、效果等仅是示例而非限制,不能认为这些优点、优势、效果等是本发明的各个实施例必须具备的。另外,上述实施例的具体细节仅是为了示例的作用和便于理解的作用,而非限制,上述细节并不限制本发明为必须采用上述具体的细节来实现。

[0068] 在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中没有详述或记载的部分,可以参见其它实施例的相关描述。在本发明所提供的几个实施例中,应该理解到,所揭露的系统和方法,可以通过其它的方式实现。例如,以上所描述的系统实施例仅仅是示意性的,例如,所述单元划分,仅仅为一种逻辑功能划分,实际实现时可以有另外的划分方式。所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0069] 对于本领域技术人员而言,显然本发明不限于上述示范性实施例的细节,而且在不背离本发明的精神或基本特征的情况下,能够以其他的具体形式实现本发明。因此,无论从哪一点来看,均应将实施例看作是示范性的,而且是非限制性的,本发明的范围由所附权利要求而不是上述说明限定,因此旨在将落在权利要求的等同要件的含义和范围内的所有变化涵括在本发明内。不应将权利要求中的任何附关联图标记视为限制所涉及的权利要求。

[0070] 此外,显然“包括”一词不排除其他单元或步骤,单数不排除复数。系统权利要求中陈述的多个单元也可以由一个单元通过软件或者硬件来实现。

[0071] 最后应说明的是,为了例示和描述的目的已经给出了以上描述。此外,以上实施例仅用以说明本发明的技术方案而非限制,尽管参照较佳实施例对本发明进行了详细说明,本领域的普通技术人员应当理解,可以对本发明的技术方案进行修改或等同替换,而不脱离本发明技术方案的精神和范围。

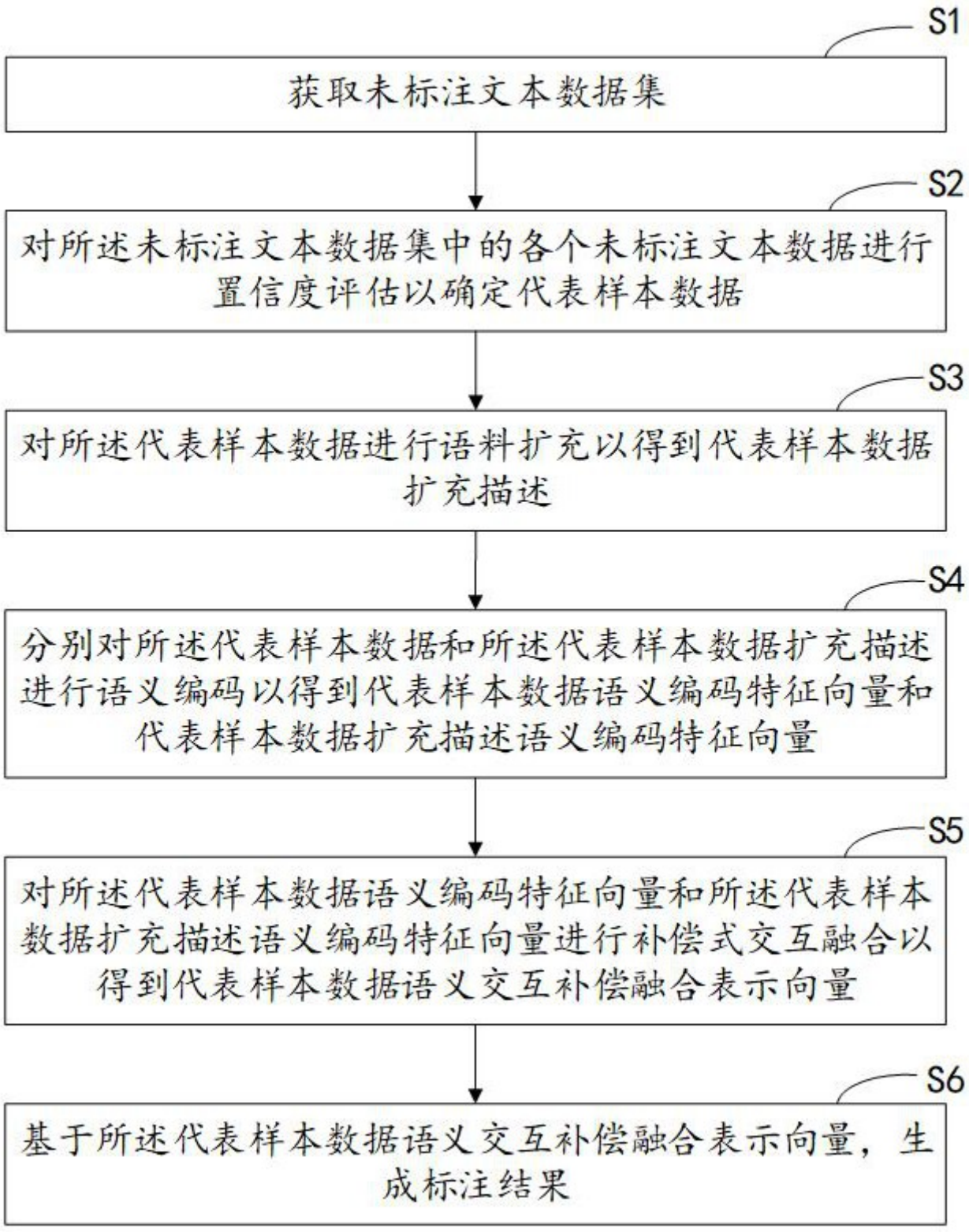
S1
获取未标注文本数据集
S2
对所述未标注文本数据集中的各个未标注文本数据进行置信度评估以确定代表样本数据
S3
对所述代表样本数据进行语料扩充以得到代表样本数据扩充描述
S4
分别对所述代表样本数据和所述代表样本数据扩充描述进行语义编码以得到代表样本数据语义编码特征向量和代表样本数据扩充描述语义编码特征向量
S5
对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行补偿式交互融合以得到代表样本数据语义交互补偿融合表示向量
S6
基于所述代表样本数据语义交互补偿融合表示向量，生成标注结果

图 1

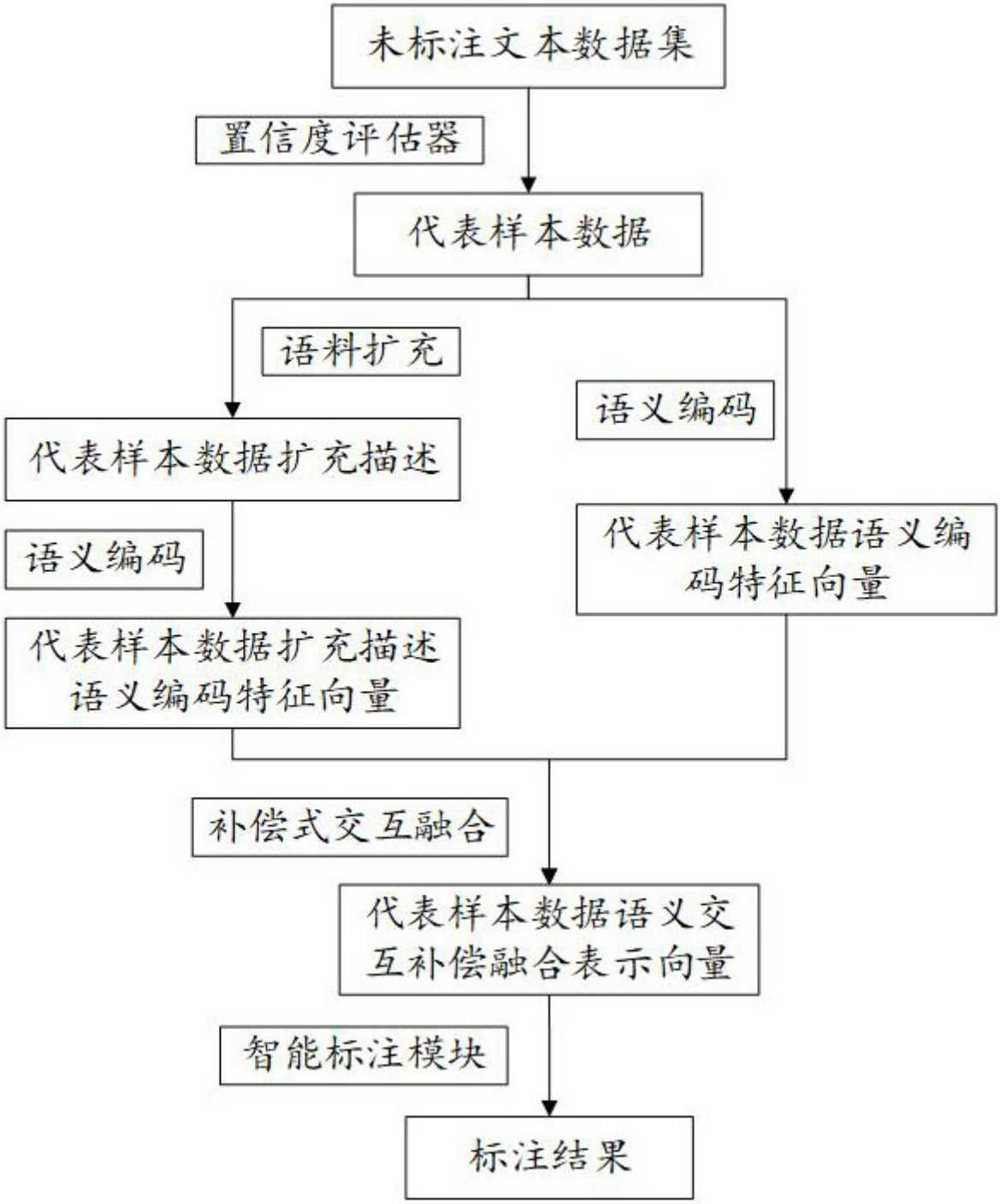
未标注文本数据集
置信度评估器
代表样本数据
语料扩充
语义编码
代表样本数据扩充描述
代表样本数据语义编码特征向量
语义编码
代表样本数据扩充描述语义编码特征向量
补偿式交互融合
代表样本数据语义交互补偿融合表示向量
智能标注模块
标注结果

图 2

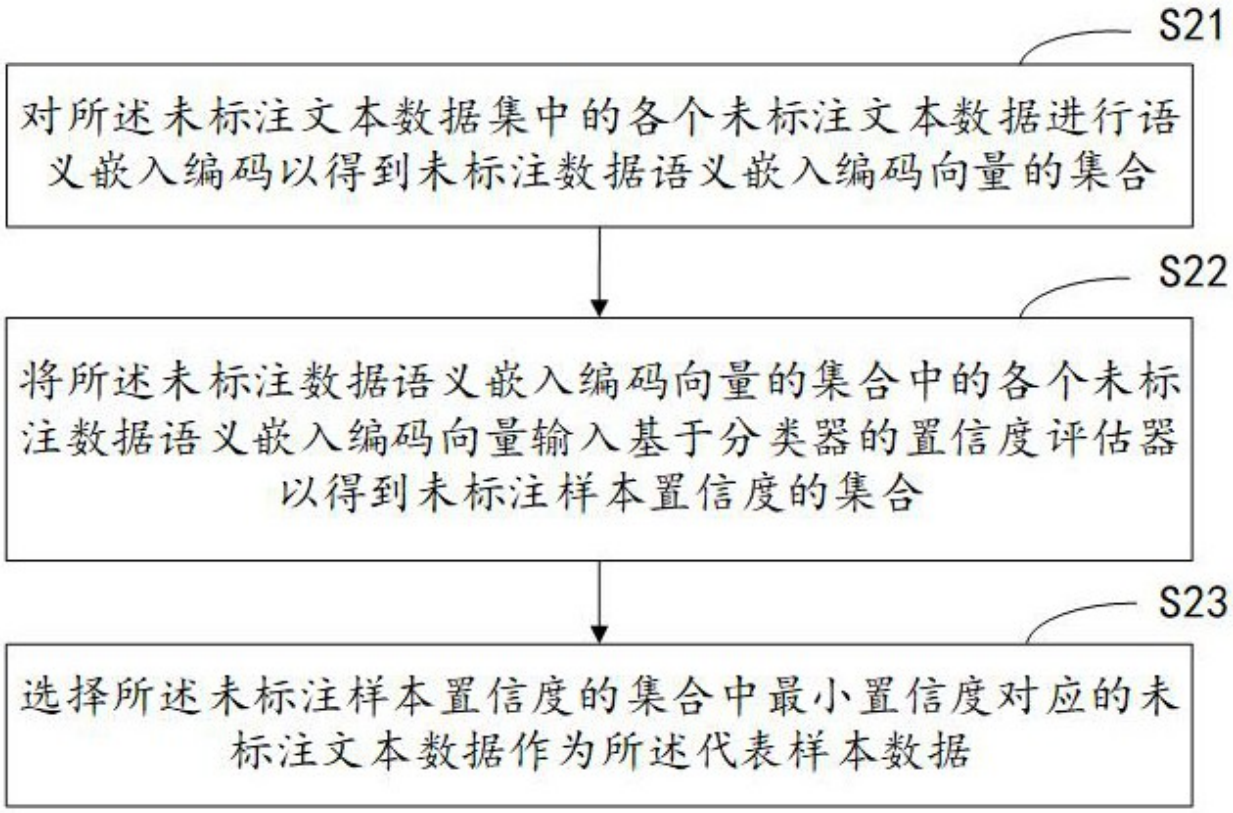
S21
对所述未标注文本数据集中的各个未标注文本数据进行语义嵌入编码以得到未标注数据语义嵌入编码向量的集合
S22
将所述未标注数据语义嵌入编码向量的集合中的各个未标注数据语义嵌入编码向量输入基于分类器的置信度评估器以得到未标注样本置信度的集合
S23
选择所述未标注样本置信度的集合中最小置信度对应的未标注文本数据作为所述代表样本数据

图 3

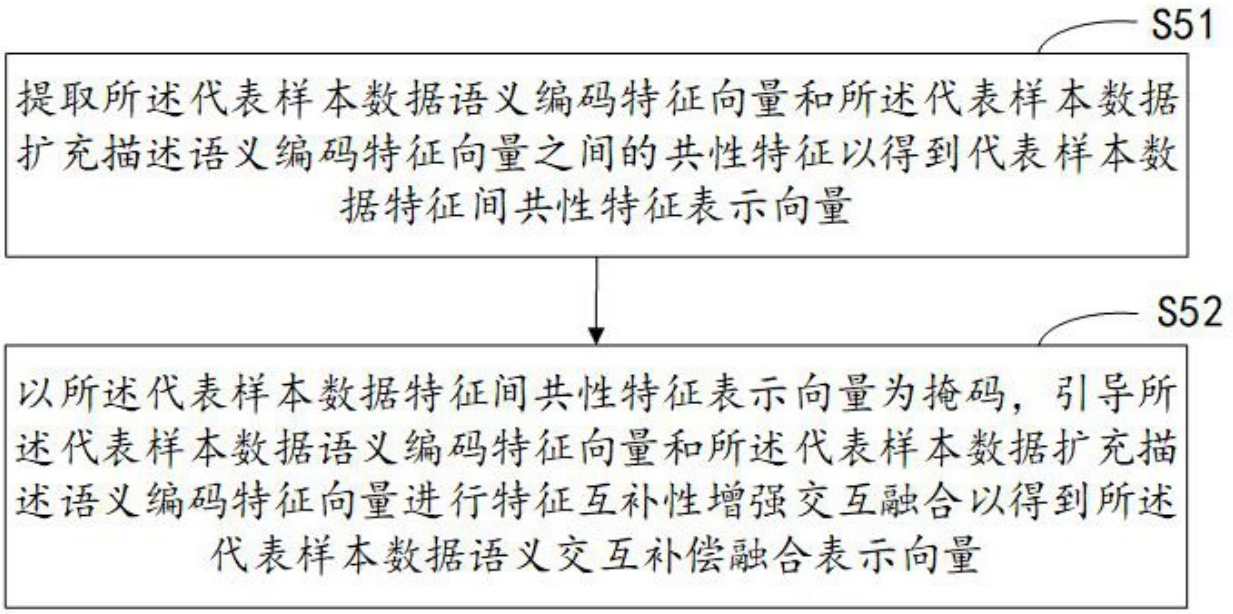
S51
提取所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量之间的共性特征以得到代表样本数据特征间共性特征表示向量
S52
以所述代表样本数据特征间共性特征表示向量为掩码，引导所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行特征互补性增强交互融合以得到所述代表样本数据语义交互补偿融合表示向量

图 4

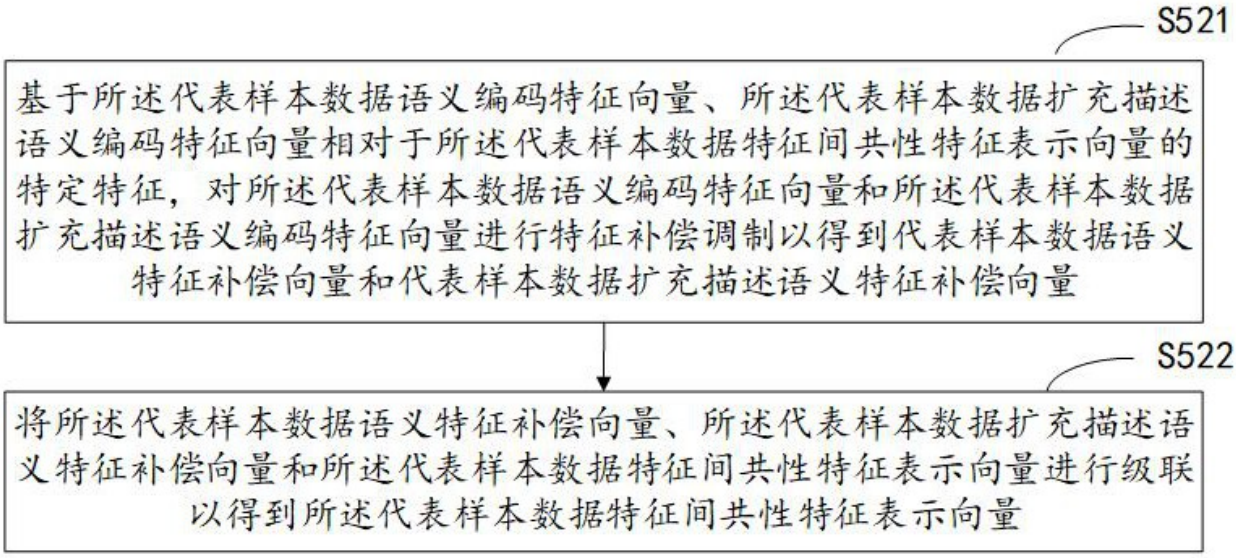
S521
基于所述代表样本数据语义编码特征向量、所述代表样本数据扩充描述语义编码特征向量相对于所述代表样本数据特征间共性特征表示向量的特定特征，对所述代表样本数据语义编码特征向量和所述代表样本数据扩充描述语义编码特征向量进行特征补偿调制以得到代表样本数据语义特征补偿向量和代表样本数据扩充描述语义特征补偿向量
S522
将所述代表样本数据语义特征补偿向量、所述代表样本数据扩充描述语义特征补偿向量和所述代表样本数据特征间共性特征表示向量进行级联以得到所述代表样本数据特征间共性特征表示向量

图 5

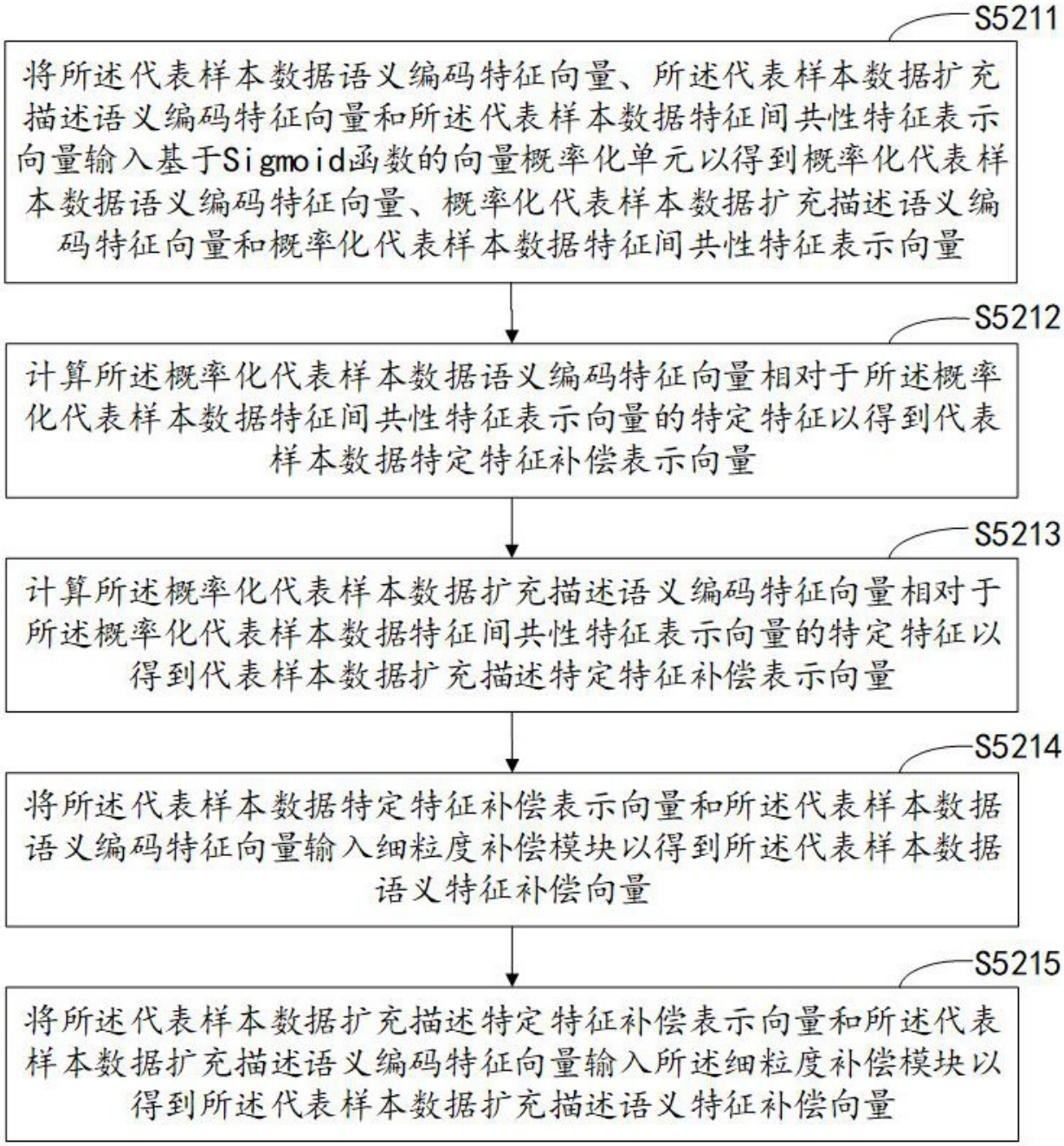
S5211
将所述代表样本数据语义编码特征向量、所述代表样本数据扩充描述语义编码特征向量和所述代表样本数据特征间共性特征表示向量输入基于Sigmoid函数的向量概率化单元以得到概率化代表样本数据语义编码特征向量、概率化代表样本数据扩充描述语义编码特征向量和概率化代表样本数据特征间共性特征表示向量
S5212
计算所述概率化代表样本数据语义编码特征向量相对于所述概率化代表样本数据特征间共性特征表示向量的特定特征以得到代表样本数据特定特征补偿表示向量
S5213
计算所述概率化代表样本数据扩充描述语义编码特征向量相对于所述概率化代表样本数据特征间共性特征表示向量的特定特征以得到代表样本数据扩充描述特定特征补偿表示向量
S5214
将所述代表样本数据特定特征补偿表示向量和所述代表样本数据语义编码特征向量输入细粒度补偿模块以得到所述代表样本数据语义特征补偿向量
S5215
将所述代表样本数据扩充描述特定特征补偿表示向量和所述代表样本数据扩充描述语义编码特征向量输入所述细粒度补偿模块以得到所述代表样本数据扩充描述语义特征补偿向量

图 6

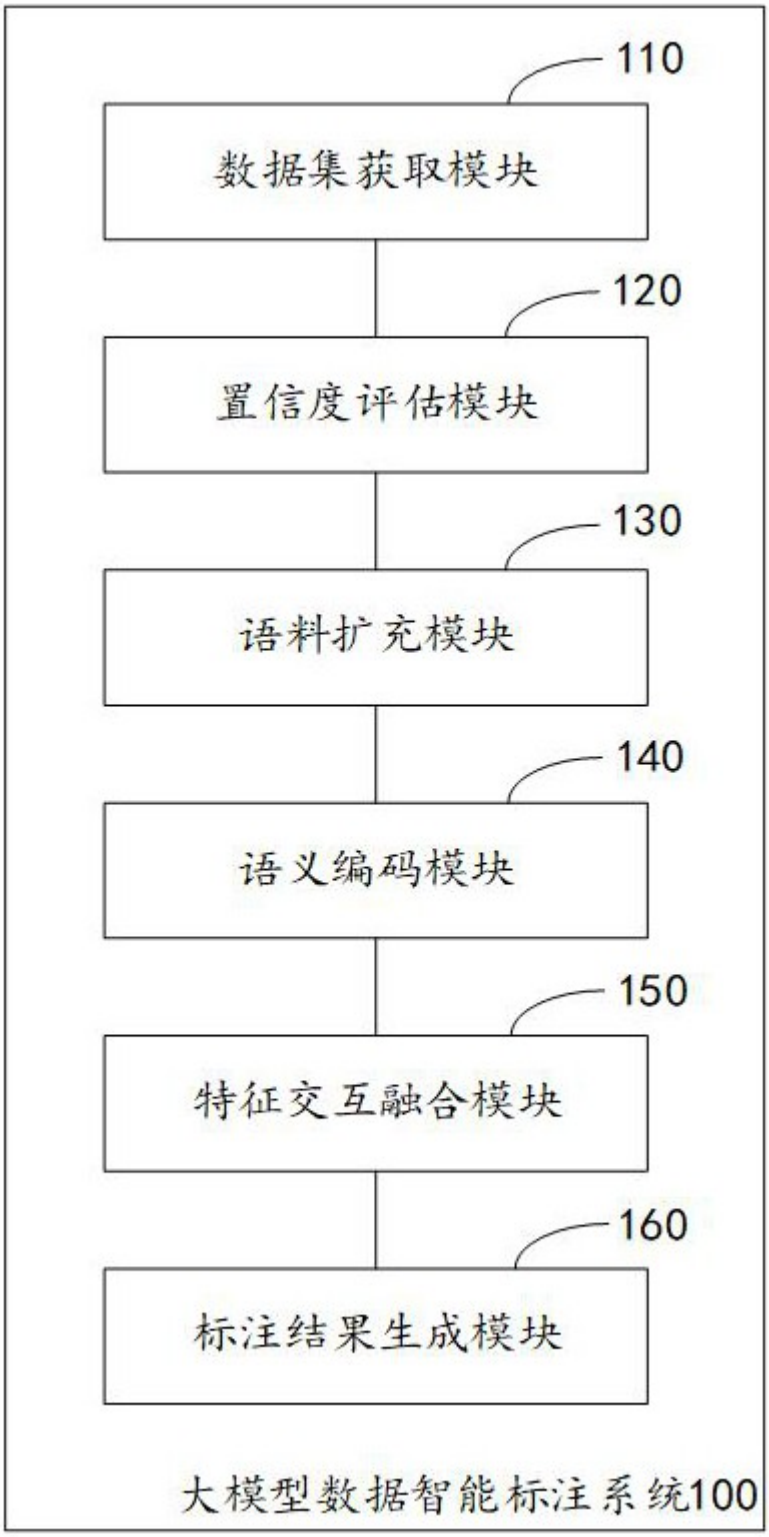
110
数据集获取模块
120
置信度评估模块
130
语料扩充模块
140
语义编码模块
150
特征交互融合模块
160
标注结果生成模块
大模型数据智能标注系统100

图 7