(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112580298 B
(45)授权公告日 2024.05.07

(21)申请号 201910930956.4

(22)申请日 2019.09.29

(65)同一申请的已公布的文献号
申请公布号 CN 112580298 A

(43)申请公布日 2021.03.30

(73)专利权人 大众问问(北京)信息科技有限公司
地址 100098 北京市海淀区北三环西路25号27号楼三层3011室

(72)发明人 杜京钢

(74)专利代理机构 北京乾成律信知识产权代理有限公司 11927
专利代理师 苏捷 姚志远

(51)Int.Cl.
G06F 40/117(2020.01)
G06F 40/30(2020.01)

(56)对比文件
CN 101151843 A,2008.03.26
CN 108304372 A,2018.07.20
CN 108959257 A,2018.12.07
CN 109388700 A,2019.02.26
CN 109918680 A,2019.06.21
CN 109949799 A,2019.06.28
CN 110222328 A,2019.09.10
US 2018293148 A1,2018.10.11
WO 2018149326 A1,2018.08.23
陈洪平.面向Deep Web的数据抽取与语义标注技术研究.中国优秀硕士论文电子期刊网.2011,第I138-1102页.
刘炜;王旭;张雨嘉;刘宗田.一种面向突发事件的文本语料自动标注方法.中文信息学报.2017,(02),第81-90页.
周彬彬;张宏军;张睿;冯蕴天;徐有为.军事语料实体标注系统的设计与实现.信息系统工程.2018,(08),第58-62页.
潘炜.面向层次分类标签的词性标注系统.中国优秀硕士论文电子期刊网.2009,第I138-1551页.
Vincentius .Regular Expression Matching for XWraps Action Level Data.CEED.2007,第117-124页.

审查员 熊林

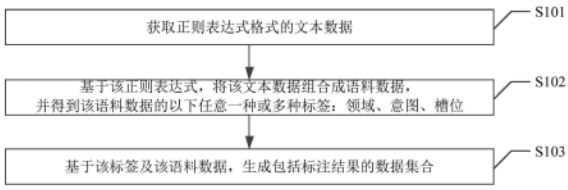
权利要求书2页 说明书12页 附图2页

(54)发明名称

一种标注数据获取方法、装置及设备

(57)摘要

本发明实施例公开了一种标注数据获取方法、装置及设备,方法包括:获取正则表达式格式的文本数据;基于正则表达式,将文本数据组合成语料数据,并得到语料数据的以下任意一种或多种标签:领域、意图、槽位;基于标签及语料数据,生成包括标注结果的数据集合;本方案中,文本数据符合正则表达式,对于设备来说,其可以基于该正则表达式将文本数据组合成语料数据,并得到相应的标签,这样,相比于人工标注,节省了较多人力。



CN 112580298 B

1.一种标注数据获取方法,其特征在于,包括:

获取正则表达式格式的文本数据;

基于所述正则表达式,将所述文本数据组合成语料数据,并得到所述语料数据的以下任意一种或多种标签:领域、意图、槽位;

基于所述标签及所述语料数据,生成包括标注结果的数据集合,包括:

针对每份槽位数据,确定该份槽位数据对应的每种表达方式下的语料数据;基于所述每种表达方式下的语料数据对应的领域标签、意图标签和槽位标签,生成该份槽位数据对应的数据子集;得到包括每份槽位数据对应的数据子集的数据集合;

或者,针对每种表达方式,确定该种表达方式下的每份槽位数据对应的语料数据;基于所述每份槽位数据对应的语料数据的领域标签、意图标签和槽位标签,生成该种表达方式对应的数据子集;得到包括每种表达方式对应的数据子集的数据集合。

2.根据权利要求1所述的方法,其特征在于,所述正则表达式中:将可相互替换的多份内容填充于同一标号区间中,同一标号区间中的多份内容之间由分隔符隔开;

所述基于所述正则表达式,将所述文本数据组合成语料数据,包括:

识别所述文本数据中的标号区间;

基于标号区间内的分隔符,识别标号区间内的多份语料子数据;

将识别到的多份语料子数据组合成语料数据。

3.根据权利要求2所述的方法,其特征在于,所述正则表达式中:将槽位数据对应的变量参数填充于预设标号区间中,所述变量参数指代多份槽位数据;

所述将识别到的多份语料子数据组合成语料数据,包括:

将识别到的多份语料子数据、以及所述多份槽位数据组合成语料数据。

4.根据权利要求3所述的方法,其特征在于,得到所述语料数据的领域标签,包括:基于用户定义的领域类型,得到所述语料数据的领域标签;

得到所述语料数据的意图标签,包括:基于用户定义的意图类型,得到所述语料数据的意图标签;

得到所述语料数据的槽位标签,包括:识别所述文本数据中的槽位标签。

5.根据权利要求1所述的方法,其特征在于,所述方法还包括:

判断表达方式的数量是否大于槽位数据的数量;

如果大于,执行所述针对每份槽位数据,确定该份槽位数据对应的每种表达方式下的语料数据;基于所述每种表达方式下的语料数据对应的领域标签、意图标签和槽位标签,生成该份槽位数据对应的数据子集;得到包括每份槽位数据对应的数据子集的数据集合的步骤;

如果不大于,执行所述针对每种表达方式,确定该种表达方式下的每份槽位数据对应的语料数据;基于所述每份槽位数据对应的语料数据的领域标签、意图标签和槽位标签,生成该种表达方式对应的数据子集;得到包括每种表达方式对应的数据子集的数据集合的步骤。

6.一种标注数据获取装置,其特征在于,包括:

获取模块,用于获取正则表达式格式的文本数据;

组合模块,用于基于所述正则表达式,将所述文本数据组合成语料数据;

获得模块,用于得到所述语料数据的以下任意一种或多种标签:领域、意图、槽位;

生成模块,用于基于所述标签及所述语料数据,生成包括标注结果的数据集合,所述生成模块包括:第一生成子模块或者第二生成子模块,其中,

所述第一生成子模块,用于针对每份槽位数据,确定该份槽位数据对应的每种表达方式下的语料数据;基于所述每种表达方式下的语料数据对应的领域标签、意图标签和槽位标签,生成该份槽位数据对应的数据子集;得到包括每份槽位数据对应的数据子集的数据集合;

所述第二生成子模块,用于针对每种表达方式,确定该种表达方式下的每份槽位数据对应的语料数据;基于所述每份槽位数据对应的语料数据的领域标签、意图标签和槽位标签,生成该种表达方式对应的数据子集;得到包括每种表达方式对应的数据子集的数据集合。

7.根据权利要求6所述的装置,其特征在于,所述正则表达式中:将可相互替换的多份内容填充于同一标号区间中,同一标号区间中的多份内容之间由分隔符隔开;

所述组合模块,具体用于:识别所述文本数据中的标号区间;基于标号区间内的分隔符,识别标号区间内的多份语料子数据;将识别到的多份语料子数据组合成语料数据。

8.根据权利要求7所述的装置,其特征在于,所述正则表达式中:将槽位数据对应的变量参数填充于预设标号区间中,所述变量参数指代多份槽位数据;

所述组合模块,还用于将识别到的多份语料子数据、以及所述多份槽位数据组合成语料数据。

9.根据权利要求8所述的装置,其特征在于,所述获得模块,具体用于:

基于用户定义的领域类型,得到所述语料数据的领域标签;

基于用户定义的意图类型,得到所述语料数据的意图标签;

识别所述文本数据中的槽位标签。

10.根据权利要求6所述的装置,其特征在于,所述装置还包括:

判断模块,用于判断表达方式的数量是否大于槽位数据的数量;如果大于,触发所述第一生成子模块;如果不大于,触发所述第二生成子模块。

11.一种电子设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,其特征在于,所述处理器执行所述程序时实现如权利要求1至5任意一项所述的方法。

一种标注数据获取方法、装置及设备

技术领域

[0001] 本发明涉及自然语言技术领域,特别是指一种标注数据获取方法、装置及设备。

背景技术

[0002] 一些场景中,用户可以与智能设备如车载设备、智能家居设备、或者手机、电脑等终端设备进行语音交互。这些智能设备对用户发出的语音指令进行语音识别,语音识别过程中,将语音数据转化为语料数据,将语料数据输入至训练得到的识别模型进行语义解析。

[0003] 训练得到该识别模型需要大量标注过的文本数据,比如,可以基于领域(domain)、意图(intent)、槽位(slot)对文本数据进行标注。领域(domain)是指同一类型的数据或者资源,以及围绕这些数据或资源提供的服务,比如“餐厅”,“酒店”,“飞机票”、“火车票”、“电影院”等;意图(intent)是指对于领域数据的操作,一般以动宾短语来命名,例如打电话、设置温度等;槽位(slot)用来存放领域(domain)的某些属性,比如餐厅领域中,槽位可以包括:位置、餐厅名、距离等等,再比如飞机票领域中,槽位可以包括:出发时间、出发地、目的地等等。

[0004] 目前,获取标注数据的方案一般包括:获取未标注过的原始数据,然后由人工逐一对这些原始数据进行标注,原始数据的数据量大,耗费较多的人力。

发明内容

[0005] 有鉴于此,本发明的目的在于提出一种标注数据获取方法、装置及设备,以节省人力。

[0006] 基于上述目的,本发明实施例提供了一种标注数据获取方法,包括:

[0007] 获取正则表达式格式的文本数据;

[0008] 基于所述正则表达式,将所述文本数据组合成语料数据,并得到所述语料数据的以下任意一种或多种标签:领域、意图、槽位;

[0009] 基于所述标签及所述语料数据,生成包括标注结果的数据集合。

[0010] 可选的,所述正则表达式中:将可相互替换的多份内容填充于同一标号区间中,同一标号区间中的多份内容之间由分隔符隔开;

[0011] 所述基于所述正则表达式,将所述文本数据组合成语料数据,包括:

[0012] 识别所述文本数据中的标号区间;

[0013] 基于标号区间内的分隔符,识别标号区间内的多份语料子数据;

[0014] 将识别到的多份语料子数据组合成语料数据。

[0015] 可选的,所述正则表达式中:将槽位数据对应的变量参数填充于预设标号区间中,所述变量参数指代多份槽位数据;

[0016] 所述将识别到的多份语料子数据组合成语料数据,包括:

[0017] 将识别到的多份语料子数据、以及所述多份槽位数据组合成语料数据。

[0018] 可选的,得到所述语料数据的领域标签,包括:基于用户定义的领域类型,得到所

述语料数据的领域标签；

[0019] 得到所述语料数据的意图标签，包括：基于用户定义的意图类型，得到所述语料数据的意图标签；

[0020] 得到所述语料数据的槽位标签，包括：识别所述文本数据中的槽位标签。

[0021] 可选的，基于所述标签及所述语料数据，生成包括标注结果的数据集合，包括：

[0022] 针对每份槽位数据，确定该份槽位数据对应的每种表达方式下的语料数据；基于所述每种表达方式下的语料数据对应的领域标签、意图标签和槽位标签，生成该份槽位数据对应的数据子集；得到包括每份槽位数据对应的数据子集的数据集合；

[0023] 或者，针对每种表达方式，确定该种表达方式下的每份槽位数据对应的语料数据；基于所述每份槽位数据对应的语料数据的领域标签、意图标签和槽位标签，生成该种表达方式对应的数据子集；得到包括每种表达方式对应的数据子集的数据集合。

[0024] 可选的，所述方法还包括：

[0025] 判断表达方式的数量是否大于槽位数据的数量；

[0026] 如果大于，执行所述针对每份槽位数据，确定该份槽位数据对应的每种表达方式下的语料数据；基于所述每种表达方式下的语料数据对应的领域标签、意图标签和槽位标签，生成该份槽位数据对应的数据子集；得到包括每份槽位数据对应的数据子集的数据集合的步骤；

[0027] 如果不大于，执行所述针对每种表达方式，确定该种表达方式下的每份槽位数据对应的语料数据；基于所述每份槽位数据对应的语料数据的领域标签、意图标签和槽位标签，生成该种表达方式对应的数据子集；得到包括每种表达方式对应的数据子集的数据集合的步骤。

[0028] 基于上述目的，本发明实施例还提供了一种标注数据获取装置，包括：

[0029] 获取模块，用于获取正则表达式格式的文本数据；

[0030] 组合模块，用于基于所述正则表达式，将所述文本数据组合成语料数据；

[0031] 获得模块，用于得到所述语料数据的以下任意一种或多种标签：领域、意图、槽位；

[0032] 生成模块，用于基于所述标签及所述语料数据，生成包括标注结果的数据集合。

[0033] 可选的，所述正则表达式中：将可相互替换的多份内容填充于同一标号区间中，同一标号区间中的多份内容之间由分隔符隔开；

[0034] 所述组合模块，具体用于：识别所述文本数据中的标号区间；基于标号区间内的分隔符，识别标号区间内的多份语料子数据；将识别到的多份语料子数据组合成语料数据。

[0035] 可选的，所述正则表达式中：将槽位数据对应的变量参数填充于预设标号区间中，所述变量参数指代多份槽位数据；

[0036] 所述组合模块，还用于将识别到的多份语料子数据、以及所述多份槽位数据组合成语料数据。

[0037] 可选的，所述获得模块，具体用于：

[0038] 基于用户定义的领域类型，得到所述语料数据的领域标签；

[0039] 基于用户定义的意图类型，得到所述语料数据的意图标签；

[0040] 识别所述文本数据中的槽位标签。

[0041] 可选的，所述生成模块包括：第一生成子模块或者第二生成子模块，其中，

[0042] 所述第一生成子模块,用于针对每份槽位数据,确定该份槽位数据对应的每种表达方式下的语料数据;基于所述每种表达方式下的语料数据对应的领域标签、意图标签和槽位标签,生成该份槽位数据对应的数据子集;得到包括每份槽位数据对应的数据子集的数据集合;

[0043] 所述第二生成子模块,用于针对每种表达方式,确定该种表达方式下的每份槽位数据对应的语料数据;基于所述每份槽位数据对应的语料数据的领域标签、意图标签和槽位标签,生成该种表达方式对应的数据子集;得到包括每种表达方式对应的数据子集的数据集合。

[0044] 可选的,所述装置还包括:

[0045] 判断模块,用于判断表达方式的数量是否大于槽位数据的数量;如果大于,触发所述第一生成子模块;如果不大于,触发所述第二生成子模块。

[0046] 基于上述目的,本发明实施例还提供了一种电子设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,所述处理器执行所述程序时实现上述任一种标注数据获取方法。

[0047] 应用本发明所示实施例,获取正则表达式格式的文本数据;基于正则表达式,将文本数据组合成语料数据,并得到语料数据的以下任意一种或多种标签:领域、意图、槽位;基于标签及语料数据,生成包括标注结果的数据集合;本方案中,文本数据符合正则表达式,对于设备来说,其可以基于该正则表达式将文本数据组合成语料数据,并得到相应的标签,这样,相比于人工标注,节省了较多人力。

附图说明

[0048] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术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发明的一些实施例,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。

[0049] 图1为本发明实施例提供的标注数据获取方法的第一种流程示意图;

[0050] 图2为本发明实施例提供的标注数据获取方法的第二种流程示意图;

[0051] 图3为本发明实施例提供的一种标注数据获取装置的结构示意图;

[0052] 图4为本发明实施例提供的一种电子设备的结构示意图。

具体实施方式

[0053] 为使本发明的目的、技术方案和优点更加清楚明白,以下结合具体实施例,并参照附图,对本发明进一步详细说明。

[0054] 需要说明的是,本发明实施例中所有使用“第一”和“第二”的表述均是为了区分两个相同名称非相同的实体或者非相同的参量,可见“第一”“第二”仅为了表述的方便,不应理解为对本发明实施例的限定,后续实施例对此不再一一说明。

[0055] 为了解决上述技术问题,本发明实施例提供了一种标注数据获取方法、装置及设备,该方法和装置可以应用于各种电子设备,比如车载设备、智能家居设备、手机、电脑等终端设备,具体不做限定。下面首先对发明实施例提供的标注数据获取方法进行介绍。

[0056] 图1为本发明实施例提供的标注数据获取方法的第一种流程示意图,包括:

[0057] S101:获取正则表达式格式的文本数据。

[0058] 举例来说,正则表达式可以理解为对字符串操作的一种逻辑公式,比如,用预先定义好的一些特定字符、及这些特定字符的组合,组成一个规则字符串,该规则字符串用来表达对字符串的一种过滤逻辑。

[0059] 一种实施方式中,该正则表达式中:将可相互替换的多份内容填充于同一标号区间中,同一标号区间中的多份内容之间由分隔符隔开。

[0060] 举例来说,标号区间可以为小括号()、中括号[]、大括号{}、书名号《》、或者<>等标号中的区间,具体标号不做限定。同一标号区间中的多份内容之间可以通过分隔符“|”隔开,或者,也可以采用其他分隔符,比如“、”“;”等等,具体分隔符不做限定。

[0061] 一种情况下,该正则表达式的定义可以为:

[0062]

[]	其中的内容可以选其一也可为空,不同内容用“|”分隔开
()	其中的内容必可以选其一并且不可为空,不同内容用“|”分隔开
<>	代表内容需要根据实际需要调整
{}	其中的内容由多个可替换的变量组成,且组成方式比较多样

[0063] S102:基于该正则表达式,将该文本数据组合成语料数据,并得到语料数据的以下任意一种或多种标签:领域、意图、槽位。

[0064] 领域(domain)是指同一类型的数据或者资源,以及围绕这些数据或资源提供的服务,比如“餐厅”,“酒店”,“飞机票”、“火车票”、“电影院”等;意图(intent)是指对于领域数据的操作,一般以动宾短语来命名,例如打电话、设置温度等;槽位(slot)用来存放领域(domain)的某些属性,比如餐厅领域中,槽位可以包括:位置、餐厅名、距离等等,再比如飞机票领域中,槽位可以包括:出发时间、出发地、目的地等等。

[0065] 如上所述,该正则表达式中:将可相互替换的多份内容填充于同一标号区间中,同一标号区间中的多份内容之间由分隔符隔开;这种实施方式中,识别所述文本数据中的标号区间;基于标号区间内的分隔符,识别标号区间内的多份语料子数据;将识别到的多份语料子数据组合成语料数据。

[0066] 举例来说,假设获取的语料数据可以为:[我想|我们]出发|导航](去|去一下)<CrossRoad>;其中,[我想|我们]出发|导航]表示:可以在“我想|我们]出发|导航”中任选一个,或者也可以为空;(去|去一下)表示:可以在“去|去一下”中任选一个,并且不可以为空;<CrossRoad>表示交叉路口,具体内容可以根据实际内容进行替换。

[0067] 识别出各标号区间:[我想|我们]出发|导航]、(去|去一下)、<CrossRoad>;针对[我想|我们]出发|导航],基于分隔符“|”,识别出多份领语料子数据为“我想”、“我们”、“出发”“导航”。针对(去|去一下),基于分隔符“|”,识别出多份语料子数据为“去”、“去一下”。

[0068] 一种情况下,该正则表达式中:将槽位数据对应的变量参数填充于预设标号区间中,所述变量参数指代多份槽位数据;。上述CrossRoad可以仅表示变量参数,其指代的具体内容可以如表1所示:

[0069] 表1

[0070]

CrossRoad
大望路与建国路十字路口

大望路与建国路道路交叉口
长安路口
大望路与建国路交叉路口
大望路与建国路交叉点
大望路与建国路交叉口
长安街与建国路交叉
……

[0071] 这样,可以将识别到的多份语料子数据、以及该多份槽位数据组合成语料数据,比如:我想去大望路与建国路十字路口、我想去长安路口、我们去一下长安路口、导航去大望路与建国路交叉口……可以基于上述正则表达式的定义,进行任意组合,不再一一列举。

[0072] 再举一例,槽位数据对应的变量参数可以为streetNo.(街道),这样,获取的语料数据可以为:[我想|我们|出发|导航](去|去一下)<streetNo.>,其指代的具体内容可以如表2所示:

[0073] 表2

[0074]

streetNo.
景山街道
金融街街道
朝外街道
丰台街道
鲁谷街道
海淀街道
大峪街道
……

[0075] 或者,槽位数据对应的变量参数可以为POI(Point of Interest兴趣点),这样,获取的语料数据可以为:[我想|我们|出发|导航](去|去一下)<POI>,其指代的具体内容可以如表3所示:

[0076] 表3

[0077]

POI
游乐场
家乐福超市
大悦城
苏宁购物广场
未来汇
永辉超市
……

[0078] 槽位数据的具体内容不做限定,不再一一列举。

[0079] 一种实施方式中,语料数据的标签包括领域标签;这种实施方式中,可以基于用户定义的领域类型,得到所述语料数据的领域标签。

[0080] 举例来说,可以由用户定义领域类型,基于用户定义的领域类型生成领域标签。比

如,用户可以定义领域类型为“导航”,这样,领域标签可以“导航”、或者与导航类似的相关词语。

[0081] 一种实施方式中,语料数据的标签包括意图标签;这种实施方式中,可以基于用户定义的意图类型,得到所述语料数据的意图标签。

[0082] 举例来说,可以由用户定义意图类型,基于用户定义的意图类型生成意图标签。比如,用户可以定义意图类型为“打电话”,这样,意图标签可以“打电话”、或者与“打电话”类似的相关词语。或者,一些情况下,可以省略“用户定义意图类型”这一步骤。或者,用户也可以定义意图类型为“other”,表示没有特定意义的意图。

[0083] 一种实施方式中,S101中获取的文本数据中可以包括槽位标签,这种实施方式中,可以识别所述文本数据中的槽位标签。

[0084] 比如,获取的文本数据可以为:[我想|我们|出发|导航](去|去一下)《to》<CrossRoad>《/to》;其中,[我想|我们|出发|导航]表示:可以在“我想|我们|出发|导航”中任选一个,或者也可以为空;(去|去一下)表示:可以在“去|去一下”中任选一个,并且不可以为空;<CrossRoad>表示交叉路口,具体内容可以根据实际内容进行替换;《to》<CrossRoad>《/to》表示对<CrossRoad>添加了槽位标签。根据槽位标签《to》《/to》的位置,识别出槽位标签指向的标号区间为<CrossRoad>,识别出的槽位数据为CrossRoad。

[0085] 一种情况下,槽位数据可以表示为表格的形式,比如CrossRoad可以如上述表1所示。

[0086] 一种实施方式中,正则表达式中:将多种可替换的槽位数据填充于预设标号区间中。上述例子,槽位数据为一种:CrossRoad(交叉路口),CrossRoad为可替换的槽位数据。再举一例,假设槽位数据包括三种:省份(Province)、城市(City)、交叉路口(CrossRoad),这三种均为可替换的槽位数据,这种情况下,槽位数据可以表达为{<Province><City><CrossRoad>},这样,预设标号区间可以为{}中的区间,可以识别出槽位标签指向的标号区间为{<Province><City><CrossRoad>},识别出的槽位数据为Province、City、CrossRoad。

[0087] 或者,槽位数据可以表达为{<Province>|<City>|<CrossRoad>},这样,预设标号区间可以为{}中的区间,可以识别出槽位标签指向的标号区间为{<Province>|<City>|<CrossRoad>},识别出的槽位数据为Province、City、CrossRoad。

[0088] 一种情况下,槽位数据可以表示为表格的形式,比如可以如表4所示:

[0089] 表4

[0090]

Province	City	CrossRoad
北京	北京	长安路口
上海	上海	南京路口
……	……	……

[0091] 如上所述,语料数据的预设标号区间中填充的可以仅为变量参数,一种情况下,电子设备可以基于该变量参数,获取变量参数对应的实际数据。比如,槽位标号区间中填充的可以仅为CrossRoad,电子设备基于CrossRoad在地图上查找实际的交叉路口信息,比如表1所示出的信息。

[0092] 再比如,槽位标号区间中填充的可以仅为<Province>|<City>|<CrossRoad>,电子设备基于Province、City、CrossRoad,在地图上查找相关省份城市下的相关交叉路口信息,

比如表4所示出的信息。

[0093] 或者,也可以由用户输入相应的槽位数据。

[0094] S103:基于该标签及该语料数据,生成包括标注结果的数据集合。

[0095] 一种实施方式中,可以针对每份槽位数据,确定该份槽位数据对应的每种表达方式下的语料数据;基于所述每种表达方式下的语料数据对应的领域标签、意图标签和槽位标签,生成该份槽位数据对应的数据子集;得到包括每份槽位数据对应的数据子集的数据集合。

[0096] 以表1为例进行说明,表1中包括多种槽位数据;本实施方式中,一份槽位数据对应一个数据子集,该数据子集可以如表5所示:

[0097] 表5

[0098]

语料数据	领域标签	意图标签	带槽位标签的语料数据
去长安路口	公共导航	other	去<to>长安路口</to>
去一下长安路口	公共导航	other	去一下<to>长安路口</to>
我想去长安路口	公共导航	other	我想去<to>长安路口</to>
我想去一下长安路口	公共导航	other	我想去一下<to>长安路口</to>
我们去长安路口	公共导航	other	我们去<to>长安路口</to>
我们去一下长安路口	公共导航	other	我们去一下<to>长安路口</to>
出发去长安路口	公共导航	other	出发去<to>长安路口</to>
出发去一下长安路口	公共导航	other	出发去一下<to>长安路口</to>
导航去长安路口	公共导航	other	导航去<to>长安路口</to>
导航去一下长安路口	公共导航	other	导航去一下<to>长安路口</to>
……	……	……	……

[0099] 表5可以理解为槽位数据“长安路口”对应的数据子集,表5中包括“长安路口”对应的各种表达方式下的语料数据。表5中,意图标签定义为“other”,或者也可以定义为其他内容,具体不做限定。其他槽位数据也可以对应表5类似的数据子集,这些数据子集组成数据集合。

[0100] 或者,另一种实施方式中,可以针对每种表达方式,确定该种表达方式下的每份槽位数据对应的语料数据;基于所述每份槽位数据对应的语料数据的领域标签、意图标签和槽位标签,生成该种表达方式对应的数据子集;得到包括每种表达方式对应的数据子集的数据集合。

[0101] 以表1为例进行说明,表1中包括多种槽位数据;本实施方式中,一种表达方式对应一个数据子集,该数据子集可以如表6所示:

[0102] 表6

[0103]

语料数据	领域标签	意图标签	带槽位标签的语料数据
去一下长安路口	公共导航	other	去一下<to>长安路口</to>
去一下长安街与建国路交叉	公共导航	other	去一下<to>长安街与建国路交叉</to>
去一下大望路与建国路交叉点	公共导航	other	去一下<to>大望路与建国路交叉点</to>
去一下大望路与建国路十字路口	公共导航	other	去一下<to>大望路与建国路十字路口 </to>
去一下大望路与建国路交叉口	公共导航	other	去一下<to>大望路与建国路交叉口</to>
去一下大望路与建国路道路交叉口	公共导航	other	去一下<to>大望路与建国路道路交叉口</to>
……	……	……	……

[0104] 表6可以理解为表达方式“去一下……”对应的数据子集,表6中包括“去一下……”对应的各种槽位数据对应的的语料数据。表6中,意图标签定义为“other”,或者也可以定义为其他内容,具体不做限定。表达方式可以包括:“我们去一下……”、“导航去……”等等,也可以理解为领域数据与意图数据的任意组合。其他表达方式也可以对应表6类似的数据子集,这些数据子集组成数据集合。

[0105] 举例来说,可以采用上述任意一种实施方式(表5或表6对应的实施方式)生成数据集合。或者,一种情况下,也可以先判断表达方式的数量是否大于槽位数据的数量;如果大于,采用表5对应的实施方式生成数据集合;如果不大于,采用表6对应的实施方式生成数据集合。

[0106] 具体来说,也就是判断表达方式的数量是否大于槽位数据的数量;

[0107] 如果大于,针对每份槽位数据,确定该份槽位数据对应的每种表达方式下的语料数据;基于所述每种表达方式下的语料数据对应的领域标签、意图标签和槽位标签,生成该份槽位数据对应的数据子集;得到包括每份槽位数据对应的数据子集的数据集合;

[0108] 如果不大于,针对每种表达方式,确定该种表达方式下的每份槽位数据对应的语料数据;基于所述每份槽位数据对应的语料数据的领域标签、意图标签和槽位标签,生成该种表达方式对应的数据子集;得到包括每种表达方式对应的数据子集的数据集合。

[0109] 这种情况下,如果表达方式较多,则一份槽位数据对应一个数据子集,数据子集中

包括该份槽位数据对应的每种表达方式下的语料数据、以及领域标签、意图标签和槽位标签;如果槽位数据较多,则一种表达方式对应一个数据子集,数据子集中包括该种表达方式对应的每份槽位数据、以及领域标签、意图标签和槽位标签。这样,数据子集的数量会较少,但数据子集中包括的信息量会较大,这种数据集合更有利于后续检索、训练模型等等。

[0110] 生成的数据集合为标注过的数据,可以利用该数据集合进行NLU(Natural Language Processing,自然语言处理)模型训练、或者进行其他识别模型的训练,具体不做限定。

[0111] 上述内容中,以简体中文为例对本发明实施例进行了介绍,此外,利用本发明实施例还可以对英文、繁体中文或者其他语言进行处理,生成其对应的数据集合。本发明实施例并不对语言类型进行限定。

[0112] 应用本发明图1所示实施例,获取正则表达式格式的文本数据;基于正则表达式,将文本数据组合成语料数据,并得到语料数据的以下任意一种或多种标签:领域、意图、槽位;基于标签及语料数据,生成包括标注结果的数据集合;本方案中,文本数据符合正则表达式,对于设备来说,其可以基于该正则表达式将文本数据组合成语料数据,并得到相应的标签,这样,相比于人工标注,节省了较多人力。

[0113] 下面参考图2介绍一种具体的实施方式:

[0114] S201:预先定义正则表达式:

[0115]

[]	其中的内容可以选其一也可为空,不同内容用“|”分隔开
()	其中的内容必可以选其一并且不可为空,不同内容用“|”分隔开
<>	代表内容需要根据实际需要调整
{}	其中的内容由多个可替换的变量组成,且组成方式比较多样

[0116] S202:用户定义领域类型和意图类型。

[0117] 假设用户定义的领域类型为“导航”。

[0118] 一些情况下,可以省略“用户定义意图类型”这一步骤。或者,用户也可以定义意图类型为“other”,表示没有特定意义的意图。

[0119] S203:用户输入正则表达式格式的文本数据:

[0120] 假设用户输入的文本数据为:[我想|我们|出发|导航](去|去一下)<CrossRoad>;其中,[我想|我们|出发|导航]表示:可以在“我想|我们|出发|导航”中任选一个,或者也可以为空;(去|去一下)表示:可以在“去|去一下”中任选一个,并且不可以为空;<CrossRoad>表示交叉路口,具体内容可以根据实际内容进行替换。

[0121] S204:为槽位数据添加槽位标签。

[0122] 举例来说,可以由用户添加槽位标签。延续上述例子,可以对<CrossRoad>添加《to》《/to》的标签,to表示去往某个地方。可以针对不同内容添加不同的标签,比如,时间标签可以为《time》《/time》等等,不再一一列举。

[0123] 这样,电子设备便得到了正则表达式格式的文本数据:[我想|我们|出发|导航](去|去一下)《to》<CrossRoad>《/to》;其中,《to》<CrossRoad>《/to》表示对<CrossRoad>添加了槽位标签。

[0124] 对于电子设备来说,便可以基于用户定义的领域类型,得到语料数据的领域标签;基于用户定义的意图类型,得到语料数据的意图标签;识别文本数据中的槽位标签。这样,

电子设备便得到了领域标签、意图标签和槽位标签。

[0125] S205:将识别出的多份领域数据、多份意图数据和多份槽位数据进行组合,得到语料数据。

[0126] 延续上述例子,电子设备得到了正则表达式格式的文本数据:[我想|我们|出发|导航](去|去一下)《to》<CrossRoad>《/to》;电子设备识别出各标号区间:[我想|我们|出发|导航]、(去|去一下)、<CrossRoad>;针对[我想|我们|出发|导航],基于分隔符“|”,识别出多份领语料子数据为“我想”、“我们”、“出发”“导航”。针对(去|去一下),基于分隔符“|”,识别出多份语料子数据为“去”、“去一下”。

[0127] 另外,CrossRoad可以表示变量参数,其指代的具体内容可以如上述表1所示。

[0128] 将识别到的多份语料子数据、以及该多份槽位数据组合成语料数据,比如:我想去大望路与建国路十字路口、我想去长安路口、我们去一下长安路口、导航去大望路与建国路交叉口……可以基于上述正则表达式的定义,进行任意组合,不再一一列举。

[0129] S206:针对每份槽位数据,生成该份槽位数据对应的数据子集;得到包括每份槽位数据对应的数据子集的数据集合;或者,针对每种表达方式,生成该种表达方式对应的数据子集;得到包括每种表达方式对应的数据子集的数据集合。

[0130] 本步骤中,槽位标签仅用于对槽位数据进行标注,而领域标签和意图标签用于对整个语料数据进行标注。

[0131] 一种实施方式中,可以针对每份槽位数据,确定该份槽位数据对应的每种表达方式下的语料数据;基于所述每种表达方式下的语料数据对应的领域标签、意图标签和槽位标签,生成该份槽位数据对应的数据子集;得到包括每份槽位数据对应的数据子集的数据集合。

[0132] 以表1为例进行说明,表1中包括多种槽位数据;本实施方式中,一份槽位数据对应一个数据子集,该数据子集可以如上述表5所示。

[0133] 表5可以理解为槽位数据“长安路口”对应的数据子集,表5中包括“长安路口”对应的各种表达方式下的语料数据。表5中,意图标签定义为“other”,或者也可以定义为其他内容,具体不做限定。其他槽位数据也可以对应表5类似的数据子集,这些数据子集组成数据集合。

[0134] 或者,另一种实施方式中,可以针对每种表达方式,确定该种表达方式下的每份槽位数据对应的语料数据;基于所述每份槽位数据对应的语料数据的领域标签、意图标签和槽位标签,生成该种表达方式对应的数据子集;得到包括每种表达方式对应的数据子集的数据集合。

[0135] 以表1为例进行说明,表1中包括多种槽位数据;本实施方式中,一种表达方式对应一个数据子集,该数据子集可以如上述表6所示。

[0136] 表6可以理解为表达方式“去一下……”对应的数据子集,表6中包括“去一下……”对应的各种槽位数据对应的的语料数据。表6中,意图标签定义为“other”,或者也可以定义为其他内容,具体不做限定。表达方式可以包括:“我们去一下……”、“导航去……”等等,也可以理解为领域数据与意图数据的任意组合。其他表达方式也可以对应表6类似的数据子集,这些数据子集组成数据集合。

[0137] 举例来说,可以采用上述任意一种实施方式(表5或表6对应的实施方式)生成数据

集合。或者,一种情况下,也可以先判断表达方式的数量是否大于槽位数据的数量;如果大于,采用表5对应的实施方式生成数据集合;如果不大于,采用表6对应的实施方式生成数据集合。

[0138] 具体来说,也就是判断表达方式的数量是否大于槽位数据的数量;

[0139] 如果大于,针对每份槽位数据,确定该份槽位数据对应的每种表达方式下的语料数据;基于所述每种表达方式下的语料数据对应的领域标签、意图标签和槽位标签,生成该份槽位数据对应的数据子集;得到包括每份槽位数据对应的数据子集的数据集合;

[0140] 如果不大于,针对每种表达方式,确定该种表达方式下的每份槽位数据对应的语料数据;基于所述每份槽位数据对应的语料数据的领域标签、意图标签和槽位标签,生成该种表达方式对应的数据子集;得到包括每种表达方式对应的数据子集的数据集合。

[0141] 这种情况下,如果表达方式较多,则一份槽位数据对应一个数据子集,数据子集中包括该份槽位数据对应的每种表达方式下的语料数据、以及领域标签、意图标签和槽位标签;如果槽位数据较多,则一种表达方式对应一个数据子集,数据子集中包括该种表达方式对应的每份槽位数据、以及领域标签、意图标签和槽位标签。这样,数据子集的数量会较少,但数据子集中包括的信息量会较大,这种数据集合更有利于后续检索、训练模型等等。

[0142] 与上述方法实施例相对应,本发明实施例还提供了一种标注数据获取装置,如图3所示,包括:

[0143] 获取模块301,用于获取正则表达式格式的文本数据;

[0144] 组合模块302,用于基于所述正则表达式,将所述文本数据组合成语料数据;

[0145] 获得模块303,用于得到所述语料数据的以下任意一种或多种标签:领域、意图、槽位;

[0146] 生成模块304,用于基于所述标签及所述语料数据,生成包括标注结果的数据集合。

[0147] 作为一种实施方式,所述正则表达式中:将可相互替换的多份内容填充于同一标号区间中,同一标号区间中的多份内容之间由分隔符隔开;

[0148] 组合模块302具体用于:识别所述文本数据中的标号区间;基于标号区间内的分隔符,识别标号区间内的多份语料子数据;将识别到的多份语料子数据组合成语料数据。

[0149] 作为一种实施方式,所述正则表达式中:将槽位数据对应的变量参数填充于预设标号区间中,所述变量参数指代多份槽位数据;

[0150] 组合模块302还用于将识别到的多份语料子数据、以及所述多份槽位数据组合成语料数据。

[0151] 作为一种实施方式,获得模块303具体用于:

[0152] 基于用户定义的领域类型,得到所述语料数据的领域标签;

[0153] 基于用户定义的意图类型,得到所述语料数据的意图标签;

[0154] 识别所述文本数据中的槽位标签。

[0155] 作为一种实施方式,生成模块304包括:第一生成子模块或者第二生成子模块(图中未示出),其中,

[0156] 所述第一生成子模块,用于针对每份槽位数据,确定该份槽位数据对应的每种表达方式下的语料数据;基于所述每种表达方式下的语料数据对应的领域标签、意图标签和

槽位标签,生成该份槽位数据对应的数据子集;得到包括每份槽位数据对应的数据子集的数据集合;

[0157] 所述第二生成子模块,用于针对每种表达方式,确定该种表达方式下的每份槽位数据对应的语料数据;基于所述每份槽位数据对应的语料数据的领域标签、意图标签和槽位标签,生成该种表达方式对应的数据子集;得到包括每种表达方式对应的数据子集的数据集合。

[0158] 作为一种实施方式,所述装置还包括:

[0159] 判断模块(图中未示出),用于判断表达方式的数量是否大于槽位数据的数量;如果大于,触发所述第一生成子模块;如果不大于,触发所述第二生成子模块。

[0160] 上述实施例的装置用于实现前述实施例中相应的方法,并且具有相应的方法实施例的有益效果,在此不再赘述。

[0161] 本发明实施例还提供了一种电子设备,如图4所示,包括存储器402、处理器401及存储在存储器402上并可在处理器401上运行的计算机程序,处理器401执行所述程序时实现上述任一种标注数据获取方法。

[0162] 本发明实施例还提供了一种非暂态计算机可读存储介质,所述非暂态计算机可读存储介质存储计算机指令,所述计算机指令用于使所述计算机执行上述任一种标注数据获取方法。

[0163] 所属领域的普通技术人员应当理解:以上任何实施例的讨论仅为示例性的,并非旨在暗示本公开的范围(包括权利要求)被限于这些例子;在本发明的思路下,以上实施例或者不同实施例中的技术特征之间也可以进行组合,步骤可以以任意顺序实现,并存在如上所述的本发明的不同方面的许多其它变化,为了简明它们没有在细节中提供。

[0164] 另外,为简化说明和讨论,并且为了不会使本发明难以理解,在所提供的附图中可以示出或可以不示出与集成电路(IC)芯片和其它部件的公知的电源/接地连接。此外,可以以框图的形式示出装置,以便避免使本发明难以理解,并且这也考虑了以下事实,即关于这些框图装置的实施方式的细节是高度取决于将要实施本发明的平台的(即,这些细节应当完全处于本领域技术人员的理解范围内)。在阐述了具体细节(例如,电路)以描述本发明的示例性实施例的情况下,对本领域技术人员来说显而易见的是,可以在没有这些具体细节的情况下或者这些具体细节有变化的情况下实施本发明。因此,这些描述应被认为是说明性的而不是限制性的。

[0165] 尽管已经结合了本发明的具体实施例对本发明进行了描述,但是根据前面的描述,这些实施例的很多替换、修改和变型对本领域普通技术人员来说将是显而易见的。例如,其它存储器架构(例如,动态RAM(DRAM))可以使用所讨论的实施例。

[0166] 本发明的实施例旨在涵盖落入所附权利要求的宽泛范围之内的所有这样的替换、修改和变型。因此,凡在本发明的精神和原则之内,所做的任何省略、修改、等同替换、改进等,均应包含在本发明的保护范围之内。

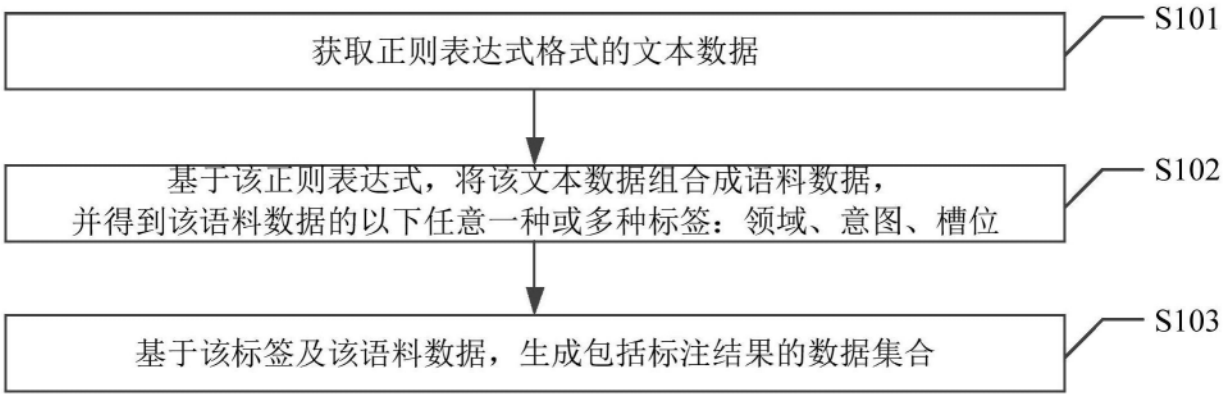

获取正则表达式格式的文本数据
S101
基于该正则表达式，将该文本数据组合成语料数据，
并得到该语料数据的以下任意一种或多种标签：领域、意图、槽位
S102
基于该标签及该语料数据，生成包括标注结果的数据集合
S103

图1

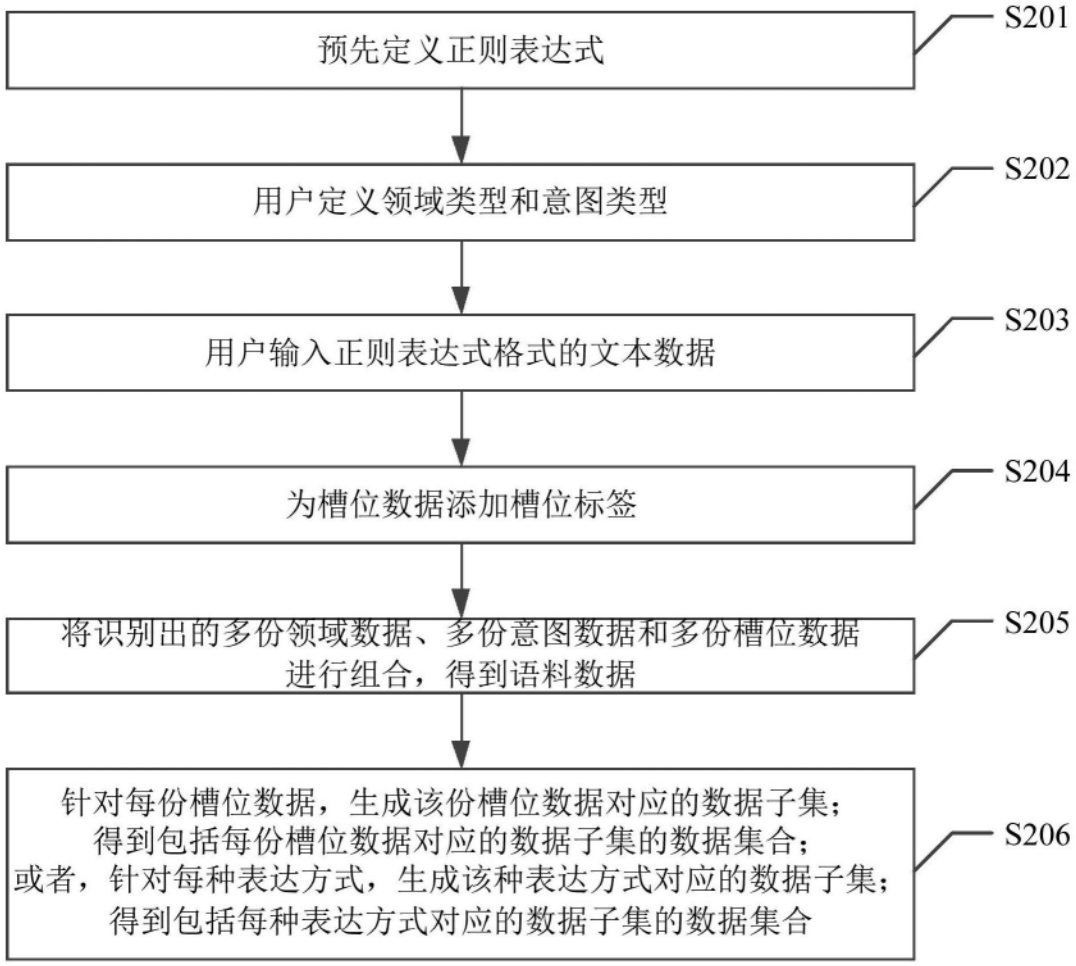

预先定义正则表达式
S201
用户定义领域类型和意图类型
S202
用户输入正则表达式格式的文本数据
S203
为槽位数据添加槽位标签
S204
将识别出的多份领域数据、多份意图数据和多份槽位数据
进行组合，得到语料数据
S205
针对每份槽位数据，生成该份槽位数据对应的数据子集；
得到包括每份槽位数据对应的数据子集的数据集合；
或者，针对每种表达方式，生成该种表达方式对应的数据子集；
得到包括每种表达方式对应的数据子集的数据集合
S206

图2

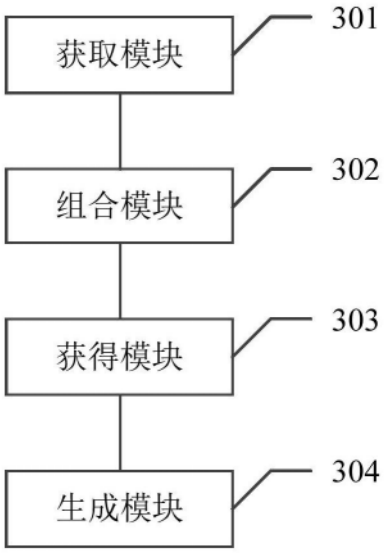


图3

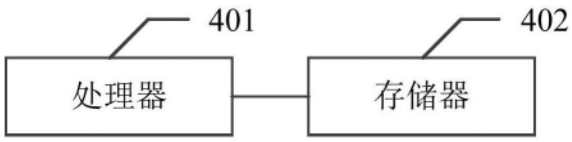


图4