



(12) 发明专利申请

(10) 申请公布号 CN 114817455 A

(43) 申请公布日 2022. 07. 29

(21) 申请号 202210229151.9

(22) 申请日 2022.03.08

(71) 申请人 招商银行股份有限公司

地址 518000 广东省深圳市福田区深南大道7088招商银行大厦

(72) 发明人 赵高枫 文俊杰 李金龙

(74) 专利代理机构 深圳市世纪恒程知识产权代理有限公司 44287

专利代理师 李幸芳

(51) Int. Cl.

G06F 16/33 (2019.01)

G06F 16/332 (2019.01)

G06F 16/35 (2019.01)

G06K 9/62 (2022.01)

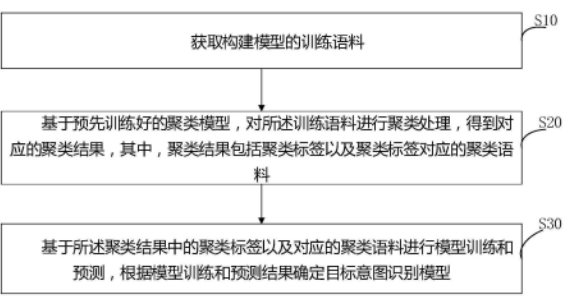
权利要求书2页 说明书11页 附图6页

(54) 发明名称

模型构建方法、装置、设备与介质

(57) 摘要

本发明涉及人工智能技术领域,公开了一种模型构建方法、装置、设备与介质。本发明通过获取构建模型的训练语料;基于预先训练好的聚类模型,对所述训练语料进行聚类处理,得到对应的聚类结果,其中,聚类结果包括聚类标签以及聚类标签对应的聚类语料;基于所述聚类结果中的聚类标签以及对应的聚类语料进行模型训练和预测,根据模型训练和预测结果确定目标意图识别模型。上述自动生成目标意图识别模型的方法,减少了在熟悉业务点以及数据标注过程中投入的时间,加速梳理业务点以及标注业务语料,提高了构建目标意图识别模型的效率,降低了人力成本。



1. 一种模型构建方法,其特征在于,所述模型构建方法包括如下步骤:

获取构建模型的训练语料;

基于预先训练好的聚类模型,对所述训练语料进行聚类处理,得到对应的聚类结果,其中,聚类结果包括聚类标签以及聚类标签对应的聚类语料;

基于所述聚类结果中的聚类标签以及对应的聚类语料进行模型训练和预测,根据模型训练和预测结果确定目标意图识别模型。

2. 如权利要求1所述的模型构建方法,其特征在于,所述基于预先训练好的聚类模型,对所述训练语料进行聚类处理,得到对应的聚类结果的步骤包括:

将所述训练语料进行依次乱序、分割,得到聚类样本语料;

基于层次凝聚聚类算法HAC,对所述聚类样本语料进行聚类处理,得到聚类标签以及聚类标签对应的聚类语料。

3. 如权利要求2所述的模型构建方法,其特征在于,所述基于层次凝聚聚类算法HAC,对所述聚类样本语料进行聚类处理,得到聚类标签以及聚类标签对应的聚类语料的步骤包括:

将所述聚类样本语料进行分类,并将分类过后同种类的聚类样本语料划分为一个簇,得到不同种类的簇对应的聚类标签;

基于所述簇对应的聚类标签,确定不同种类的聚类标签对应的簇内语料;

若所述簇内语料的数量大于预设阈值N1,则所述簇内语料以所述簇为聚类标签,聚类标签对应的簇内语料为对应的聚类语料;

若所述簇内语料的数量不大于预设阈值N1,则所述簇内语料以other为聚类标签,聚类标签对应的簇内语料为对应的聚类语料。

4. 如权利要求1所述的模型构建方法,其特征在于,所述基于所述聚类结果中的聚类标签以及对应的聚类语料进行模型训练和预测,根据模型训练和预测结果确定目标意图识别模型的步骤包括:

将聚类结果中的聚类语料,根据聚类标签划分为训练语料和预测语料;

基于所述训练语料进行模型训练,得到训练好的初始分类模型;

将所述预测语料输入所述训练好的初始分类模型进行预测,得到预测分数值;

基于所述聚类结果中的聚类标签与所述预测分数值,确定所述聚类结果的精确召回率PRF值;

基于所述PRF值,确定对应的目标意图识别模型。

5. 如权利要求4所述的模型构建方法,其特征在于,所述基于所述PRF值,确定对应的目标意图识别模型的步骤包括:

判断所述PRF值是否达到预设阈值;

若所述PRF值达到预设阈值,则所述聚类结果合理,输出所述初始分类模型作为目标意图识别模型;

若所述PRF值未达到预设阈值,则所述聚类结果不合理,对所述聚类结果进行分类调整,得到分类调整后的聚类结果;

将所述分类调整后的聚类结果作为当前的聚类结果,并返回执行步骤:

将聚类结果中的聚类语料,根据聚类标签划分为训练语料和预测语料;

基于所述训练语料进行模型训练,得到训练好的初始分类模型;  
将所述预测语料输入所述训练好的初始分类模型进行预测,得到预测分数值;  
基于所述聚类结果中的聚类标签与所述预测分数值,确定所述聚类结果的精确召回率PRF值;

直到所述PRF值达到预设阈值,所述聚类结果合理,输出所述初始分类模型作为目标意图识别模型。

6.如权利要求5所述的模型构建方法,其特征在于,所述聚类结果包括聚类标签为other的other聚类语料和聚类标签为非other的非other聚类语料,

所述对所述聚类结果进行分类调整,得到分类调整后的聚类结果的步骤包括:

调整所述other聚类语料和非other聚类语料,获取调整后的聚类标签对应的聚类语料;

计算所述调整后的聚类标签对应的聚类语料的混淆度;

当所述混淆度大于预设阈值T2,则合并调整分类前后的非other的聚类语料作为当前的聚类结果的非other聚类语料,其他语料则作为聚类标签为other的other聚类语料。

7.如权利要求6所述的模型构建方法,其特征在于,所述调整所述other聚类语料和非other聚类语料,获取调整后的聚类标签对应的聚类语料的步骤包括:

获取所述非other聚类语料的预测分数值;

若所述非other聚类语料的预测分数值低于预设阈值T1,则将所述非other聚类语料的聚类标签更改为other;

当所述聚类标签为other的other聚类语料数量超过预设阈值N2,则获取到调整后other聚类语料以及调整后的非other聚类语料。

8.如权利要求1所述的模型构建方法,其特征在于,所述获取构建模型的训练语料的步骤包括:

从业务端获取原始语料;

对所述原始语料进行预处理,得到用于模型构建的训练语料;

其中,预处理的方式包括剔除停用词、全角转半角、剔除表情符号、剔除招呼用语和无意义问题、统一用标点符号以及剔除非常用标点符号中的一种或多种。

9.一种模型构建装置,其特征在于,所述模型构建装置包括:

获取模块,用于获取构建模型的训练语料;

聚类模块,用于基于预先训练好的聚类模型,对所述训练语料进行聚类处理,得到对应的聚类结果,其中,聚类结果包括聚类标签以及聚类标签对应的聚类语料;

确定模块,用于基于所述聚类结果中的聚类标签以及对应的聚类语料进行模型训练和预测,根据模型训练和预测结果确定目标意图识别模型。

10.一种模型构建设备,其特征在于,所述模型构建设备包括:存储器、处理器及存储在所述存储器上并可在所述处理器上运行的模型构建程序,所述模型构建程序被所述处理器执行时实现如权利要求1至8中任一项所述的模型构建方法的步骤。

11.一种介质,所述介质为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上存储有模型构建程序,所述模型构建程序被处理器执行时实现如权利要求1至8中任一项所述的模型构建方法的步骤。

## 模型构建方法、装置、设备与介质

### 技术领域

[0001] 本发明涉及人工智能技术领域,尤其涉及一种模型构建方法、装置、设备与介质。

### 背景技术

[0002] 随着人工智能技术的不断发展,对话系统的应用越来越广泛,比如无人客服系统,在无人客服系统中,意图识别是其重要组成部分,意图识别的一种常见算法是通过文本分类来识别用户的意图,具体的,将用户的意图划分为几个类别,在类别下配以对应的应答方案。

[0003] 在搭建一个对话系统时,文本分类通常是构建用户意图识别模型的一种最简单有效的手段,现有的用以意图识别的文本分类模型主要是基于对话系统中需承担的业务制定,制定过程往往需要运营人员去了解对话系统需承担的业务点,梳理出相关的业务点信息,然后对业务方提供的语料大量标注,标注完了之后借助分类调整工具对分类进行调整,如果语料不够充足,会考虑采用工具扩充,最后得到用以识别用户意图的文本分类模型,在这个过程中,熟悉业务点以及数据的标注耗时较多,运营人员通过梳理业务点以及标注业务语料构建文本分类模型的效率低,耗费大量人力。

### 发明内容

[0004] 本发明的主要目的在于提出一种模型构建方法、装置、设备与介质,旨在降低人工梳理、标注的人力成本,提高分类模型构建效率。

[0005] 为实现上述目的,本发明提供一种模型构建方法,所述模型构建方法包括如下步骤:

[0006] 获取构建模型的训练语料;

[0007] 基于预先训练好的聚类模型,对所述训练语料进行聚类处理,得到对应的聚类结果,其中,聚类结果包括聚类标签以及聚类标签对应的聚类语料;

[0008] 基于所述聚类结果中的聚类标签以及对应的聚类语料进行模型训练和预测,根据模型训练和预测结果确定目标意图识别模型。

[0009] 优选地,所述基于预先训练好的聚类模型,对所述训练语料进行聚类处理,得到对应的聚类结果的步骤包括:

[0010] 将所述训练语料进行依次乱序、分割,得到聚类样本语料;

[0011] 基于层次凝聚聚类算法HAC,对所述聚类样本语料进行聚类处理,得到聚类标签以及聚类标签对应的聚类语料。

[0012] 优选地,所述基于层次凝聚聚类算法HAC,对所述聚类样本语料进行聚类处理,得到聚类标签以及聚类标签对应的聚类语料的步骤包括:

[0013] 将所述聚类样本语料进行分类,并将分类过后同种类的聚类样本语料划分为一个簇,得到不同种类的簇对应的聚类标签;

[0014] 基于所述簇对应的聚类标签,确定不同种类的聚类标签对应的簇内语料;

- [0015] 若所述簇内语料的数量大于预设阈值N1,则所述簇内语料以所述簇为聚类标签,聚类标签对应的簇内语料为对应的聚类语料;
- [0016] 若所述簇内语料的数量不大于预设阈值N1,则所述簇内语料以other为聚类标签,聚类标签对应的簇内语料为对应的聚类语料。
- [0017] 优选地,所述基于所述聚类结果中的聚类标签以及对应的聚类语料进行模型训练和预测,根据模型训练和预测结果确定目标意图识别模型的步骤包括:
- [0018] 将聚类结果中的聚类语料,根据聚类标签划分为训练语料和预测语料;
- [0019] 基于所述训练语料进行模型训练,得到训练好的初始分类模型;
- [0020] 将所述预测语料输入所述训练好的初始分类模型进行预测,得到预测分数值;
- [0021] 基于所述聚类结果中的聚类标签与所述预测分数值,确定所述聚类结果的精确召回率PRF值;
- [0022] 基于所述PRF值,确定对应的目标意图识别模型。
- [0023] 优选地,所述基于所述PRF值,判断所述聚类结果是否合理;
- [0024] 判断所述PRF值是否达到预设阈值;
- [0025] 若所述PRF值达到预设阈值,则所述聚类结果合理,输出所述初始分类模型作为目标意图识别模型;
- [0026] 若所述PRF值未达到预设阈值,则所述聚类结果不合理,对所述聚类结果进行分类调整,得到分类调整后的聚类结果;
- [0027] 将所述分类调整后的聚类结果作为当前的聚类结果;
- [0028] 将聚类结果中的聚类语料,根据聚类标签划分为训练语料和预测语料;
- [0029] 基于所述训练语料进行模型训练,得到训练好的初始分类模型;
- [0030] 将所述预测语料输入所述训练好的初始分类模型进行预测,得到预测分数值;
- [0031] 基于所述聚类结果中的聚类标签与所述预测分数值,确定所述聚类结果的精确召回率PRF值;
- [0032] 直到所述PRF值达到预设阈值,所述聚类结果合理,输出所述初始分类模型作为目标意图识别模型。
- [0033] 优选地,所述聚类结果包括聚类标签为other的other聚类语料和聚类标签为非other的非other聚类语料,
- [0034] 所述对所述聚类结果进行分类调整,得到分类调整后的聚类结果的步骤包括:
- [0035] 调整所述other聚类语料和非other聚类语料,获取调整后的聚类标签对应的聚类语料;
- [0036] 计算所述调整后的聚类标签对应的聚类语料的混淆度;
- [0037] 当所述混淆度大于预设阈值T2,则合并调整分类前后的非other的聚类语料作为当前的聚类结果的非other聚类语料,其他语料则作为聚类标签为other的other聚类语料。
- [0038] 优选地,所述调整所述other聚类语料和非other聚类语料,获取调整后的聚类标签对应的聚类语料的步骤包括:
- [0039] 获取所述非other聚类语料的预测分数值;
- [0040] 若所述非other聚类语料的预测分数值低于预设阈值T1,则将所述非other聚类语料的聚类标签更改为other;

[0041] 当所述聚类标签为other的other聚类语料数量超过预设阈值N2,则获取到调整后other聚类语料以及调整后的非other聚类语料。

[0042] 优选地,所述获取构建模型的训练语料的步骤包括:

[0043] 从业务端获取原始语料;

[0044] 对所述原始语料进行预处理,得到用于模型构建的训练语料;

[0045] 其中,预处理的方式包括剔除停用词、全交转半角、剔除表情符号、剔除招呼用语和无意义问题、统一用标点符号以及剔除非常用标点符号中的一种或多种。

[0046] 此外,为实现上述目的,本发明还提供一种模型构建装置,所述模型构建装置包括:

[0047] 获取模块,用于获取构建模型的训练语料;

[0048] 聚类模块,用于基于预先训练好的聚类模型,对所述训练语料进行聚类处理,得到对应的聚类结果,其中,聚类结果包括聚类标签以及聚类标签对应的聚类语料;

[0049] 确定模块,用于基于所述聚类结果中的聚类标签以及对应的聚类语料进行模型训练和预测,根据模型训练和预测结果确定目标意图识别模型。

[0050] 优选地,获取模块还用于:

[0051] 从业务端获取原始语料;

[0052] 对所述原始语料进行预处理,得到用于模型构建的训练语料;

[0053] 其中,预处理的方式包括剔除停用词、全交转半角、剔除表情符号、剔除招呼用语和无意义问题、统一用标点符号以及剔除非常用标点符号中的一种或多种。

[0054] 优选地,聚类模块还用于:

[0055] 将所述训练语料进行依次乱序、分割,得到聚类样本语料;

[0056] 基于层次凝聚聚类算法HAC,对所述聚类样本语料进行聚类处理,得到聚类标签以及聚类标签对应的聚类语料。

[0057] 优选地,聚类模块还用于:

[0058] 将所述聚类样本语料进行分类,并将分类过后同种类的聚类样本语料划分为一个簇,得到不同种类的簇对应的聚类标签;

[0059] 基于所述簇对应的聚类标签,确定不同种类的聚类标签对应的簇内语料;

[0060] 若所述簇内语料的数量大于预设阈值N1,则所述簇内语料以所述簇为聚类标签,聚类标签对应的簇内语料为对应的聚类语料;

[0061] 若所述簇内语料的数量不大于预设阈值N1,则所述簇内语料以other为聚类标签,聚类标签对应的簇内语料为对应的聚类语料。

[0062] 优选地,确定模块还用于:

[0063] 将聚类结果中的聚类语料,根据聚类标签划分为训练语料和预测语料;

[0064] 基于所述训练语料进行模型训练,得到训练好的初始分类模型;

[0065] 将所述预测语料输入所述训练好的初始分类模型进行预测,得到预测分数值;

[0066] 基于所述聚类结果中的聚类标签与所述预测分数值,确定所述聚类结果的精确召回率PRF值;

[0067] 基于所述PRF值,确定对应的目标意图识别模型。

[0068] 优选地,确定模块还用于:

- [0069] 判断所述PRF值是否达到预设阈值；
- [0070] 若所述PRF值达到预设阈值，则所述聚类结果合理，输出所述初始分类模型作为目标意图识别模型；
- [0071] 若所述PRF值未达到预设阈值，则所述聚类结果不合理，对所述聚类结果进行分类调整，得到分类调整后的聚类结果；
- [0072] 将所述分类调整后的聚类结果作为当前的聚类结果，并执行所述步骤；
- [0073] 将分类调整后的聚类结果中的聚类语料，根据聚类标签划分为训练语料和预测语料；
- [0074] 基于所述训练语料进行模型训练，得到训练好的初始分类模型；
- [0075] 将所述预测语料输入所述训练好的初始分类模型进行预测，得到预测分数值；
- [0076] 基于所述聚类结果中的聚类标签与所述预测分数值，确定所述聚类结果的精确召回率PRF值；
- [0077] 直到所述PRF值达到预设阈值，所述聚类结果合理，输出所述初始分类模型作为目标意图识别模型。
- [0078] 优选地，确定模块还用于：
- [0079] 调整所述other聚类语料和非other聚类语料，获取调整后的聚类标签对应的聚类语料；
- [0080] 计算所述调整后的聚类标签对应的聚类语料的混淆度；
- [0081] 当所述混淆度大于预设阈值T2，则合并调整分类前后的非other的聚类语料作为当前的聚类结果的非other聚类语料，其他语料则作为聚类标签为other的other聚类语料。
- [0082] 优选地，确定模块还用于：
- [0083] 获取所述非other聚类语料的预测分数值；
- [0084] 若所述非other聚类语料的预测分数值低于预设阈值T1，则将所述非other聚类语料的聚类标签更改为other；
- [0085] 当所述聚类标签为other的other聚类语料数量超过预设阈值N2，则获取到调整后other聚类语料以及调整后的非other聚类语料。
- [0086] 此外，为实现上述目的，本发明还提供一种模型构建设备，所述模型构建设备包括：存储器、处理器及存储在所述存储器上并可在所述处理器上运行的模型构建程序，所述模型构建程序被所述处理器执行时实现如上所述的模型构建方法的步骤。
- [0087] 此外，为实现上述目的，本发明还提供一种介质，所述介质为计算机可读存储介质，所述计算机可读存储介质上存储有模型构建程序，所述模型构建程序被处理器执行时实现如上所述的模型构建方法的步骤。
- [0088] 本发明提出的模型构建方法、装置、设备及介质，通过获取构建模型的训练语料；基于预先训练好的聚类模型，对所述训练语料进行聚类处理，得到对应的聚类结果，其中，聚类结果包括聚类标签以及聚类标签对应的聚类语料；基于所述聚类结果中的聚类标签以及对应的聚类语料进行模型训练和预测，根据模型训练和预测结果确定目标意图识别模型。
- [0089] 通过对用于构建意图识别模型的训练语料进行聚类处理，获取到训练语料对应的聚类结果，上述聚类结果包括聚类语料做好分类的聚类标签以及聚类标签对应的聚类语

料,将这些包括聚类标签以及聚类标签对应的聚类语料的聚类结果进行模型训练和预测,得到聚类结果对应的PRF值,根据该PRF值确定目标意图识别模型,上述自动生成目标意图识别模型的方法,减少了在熟悉业务点以及数据标注过程中投入的时间,提高梳理业务点以及标注业务语料的效率,降低了人力成本。

## 附图说明

- [0090] 图1是本发明模型构建实施例方案涉及的硬件运行环境的设备结构示意图;
- [0091] 图2为本发明模型构建方法第一实施例的流程示意图;
- [0092] 图3为本发明模型构建方法第一实施例的具体流程示意图;
- [0093] 图4为本发明模型构建方法第二实施例的流程示意图;
- [0094] 图5为本发明模型构建方法第二实施例中步骤S22的子流程示意图;
- [0095] 图6为本发明模型构建方法第三实施例的流程示意图;
- [0096] 图7为本发明模型构建方法第四实施例的流程示意图;
- [0097] 图8为本发明模型构建方法第四实施例中步骤B3的子流程示意图;
- [0098] 图9为本发明模型构建方法第一实施例涉及到的模型构建装置的功能模块示意图。
- [0099] 本发明目的的实现、功能特点及优点将结合实施例,参照附图做进一步说明。

## 具体实施方式

- [0100] 应当理解,此处所描述的具体实施例仅仅用以解释本发明,并不用于限定本发明。
- [0101] 如图1所示,图1是本发明实施例方案涉及的硬件运行环境的设备结构示意图。
- [0102] 本发明实施例设备可以是移动终端或服务器设备。
- [0103] 如图1所示,该设备可以包括:处理器1001,例如CPU,网络接口1004,用户接口1003,存储器1005,通信总线1002。其中,通信总线1002用于实现这些组件之间的连接通信。用户接口1003可以包括显示屏(Display)、输入单元比如键盘(Keyboard),可选用户接口1003还可以包括标准的有线接口、无线接口。网络接口1004可选的可以包括标准的有线接口、无线接口(如WI-FI接口)。存储器1005可以是高速RAM存储器,也可以是稳定的存储器(non-volatile memory),例如磁盘存储器。存储器1005可选的还可以是独立于前述处理器1001的存储装置。
- [0104] 本领域技术人员可以理解,图1中示出的设备结构并不构成对设备的限定,可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件布置。
- [0105] 如图1所示,作为一种计算机存储介质的存储器1005中可以包括操作系统、网络通信模块、用户接口模块以及模型构建程序。
- [0106] 其中,操作系统是管理和控制模型构建设备与软件资源的程序,支持网络通信模块、用户接口模块、模型构建程序以及其他程序或软件的运行;网络通信模块用于管理和控制网络接口1002;用户接口模块用于管理和控制用户接口1003。
- [0107] 在图1所示的模型构建设备中,所述模型构建设备通过处理器1001调用存储器1005中存储的模型构建程序,并执行下述模型构建方法各个实施例中的操作。
- [0108] 基于上述硬件结构,提出本发明模型构建方法实施例。

[0109] 参照图2,图2为本发明模型构建方法第一实施例的流程示意图,所述方法包括:

[0110] 步骤S10,获取构建模型的训练语料;

[0111] 从业务端获取原始语料;对所述原始语料进行预处理,得到用于模型构建的训练语料;其中,预处理的方式包括剔除停用词、全交转半角、剔除表情符号、剔除招呼用语和无意义问题、统一用标点符号以及剔除非常用标点符号中的一种或多种。

[0112] 在一具体实施例中,收集来自业务场景的对话语料用于模型构建,将智能客服业务大量的用户聊天记录作为训练目标模型的原始语料,将这些原始语料进行标准化的预处理,处理方式可以是包括剔除停用词、全交转半角、剔除表情符号、剔除招呼用语和无意义问题、统一用标点符号以及剔除非常用标点符号中的一种或多种,通过对这些来自于客户端的整体原始语料进行标准化预处理,达到了减少原始语料中的噪音、提高原始语料纯净度、获取到能够用于构建模型的训练语料的效果。

[0113] 步骤S20,基于预先训练好的聚类模型,对所述训练语料进行聚类处理,得到对应的聚类结果,其中,聚类结果包括聚类标签以及聚类标签对应的聚类语料;

[0114] 对训练语料进行聚类处理得到与训练语料相对应的聚类结果的原理是:聚类处理能够运用形态学算子将临近的类似分类区域聚类并合并,在聚类、合并处理之后,生成每个类别不同的簇,而这些由聚类所生成的簇是一组数据对象的集合,这些对象与同一个簇中的对象彼此相似,与其他簇中的对象相异。

[0115] 在现有的聚类处理算法中,聚类算法大都具有处理“噪声”数据的能力,一些聚类算法对于噪音数据十分敏感,能够在对音频数据进行聚类处理之后得到对应的精确聚类结果。

[0116] 在一具体实施例中,将从业务端收集到的语料数据进行预处理,得到预处理后的训练语料,将该训练语料输入预先训练好的聚类模型,将该预处理后的训练语料打乱顺序,抽取训练语中的部分语料进行聚类,其中聚类算法可以采用层次凝聚聚类算法HAC,得到每个类别不同的簇,而这些簇又有着它们各自对应的语料,进而得到包括聚类标签以及聚类标签对应的聚类语料的聚类结果,在上述聚类结果中,若簇内语料的数量大于预设阈值N1(例:50)的簇,以簇id为聚类标签,而剩余语料以other为聚类标签。

[0117] 步骤S30,基于所述聚类结果中的聚类标签以及对应的聚类语料进行模型训练和预测,根据模型训练和预测结果确定目标意图识别模型。

[0118] 在现有的技术中,通过一个初始预训练模型对人工进行划分和标注,得到能够进行模型训练的数据集,根据该数据集初始预训练模型进行深度学习,得到一个目标的分类模型。在本实施例的构建模型方法中,通过直接获取业务端的训练语料,对这些训练语料进行聚类处理,得到包括聚类标签以及聚类标签对应的聚类语料的聚类结果,并对聚类结果进行训练和预测,根据模型训练和预测结果确定目标意图识别模型。

[0119] 在一具体的实施例中,根据上述包括聚类标签以及聚类标签对应的聚类语料的聚类结果,将聚类结果按照聚类标签平均切分成5份,例如,聚类标签(簇id)为label1的语料数量为200,则切分后每份聚类语料的聚类标签为label1的数量为40,每次取4份聚类语料对预训练分类模型进行训练,剩余1份聚类语料则进行预测,可以得到全部5份聚类语料的预测值分数值。通过聚类标签与其对应的预测分数值,可以得出聚类结果对应的PRF值,即Precision(精确率)、Recall(召回率)以及F1值(F1),并以PRF值来评估聚类结果是否合理,

如果聚类结果是合理的,则输出基于训练数据训练好的分类模型,得到对应的目标意图识别模型。

[0120] 本实施例通过将预处理后的训练语料进行聚类处理,得到具有不同特征类别的聚类标签以及聚类标签对应的聚类语料,将上述聚类标签以及聚类标签对应的聚类语料作为聚类结果进行模型训练和预测,根据模型训练和预测结果确定目标意图识别模型。通过自动构建意图识别模型,大大降低了人工标注的人力成本,减轻了运营人员的运营压力,另外自动构建模型方法能够协助运营人员理解业务,发现梳理涉及的不同业务类型,提升构建意图识别模型的效率。

[0121] 进一步地,基于本发明模型构建方法第一实施例,提出本发明模型构建方法第二实施例。

[0122] 模型构建方法的第二实施例与模型构建方法的第一实施例的区别在于,本实施例是对步骤S20,所述基于预先训练好的聚类模型,对所述训练语料进行聚类处理,得到对应的聚类结果的细化,参照图4,该步骤具体包括:

[0123] 步骤S21,将所述训练语料进行依次乱序、分割,得到聚类样本语料;

[0124] 在一具体实施例中,将上述预处理后的训练语料输入预先训练好的聚类模型,打乱该训练语料的排列顺序,抽取部分训练语料进行聚类,抽取部分训练语料进行聚类的方式具体包括:将部分训练语料同步进行聚类处理,然后逐步加入剩余训练语料,进而得到训练语料对应的聚类结果。对比将全部训练语料直接进行聚类的方式,抽取部分训练语料进行聚类的方式最终得到的分类效果更好。

[0125] 在一具体实施例中,将上述预处理后的训练语料进行乱序,并抽取该训练语料中的一部分训练语料作为聚类样本语料,比如,参照总体语料数量,可抽取总体语料数量的20%作为聚类样本语料进行聚类处理。

[0126] 步骤S22,基于层次凝聚聚类算法HAC,对所述聚类样本语料进行聚类处理,得到聚类标签以及聚类标签对应的聚类语料。

[0127] 聚类处理的方式有许多种,包括划分方法、层次方法、基于密度的方法、基于网格的方法、基于模型的方法、传递闭包法、布尔矩阵法、直接聚类法、相关性分析聚类以及基于统计的聚类方法等,而在不同方法中,也有各种不同的聚类算法,可以根据这些聚类算法得到对应的聚类结果。

[0128] 在一具体实施例中,通过基于层次凝聚聚类算法HAC,对上述聚类样本语料进行聚类处理,得到聚类样本语料属于不同种类的聚类标签,该聚类标签是每个类别对应的不同的簇,而这些簇又有着它们各自对应的语料,我们将这些簇称为聚类标签,每个不同的簇对应的语料作为聚类标签对应的聚类语料,进而,得到了聚类处理后包括聚类标签以及聚类标签对应的聚类语料的聚类结果。

[0129] 参照图5,步骤S22具体包括:

[0130] 步骤A1,将所述聚类样本语料进行分类,并将分类过后同种类的聚类样本语料划分为一个簇,得到不同种类的簇对应的聚类标签;

[0131] 在一具体实施例中,对聚类样本语料进行分类,得到分类后聚类样本语料属于不同种类的聚类标签,并且这些聚类样本语料都按照聚类标签进行分类,得到聚类标签以及聚类标签对应的聚类语料。

- [0132] 步骤A2,基于所述簇对应的聚类标签,确定不同种类的聚类标签对应的簇内语料;
- [0133] 在一具体实施例中,通过基于层次凝聚聚类算法HAC对聚类样本语料进行分类,得到每个类别对应的不同的簇,而这些簇又有着它们各自对应的语料,我们将这些簇称为聚类标签,每个不同的簇对应的语料作为聚类标签对应的聚类语料。
- [0134] 步骤A3,若所述簇内语料的数量大于预设阈值N1,则所述簇内语料以所述簇为聚类标签,聚类标签对应的簇内语料为对应的聚类语料;
- [0135] 步骤A4,若所述簇内语料的数量不大于预设阈值N1,则所述簇内语料以other为聚类标签,聚类标签对应的簇内语料为对应的聚类语料。
- [0136] 在一具体的实施例中,聚类样本语料可以分为label1、label2、label3、label4以及label5五个不同的类别,这五个不同类别的聚类标签下又有其对应的聚类语料,若聚类结果中簇内语料数量大于预设阈值50的簇包括label1、label2、label3以及label4,则这些聚类标签为label1、label2、label3以及label4的聚类语料以簇id,label1、label2、label3以及label4为聚类标签,而剩余聚类样本语料label5对应的聚类语料以other为聚类标签。
- [0137] 在本实施例中,通过将训练语料进行乱序、划分,得出进行聚类的聚类样本语料,且另外的聚类非样本语料也将进行保留,基于层次凝聚聚类算法HAC,对所述聚类样本语料进行聚类处理,得到聚类处理后的聚类结果,该聚类结果包括聚类标签以及聚类标签对应的聚类语料,通过上述方法得到的用于训练目标意图识别模型的聚类结果,引入聚类算法对训练数据进行处理,减少了构建意图识别文本分类过程中的人力投入,提高了构建意图识别模型的效率。
- [0138] 进一步地,基于本发明模型构建方法第一、第二实施例,提出本发明模型构建方法第三实施例。
- [0139] 模型构建方法的第三实施例与模型构建方法的第一、第二实施例的区别在于,本实施例是对步骤S30,所述基于所述聚类结果中的聚类标签以及对应的聚类语料进行模型训练和预测,根据模型训练和预测结果确定目标意图识别模型的细化,参照图6,该步骤具体包括:
- [0140] 步骤S31,将聚类结果中的聚类语料,根据聚类标签划分为训练语料和预测语料;
- [0141] 将聚类结果中的聚类语料按聚类标签平均切分成n份,每次取n份聚类语料中的m份聚类语料对于训练初始模型进行训练,得到对应的分类模型,并把n份聚类语料中剩余的聚类语料进行预测,进而得到全部聚类语料的预测值分数。
- [0142] 在一具体实施例中,将聚类结果中的聚类语料按聚类标签label1以及label2平均切分成5份,则在聚类标签为label1的聚类语料中,当聚类语料数量为200,切分后每份聚类语料含标签为label1的数量为40,每次取4份聚类语料进行训练,剩余一份进行预测,可以得到聚类标签为label1的全部聚类语料的预测分数值;在聚类标签为label2的聚类语料中,当聚类语料数量为400,切分后每份聚类语料含标签为label2的数量为80,每次取4份聚类语料进行训练,剩余一份进行预测,可以得到聚类标签为label2的全部聚类语料的预测分数值。
- [0143] 步骤S32,基于所述训练语料进行模型训练,得到训练好的初始分类模型;
- [0144] 步骤S33,将所述预测语料输入所述训练好的初始分类模型进行预测,得到预测分数值;

[0145] 在一具体的实施例中,将划分后聚类语料中的训练语料用于对预训练分类模型进行训练,得到一个对应的分类模型,并且将上述划分后聚类语料中的预测语料输入上述分类模型中,进行分类后得到对应的分类结果,对该分类结果进行预测测试,得到上述分类模型对应的分类结果的预测分数值。

[0146] 步骤S34,基于所述聚类结果中的聚类标签与所述预测分数值,确定所述聚类结果的精确召回率PRF值;

[0147] 通过分类模型的分类结果对应的预测分数值,得到聚类结果中的聚类标签以及聚类结果对应的Precision(精确率)、Recall(召回率)、F1值(F1),简称PRF值,并以PRF值来评估该分类模型是否合理,进而评估该分类模型的训练数据是否合理,进一步地,得出聚类结果是否合理。

[0148] 步骤S35,基于所述PRF值,确定对应的目标意图识别模型。

[0149] 在一具体实施例中,根据聚类结果的PRF值,可以确定输出的对应的目标意图识别模型,具体的,对PRF值进行值判定规则如下:

[0150] 如果PRF值达到预设阈值,聚类结果合理,则采用聚类结果中的全部语料对预训练模型进行训练,得出对应的分类模型;

[0151] 如果PRF值未达到预设阈值,聚类结果不合理,则需对聚类结果对应的聚类标签以及聚类标签对应的聚类语料进行分类调整,得到分类调整后的聚类标签以及聚类标签对应的聚类语料,重新对分类调整后的聚类标签以及聚类标签对应的聚类语料的聚类结果进行模型训练和预测,直到聚类结果对应的PRF值达到预设阈值,并输出对应的分类模型。

[0152] 在本实施例中,通过对聚类处理后的聚类结果进行训练和预测,得到聚类结果对应的分类模型的分输出结果以及聚类结果对应的预测分数值,根据上述分输出结果以及预测分数值,得出聚类结果的PRF值,根据PRF值是否达到预设阈值来判定聚类结果是否合理,并且最终确定对应的目标意图识别模型,可以提高自动创建目标意图识别模型的正确率,设置容错机制,对分类模型的分输出结果进行预测和判定,提高分类模型分输出结果的准确率。

[0153] 进一步地,基于本发明模型构建方法第一、第二、第三实施例,提出本发明模型构建方法第四实施例。

[0154] 模型构建方法的第四实施例与模型构建方法的第一、第二、第三实施例的区别在于,本实施例是对步骤S35,基于所述PRF值,确定对应的目标意图识别模型的细化,参照图7,该步骤具体包括:

[0155] 步骤B1,判断所述PRF值是否达到预设阈值;

[0156] 步骤B2,若所述PRF值达到预设阈值,则所述聚类结果合理,输出所述初始分类模型作为目标意图识别模型;

[0157] 在一具体实施例中,若分类模型输出的分类结果的PRF值是达到预设阈值的,则判断聚类处理后的训练语料是否全部加入训练预训练模型的过程,若全部加入,则直接输出分类模型作为目标意图识别模型,若未全部加入,则添加训练语料对分类模型进行进一步的训练,直到所有训练语料都加入训练,输出分类模型作为目标意图识别模型。

[0158] 步骤B3,若所述PRF值未达到预设阈值,则所述聚类结果不合理,对所述聚类结果进行分类调整,得到分类调整后的聚类结果;

[0159] 参照图8,步骤B3具体包括:

[0160] 所述聚类结果包括聚类标签为other的other聚类语料和聚类标签为非other的非other聚类语料,所述对所述聚类结果进行分类调整,得到分类调整后的聚类结果的步骤包括:

[0161] 步骤b1,调整所述other聚类语料和非other聚类语料,获取调整后的聚类标签对应的聚类语料;

[0162] 步骤b2,计算所述调整后的聚类标签对应的聚类语料的混淆度;

[0163] 步骤b3,当所述混淆度大于预设阈值T2,则合并调整分类前后的非other的聚类语料作为当前的聚类结果的非other聚类语料,其他语料则作为聚类标签为other的other聚类语料。

[0164] 在一具体实施例中,对所述聚类结果进行分类调整的步骤包含三部分:

[0165] 首先,第一部分是将从非other聚类语料中不可信的部分更改标签至聚类标签为other的聚类语料中,通过上述步骤会得到所有聚类语料的聚类标签以及预测分数值。将预测分数值低于某一阈值0.3,并且预测值和聚类标签为非other的聚类语料中不一致的聚类语料调换至聚类标签为other的聚类语料中,更改标签综合了分类和聚类的优点,将聚类标签为非other的聚类语料中不可信部分放入other中。

[0166] 然后,判断聚类标签为other的语料数量是否超过阈值整体训练语料的10%,若超过,则将新的聚类结果执行上述聚类算法,将聚类结果中聚类语料数量超过某一预设阈值N2的簇内语料标记为新的聚类语料,并且将该聚类语料加入到本次训练训练语料中。

[0167] 最后,对新的聚类语料以及分类调整前的聚类语料进行分类合并,分类合并的过程需要计算分类调整前后的聚类语料的混淆度,计算混淆度的公式如下:

$$[0168] \quad confusion\_percent_{catei,catej} = \left[ \left( \frac{N_{catei,catej}}{N_{catei}} + \frac{N_{catei,catej}^{pre}}{N_{catej}^{pre}} \right) + \left( \frac{N_{catej,catei}}{N_{catej}} + \frac{N_{catej,catei}^{pre}}{N_{catei}^{pre}} \right) \right] \div 4$$

[0169] 其中, $N_{catei,catej}$ 表示实际意图为catei,但是误分到catej的数量; $N_{catei}$ 表示catei实际数量; $N_{catei}^{pre}$ 表示预测到catei的数量;合并混淆度大于阈值T2(例:0.25)的两个聚类标签为非other的聚类语料,得到分类调整后的聚类结果。

[0170] 步骤B4,将所述分类调整后的聚类结果作为当前的聚类结果;

[0171] 获取当前的聚类结果,并执行以下步骤:

[0172] 将聚类结果中的聚类语料,根据聚类标签划分为训练语料和预测语料;

[0173] 基于所述训练语料进行模型训练,得到训练好的初始分类模型;

[0174] 将所述预测语料输入所述训练好的初始分类模型进行预测,得到预测分数值;

[0175] 基于所述聚类结果中的聚类标签与所述预测分数值,确定所述聚类结果的精确召回率PRF值;

[0176] 在一具体实施例中,聚类结果的PRF值未达到预设阈值,聚类结果不合理,则需要对聚类结果对应的聚类标签以及聚类标签对应的聚类语料进行分类调整,得到分类调整后的聚类结果后,需要将分类调整后的聚类结果作为当前的聚类结果,重复执行基于所述聚类结果中的聚类标签以及对应的聚类语料进行模型训练和预测,根据模型训练和预测结果确定目标意图识别模型的步骤。

[0177] 进一步地,将聚类结果中的聚类语料根据聚类标签划分为训练语料和预测预料,通过训练语料得到分类模型的分类结果,通过预测预料得到该聚类结果的预测分数值,基于分类结果以及预测分数值确定聚类结果对应的PRF值,再根据PRF值确定对应的目标意图识别模型。

[0178] 步骤B5,直到所述PRF值达到预设阈值,所述聚类结果合理,输出所述初始分类模型作为目标意图识别模型。

[0179] 若所述PRF值达到预设阈值,聚类结果是合理的,则输出上述聚类结果训练得到的分类模型作为目标意图识别模型。

[0180] 在本实施例中,通过判断分类模型的分类结果的PRF值,确定分类模型是否合理,再进一步的确定分类模型的训练数据是否合理,更进一步的确定包含训练数据的聚类结果是否合理,若合理,则直接输出分类模型得到目标意图识别模型,若不合理则需对聚类结果进行分类调整,得到对应的模型训练数据,进而得到目标意图识别模型,提高自动创建目标意图识别模型的正确率,提高自动创建目标意图识别模型的效率。

[0181] 本发明还提供一种模型构建装置。参照图9,本发明的模型构建装置包括:

[0182] 获取模块10,用于获取构建模型的训练语料;

[0183] 聚类模块20,用于基于预先训练好的聚类模型,对所述训练语料进行聚类处理,得到对应的聚类结果,其中,聚类结果包括聚类标签以及聚类标签对应的聚类语料;

[0184] 确定模块30,用于基于所述聚类结果中的聚类标签以及对应的聚类语料进行模型训练和预测,根据模型训练和预测结果确定目标意图识别模型。

[0185] 此外,本发明还提供一种计算机可读存储介质,所述介质优选为计算机可读存储介质,其上存储有模型构建程序,所述模型构建程序被处理器执行时实现如上所述的模型构建方法的步骤。

[0186] 在本发明模型构建设备和介质的实施例中,包含了上述模型构建方法各实施例的全部技术特征,说明和解释内容与上述模型构建方法各实施例基本相同,此处不再赘述。

[0187] 需要说明的是,在本文中,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者系统不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者系统所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括该要素的过程、方法、物品或者系统中还存在另外的相同要素。

[0188] 上述本发明实施例序号仅仅为了描述,不代表实施例的优劣。

[0189] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现,当然也可以通过硬件,但很多情况下前者是更佳的实施方式。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来,该计算机软件产品存储在如上所述的一个存储介质(如ROM/RAM、磁碟、光盘)中,包括若干指令用以使得一台终端设备(可以是手机,计算机,服务器,空调器,或者网络设备等)执行本发明各个实施例所述的方法。

[0190] 以上仅为本发明的优选实施例,并非因此限制本发明的专利范围,凡是利用本发明说明书与附图内容所作的等效结构或等效流程变换,或直接或间接运用在其他相关的技术领域,均同理包括在本发明的专利保护范围内。

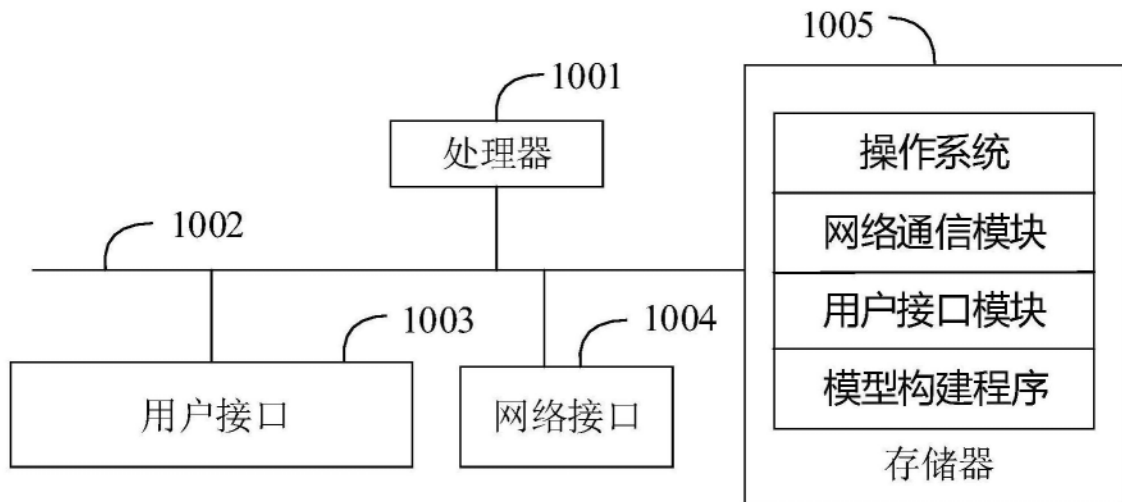


图1

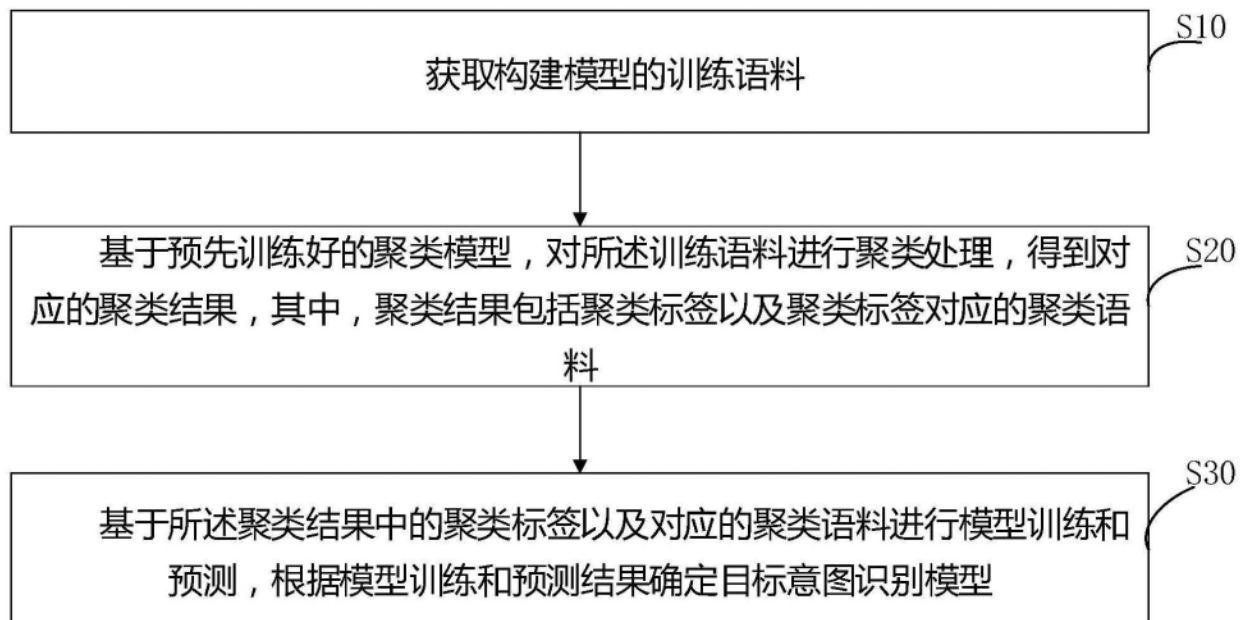


图2

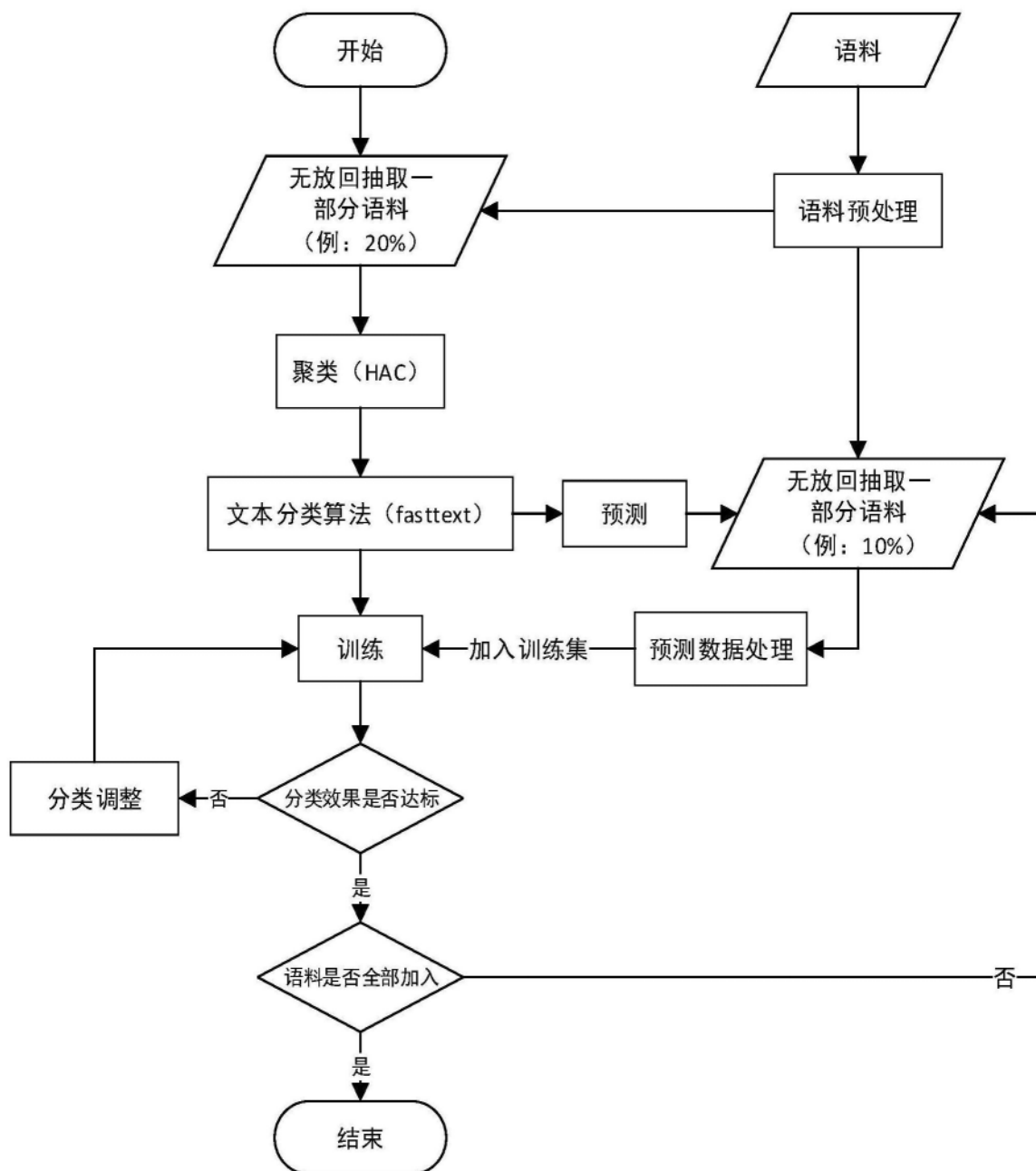


图3



图4

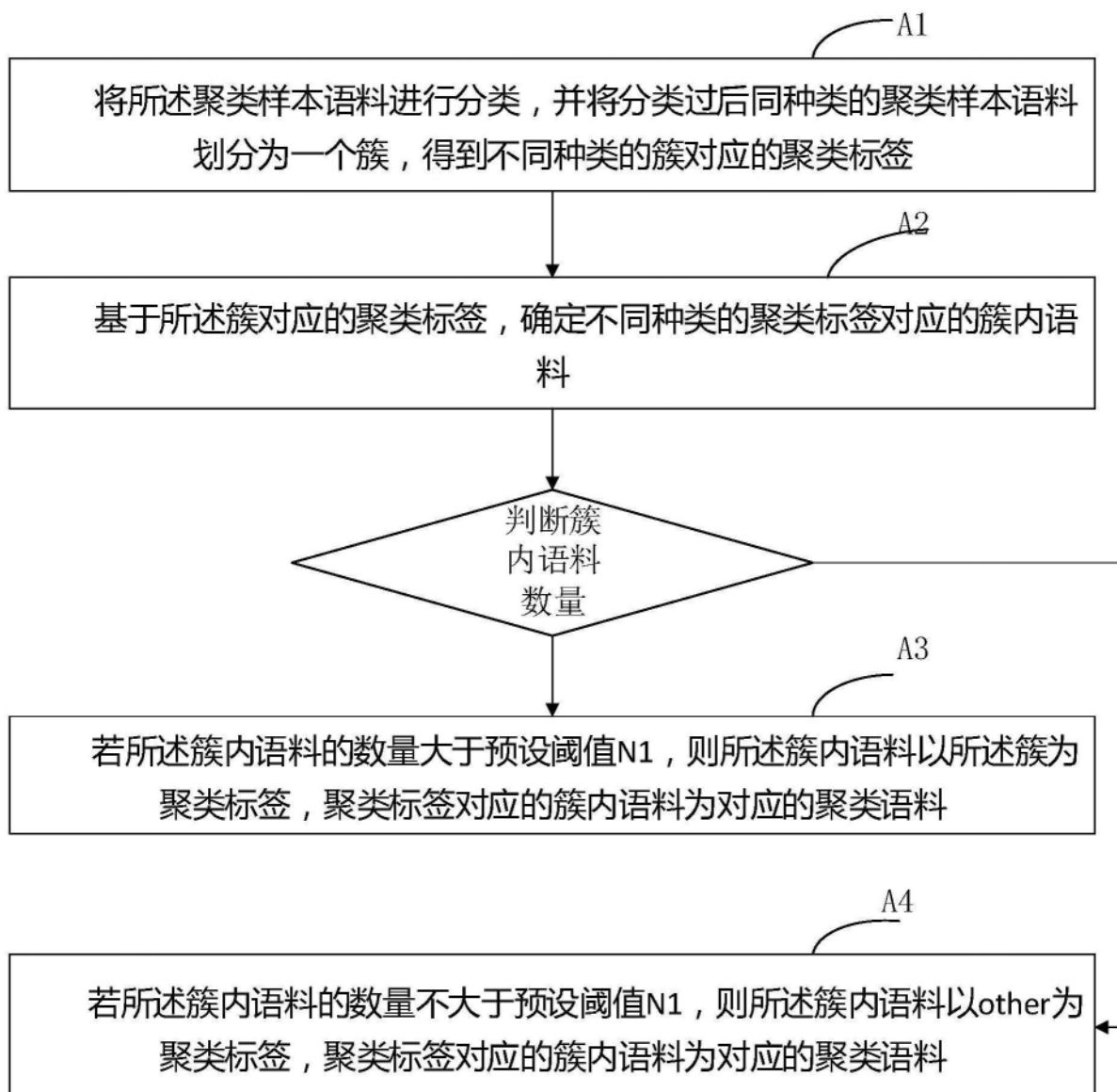


图5

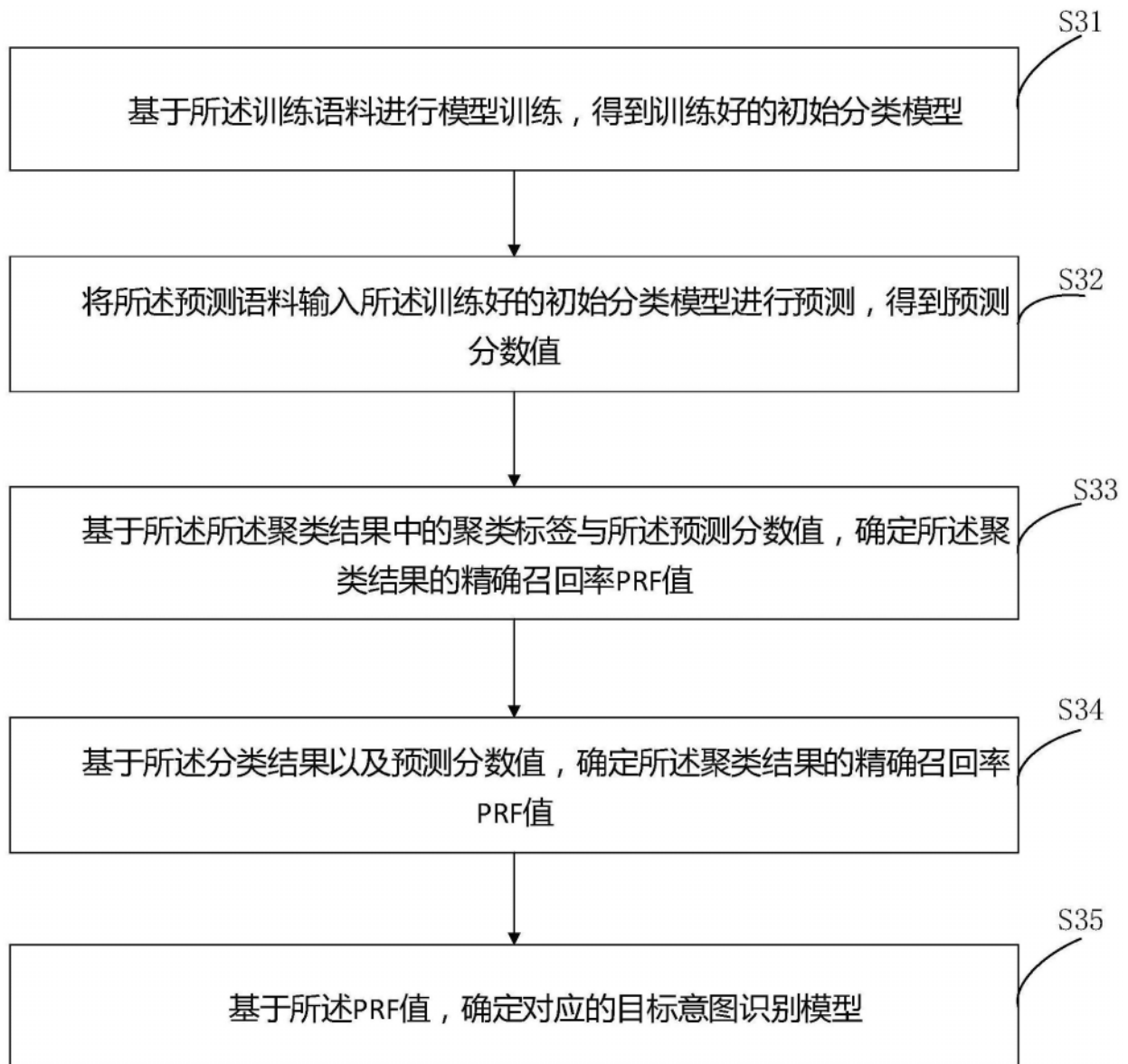


图6

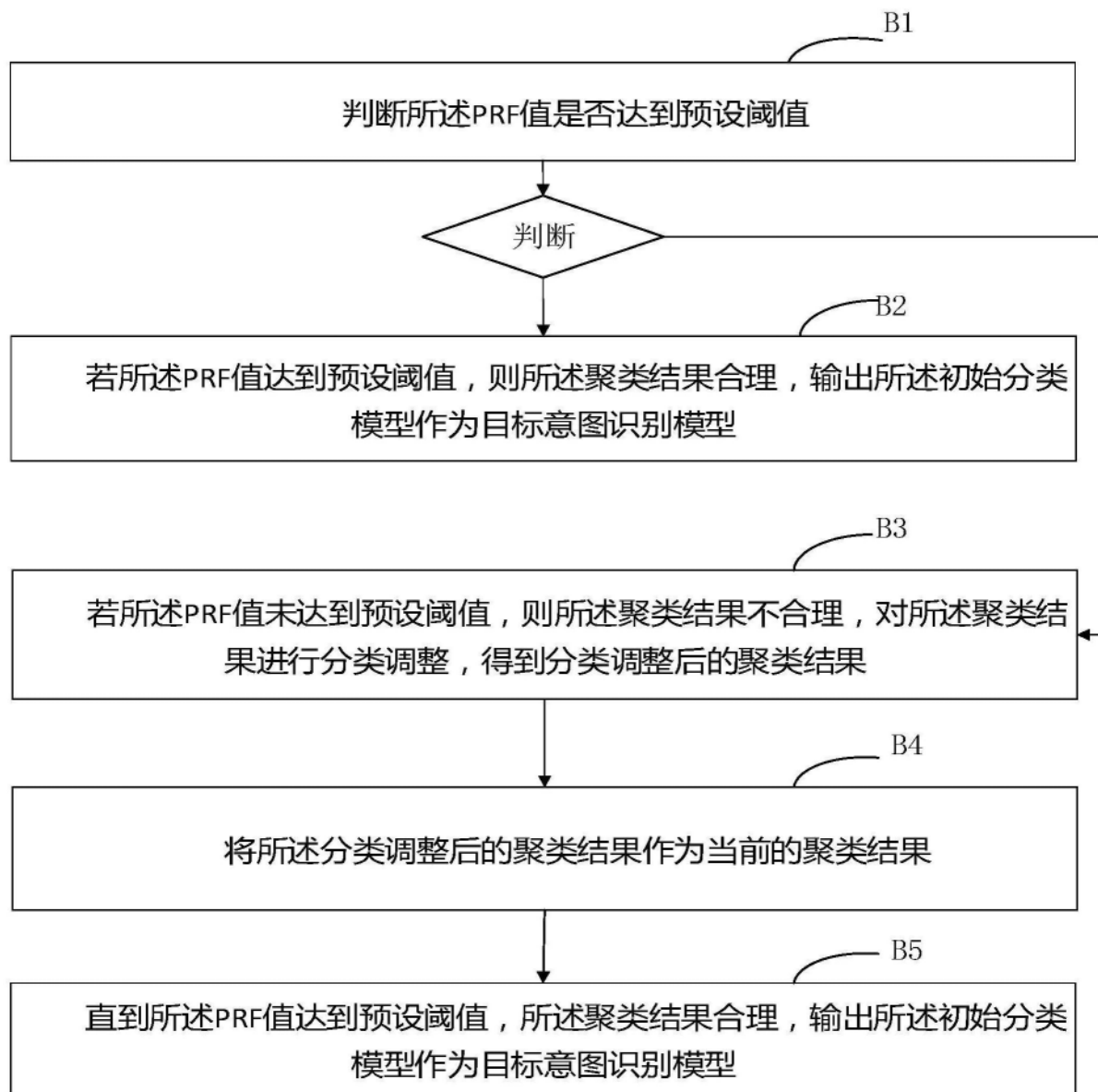


图7

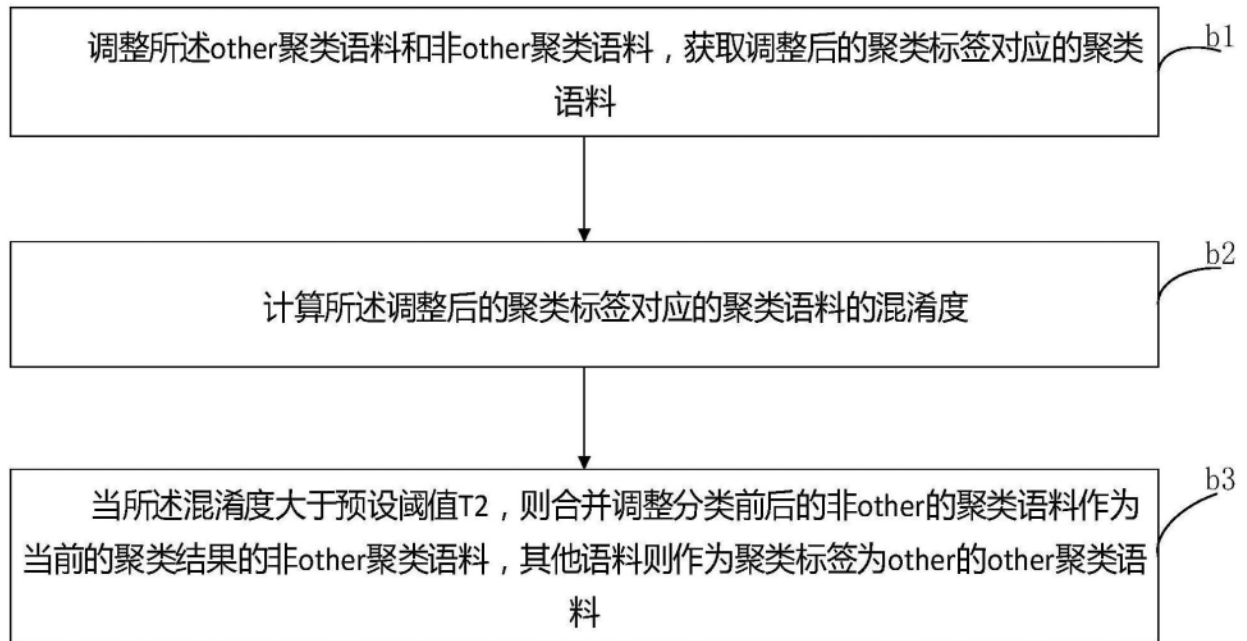


图8

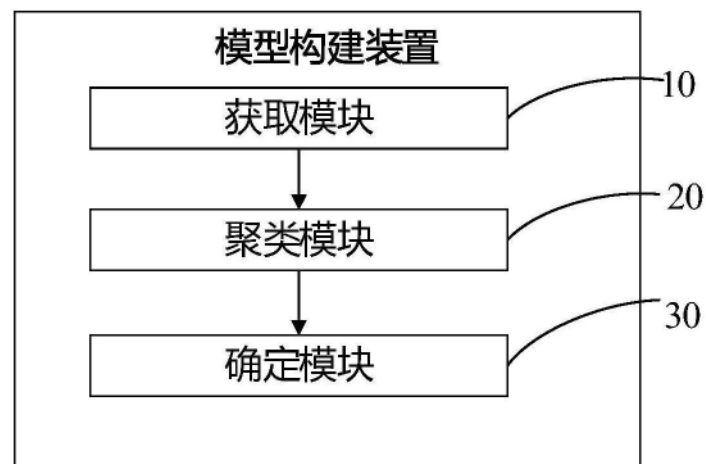


图9