(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112215005 A
(43)申请公布日 2021.01.12

(21)申请号 202011084006.3

(22)申请日 2020.10.12

(71)申请人 小红书科技有限公司
地址 200433 上海市杨浦区黄兴路2005弄2号(B楼)608-4室

(72)发明人 何永能

(74)专利代理机构 北京集佳知识产权代理有限公司 11227
代理人 钱娜

(51)Int.Cl.
G06F 40/295(2020.01)
G06F 40/30(2020.01)
G06K 9/62(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

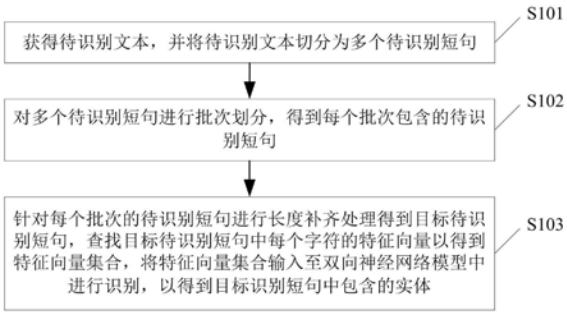
权利要求书3页 说明书9页 附图3页

(54)发明名称

实体识别方法及装置

(57)摘要

本申请提供了一种实体识别方法，该方法将待识别文本切分为多个待识别短句，将多个待识别短句进行批次划分，对同一批次内的待识别短句进行长度补齐处理，并从字向量词典中查找待识别短句中每个字符对应的特征向量，将特征向量输入至双向神经网络模型中以识别出待识别短句中包含的实体。由于双神经网络模型是由文本样本训练而成的，其考虑到文本样本内字符之间的语义关系，因此使用双向神经网络模型所识别的实体符合语义规则，识别准确度更高。另外，本申请还提供了实体识别的相关装置，用以保证所述方法在实际中的应用及实现。



CN 112215005 A

1.一种实体识别方法,其特征在于,包括:

获得待识别文本,并将所述待识别文本切分为多个待识别短句;

对多个所述待识别短句进行批次划分,得到每个批次包含的待识别短句;

针对每个批次的待识别短句,执行以下步骤:

将所述批次包含的多个待识别短句处理为相同长度的短句,以得到多个目标识别短句;

在预先生成的字向量词典中,查找所述目标识别短句中每个字符所对应的特征向量,以得到所述目标识别短句的特征向量集合;其中所述字向量词典中保存有字符对应的特征向量,具有语义关系的字符所对应的特征向量之间存在关联关系;

将多个所述目标识别短句的特征向量集合输入至预先训练完成的双向神经网络模型中,以使所述双向神经网络模型基于所述特征向量集合识别出各个所述目标识别短句中包含的实体。

2.根据权利要求1所述的实体识别方法,其特征在于,所述双向神经网络模型的训练过程包括:

获得具有实体类型标注的文本样本,并将所述文本样本切分为多个短句样本;

对多个所述短句样本进行批次划分,得到每个批次包含的短句样本;

针对各个批次的短句样本,执行以下处理步骤:

将所述批次包含的多个短句样本处理为相同长度的短句,以得到多个目标短句样本;

确定所述目标样本短句中每个字符的特征向量,以得到所述目标短句样本的特征向量集合;

使用双向神经网络训练算法,对多个所述目标短句样本的特征向量集合进行训练;

若训练结果满足预设损失函数的要求,则停止训练过程并得到训练完成的双向神经网络模型;若训练结果不满足预设损失函数的要求,则执行下一批次的短句样本的处理步骤。

3.根据权利要求2所述的实体识别方法,其特征在于,

若训练结果不满足预设损失函数的要求,在执行下一批次的短句样本的处理步骤之前,该方法还包括:对所述目标样本短句中每个字符的特征向量进行调整;

若训练结果满足预设损失函数的要求,则该方法还包括:

将所述训练结果对应的调整后的特征向量作为字符的特征向量;

保存所述文本样本中每个字符对应的特征向量,以生成字向量词典。

4.根据权利要求1所述的实体识别方法,其特征在于,所述将所述批次包含的多个待识别短句处理为相同长度的短句,包括:

确定所述批次包含的多个待识别短句中的最长短句,并计算所述最长短句的长度;

确定所述批次包含的待识别短句中不满足所述长度的短句,并在不满足所述长度的短句中添加预设的无意义字符,以将所述多个待识别短句处理为相同长度的短句。

5.根据权利要求1所述的实体识别方法,其特征在于,所述双向神经网络模型包括:双向长短期记忆网络层以及条件随机场层。

6.一种实体识别装置,其特征在于,包括:

待识别短句获取模块,用于获得待识别文本,并将所述待识别文本切分为多个待识别短句;

待识别短句划分模块，用于对多个所述待识别短句进行批次划分，得到每个批次包含的待识别短句；

待识别短句处理模块，用于针对每个批次的待识别短句，执行以下步骤：

将所述批次包含的多个待识别短句处理为相同长度的短句，以得到多个目标识别短句；

在预先生成的字向量词典中，查找所述目标识别短句中每个字符所对应的特征向量，以得到所述目标识别短句的特征向量集合；其中所述字向量词典中保存有字符对应的特征向量，具有语义关系的字符所对应的特征向量之间存在关联关系；

将多个所述目标识别短句的特征向量集合输入至预先训练完成的双向神经网络模型中，以使所述双向神经网络模型基于所述特征向量集合识别出各个所述目标识别短句中包含的实体。

7.根据权利要求6所述的实体识别装置，其特征在于，还包括：训练模块，用于训练双向神经网络模型；

所述训练模块包括：

样本短句获取子模块，用于获得具有实体类型标注的文本样本，并将所述文本样本切分为多个短句样本；

样本短句划分子模块，用于对多个所述短句样本进行批次划分，得到每个批次包含的短句样本；

样本短句处理子模块，用于针对各个批次的短句样本，执行以下处理步骤：

将所述批次包含的多个短句样本处理为相同长度的短句，以得到多个目标短句样本；

确定所述目标样本短句中每个字符的特征向量，以得到所述目标短句样本的特征向量集合；

使用双向神经网络训练算法，对多个所述目标短句样本的特征向量集合进行训练；

训练结果判断子模块，用于若训练结果满足预设损失函数的要求，则停止训练过程并得到训练完成的双向神经网络模型；若训练结果不满足预设损失函数的要求，则执行下一批次的短句样本的处理步骤。

8.根据权利要求6所述的实体识别装置，其特征在于，所述待识别短句处理模块用于将所述批次包含的多个待识别短句处理为相同长度的短句，具体包括：

待识别短句处理模块，用于确定所述批次包含的多个待识别短句中的最长短句，并计算所述最长短句的长度；以及确定所述批次包含的待识别短句中不满足所述长度的短句，并在不满足所述长度的短句中添加预设的无意义字符，以将所述多个待识别短句处理为相同长度的短句。

9.一种实体识别设备，其特征在于，包括处理器和存储器，所述处理器通过运行存储在所述存储器内的软件程序、调用存储在所述存储器内的数据，至少执行如下步骤：

获得待识别文本，并将所述待识别文本切分为多个待识别短句；

对多个所述待识别短句进行批次划分，得到每个批次包含的待识别短句；

针对每个批次的待识别短句，执行以下步骤：

将所述批次包含的多个待识别短句处理为相同长度的短句，以得到多个目标识别短句；

在预先生成的字向量词典中，查找所述目标识别短句中每个字符所对应的特征向量，以得到所述目标识别短句的特征向量集合；其中所述字向量词典中保存有字符对应的特征向量，具有语义关系的字符所对应的特征向量之间存在关联关系；

将多个所述目标识别短句的特征向量集合输入至预先训练完成的双向神经网络模型中，以使所述双向神经网络模型基于所述特征向量集合识别出各个所述目标识别短句中包含的实体。

10.一种存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时，实现如权利要求1-5任意一项所述的实体识别方法。

实体识别方法及装置

技术领域

[0001] 本申请涉及文本处理技术领域，更具体地，是实体识别方法及装置。

背景技术

[0002] 随着互联网技术的发展，越来越多的人习惯于使用互联网应用平台，来获取信息内容。信息内容的一种形式为文本，用户可以从应用平台上获取文本内容，另外应用平台也会为用户推送文本内容，这类应用平台也可以称为内容分发系统。

[0003] 内容分发系统中存储有大量的文本内容，系统需要对文本内容进行识别，以识别文本内容中包含有哪些用户可能感兴趣的实体，识别结果可以作为向用户分发文本内容的依据。例如，如果内容分发系统识别出某条文本内容中包含有“飞利浦刮胡刀”这个实体，便可以将该文本内容推送给对该实体感兴趣的用户。

[0004] 目前的实体识别方法是，建立包含实体词的词库，将文本内容与词库中的实体词进行匹配，以识别文本内容中包含的实体。这种实体识别方法的识别准确度较低。

发明内容

[0005] 有鉴于此，本申请提供了一种实体识别方法，用以提高现有技术中实体方法的识别准确度。另外，本申请还提供了实体识别的相关装置，用以保证所述方法在实际中的应用及实现。

[0006] 为实现所述目的，本申请提供的技术方案如下：

[0007] 第一方面，本发明提供了一种实体识别方法，包括：

[0008] 获得待识别文本，并将所述待识别文本切分为多个待识别短句；

[0009] 对多个所述待识别短句进行批次划分，得到每个批次包含的待识别短句；

[0010] 针对每个批次的待识别短句，执行以下步骤：

[0011] 将所述批次包含的多个待识别短句处理为相同长度的短句，以得到多个目标识别短句；

[0012] 在预先生成的字向量词典中，查找所述目标识别短句中每个字符所对应的特征向量，以得到所述目标识别短句的特征向量集合；其中所述字向量词典中保存有字符对应的特征向量，具有语义关系的字符所对应的特征向量之间存在关联关系；

[0013] 将多个所述目标识别短句的特征向量集合输入至预先训练完成的双向神经网络模型中，以使所述双向神经网络模型基于所述特征向量集合识别出各个所述目标识别短句中包含的实体。

[0014] 第二方面，本发明提供了一种实体识别装置，包括：

[0015] 待识别短句获取模块，用于获得待识别文本，并将所述待识别文本切分为多个待识别短句；

[0016] 待识别短句划分模块，用于对多个所述待识别短句进行批次划分，得到每个批次包含的待识别短句；

[0017] 待识别短句处理模块，用于针对每个批次的待识别短句，执行以下步骤：

[0018] 将所述批次包含的多个待识别短句处理为相同长度的短句，以得到多个目标识别短句；

[0019] 在预先生成的字向量词典中，查找所述目标识别短句中每个字符所对应的特征向量，以得到所述目标识别短句的特征向量集合；其中所述字向量词典中保存有字符对应的特征向量，具有语义关系的字符所对应的特征向量之间存在关联关系；

[0020] 将多个所述目标识别短句的特征向量集合输入至预先训练完成的双向神经网络模型中，以使所述双向神经网络模型基于所述特征向量集合识别出各个所述目标识别短句中包含的实体。

[0021] 第三方面，本申请提供了一种实体识别设备，包括处理器和存储器，所述处理器通过运行存储在所述存储器内的软件程序、调用存储在所述存储器内的数据，至少执行如下步骤：

[0022] 获得待识别文本，并将所述待识别文本切分为多个待识别短句；

[0023] 对多个所述待识别短句进行批次划分，得到每个批次包含的待识别短句；

[0024] 针对每个批次的待识别短句，执行以下步骤：

[0025] 将所述批次包含的多个待识别短句处理为相同长度的短句，以得到多个目标识别短句；

[0026] 在预先生成的字向量词典中，查找所述目标识别短句中每个字符所对应的特征向量，以得到所述目标识别短句的特征向量集合；其中所述字向量词典中保存有字符对应的特征向量，具有语义关系的字符所对应的特征向量之间存在关联关系；

[0027] 将多个所述目标识别短句的特征向量集合输入至预先训练完成的双向神经网络模型中，以使所述双向神经网络模型基于所述特征向量集合识别出各个所述目标识别短句中包含的实体。

[0028] 第四方面，本申请提供了一种存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时，实现任意一项所述的实体识别方法。

[0029] 由以上技术方案可知，本申请提供了一种实体识别方法，该方法将待识别文本切分为多个待识别短句，将多个待识别短句进行批次划分，对同一批次内的待识别短句进行长度补齐处理，并从字向量词典中查找待识别短句中每个字符对应的特征向量，将特征向量输入至双向神经网络模型中以识别出待识别短句中包含的实体。由于双神经网络模型是由文本样本训练而成的，其考虑到文本样本内字符之间的语义关系，因此使用双向神经网络模型所识别的实体符合语义规则，识别准确度更高。

附图说明

[0030] 为了更清楚地说明本申请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申请的实施例，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，还可以根据提供的附图获得其他的附图。

[0031] 图1为本申请提供的实体识别方法的一个流程示意图；

[0032] 图2为本申请提供的双向神经网络训练过程的一个流程示意图；

[0033] 图3为本申请提供的实体识别装置的一个结构框图；

[0034] 图4为本申请提供的双向神经网络的训练模块的一个结构框图；

[0035] 图5为本申请提供的实体识别设备的一种具体结构图。

具体实施方式

[0036] 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，都属于本申请保护的范围。

[0037] 内容分发是互联网应用中一个重要的功能，很多内容展示类互联网应用都具有这项功能，用户使用这类互联网应用时，可以浏览到系统自动推送的信息内容。

[0038] 该类互联网应用后台设置有内容分发系统，内容分发系统需要对大量的文本内容进行识别，识别出文本内容中包含的实体。实体一般为名词词性的对象，如地点、产品、时间、人名、地名等等。需要说明的是，实体是类型下的具体对象，例如童话书不是实体，安徒生童话是实体；又如刮胡刀不是实体，飞利浦刮胡刀是实体。

[0039] 目前的实体识别方法依赖于词库，具体是将文本内容与词库中的实体词语进行匹配，这种方法容易将歧义词识别为错误结果，例如词库中包含有苹果这个实体词，该实体词的本身含义为水果苹果，假设某个描述苹果手机的文本中包含“苹果”这个词，在该文本的语义中的词语“苹果”表示的是手机，词库却将该词语识别为水果苹果。

[0040] 为了提高文本中实体的识别准确率，本申请提供了一种实体识别方法，如图1所示，该实体识别方法包括步骤S101～S103。

[0041] S101：获得待识别文本，并将待识别文本切分为多个待识别短句。

[0042] 其中，在不同的应用场景中，待识别文本可以是不同内容的文本，例如该方法应用在微博这个应用平台上，则待识别文本为微博文章，又如该方法应用在知乎这个应用平台上，则待识别文本为知乎文章。

[0043] 待识别文本内容通常较多，需要将待识别文本切分为长度较短的短句，然后对待识别短句进行实体识别。一种切分方式可以是，按照标点符号进行切分。具体地，在具有结束句子意义的终止类标点符号处切分待识别文本，其中终止类标点符号可以包括句号、感叹号、问号等等。另外如果中间类标点符号所包含的文本内容长度超过预设长度，则在中间类标点符号处也切分待识别文本，其中中间类标点符号可以包括逗号、分号等等，需要说明的是，目前用户的一种书写习惯是，不使用标点符号，而是使用空格代替，因此在本申请中可以将空格等同为逗号。

[0044] S102：对多个待识别短句进行批次划分，得到每个批次包含的待识别短句。

[0045] 其中，经过切分处理之后，待识别文本可以被切分为多个待识别短句。这些待识别短句需要分批次输入至双向神经网络模型中进行识别，因此需要将这些待识别短句划分为多个批次。每个批次中包含的待识别短句的个数可以在实际应用中设置，例如可以设置为32个待识别短句为一个批次。

[0046] S103：针对每个批次的待识别短句进行长度补齐处理得到目标待识别短句，查找目标待识别短句中每个字符的特征向量以得到特征向量集合，将特征向量集合输入至双向

神经网络模型中进行识别，以得到目标识别短句中包含的实体。

[0047] 具体地，针对每个批次的待识别短句，执行以下步骤A1～A3。

[0048] A1：将批次包含的多个待识别短句处理为相同长度的短句，以得到多个目标识别短句。

[0049] 其中，每个批次中包含多个待识别短句，待识别短句的长度不一定相同，需要将这些待识别短句的长度补齐为相同长度。为了进行区分，经过长度处理的待识别短句可以称为目标识别短句。

[0050] 一种补齐方式为，将同一批次中的待识别短句的长度补齐为预设的某个长度，例如预设长度为200，则同一批次中的待识别短句的长度全部补齐为200个字符。补齐时需要在待识别短句中添加无意义的字符。这种补齐方式可能会导致双向神经网络模型识别效率较低且识别准确率较低，原因是，待识别短句的长度可能都远远低于预设长度，从而需要在待识别短句中添加较多数量的无意义字符，无意义的字符越多，对于双向神经网络模型来说，识别过程中需要处理的无效数据就越多，因此不仅降低了识别结果的准确度，而且降低了识别效率。

[0051] 另一种补齐方式为，确定批次包含的多个待识别短句中的最长短句，并计算最长短句的长度；以及确定批次包含的待识别短句中不满足长度的短句，并在不满足长度的短句中添加预设的无意义字符，以将多个待识别短句处理为相同长度的短句。

[0052] 例如，同一批次中包含有32个待识别短句，最长的待识别短句的长度为20，其余待识别短句的长度均小于该最大长度，从而将其余待识别短句的长度均补齐为20。具体的补齐方式是，在待识别短句中添加无具体语义的字符，如添加特殊符号，特殊符号例如可以是为“0”。无具体语义字符的添加位置可以是待识别短句中两个字符中间。例如，待识别短句为“基本跟zara在国内的地位一样”，该待识别短句的长度为15，则需要添加5个无具体语义字符，添加后的待识别短句为“基0本0跟0 zara在0国0内的地位一样”。

[0053] 可见，上述补齐方式是将同一批次中的多个待识别短句都补齐为这一批次中最长待识别短句的长度，可以最大限度地减少向待识别短句中添加的无意义字符的数量，进而可以提高双向神经网络模型的识别效率以及识别准确度。

[0054] A2：在预先生成的字向量词典中，查找目标识别短句中每个字符所对应的特征向量，以得到目标识别短句的特征向量集合；其中字向量词典中保存有字符对应的特征向量，具有语义关系的字符所对应的特征向量之间存在关联关系。

[0055] 其中，在实施前会预先设置有字向量词典，字向量词典中包含的是字符与特征向量之间的对应关系。特征向量表示的是字符与字符之间的语义关系，因此具有语义关系的字符所对应的特征向量之间存在关联关系。特征向量包含多个维度的特征，如果字符之间具有语义关系，则关联关系体现在特征向量的某些维度之间具有关联关系。

[0056] 例如，字向量词典中包含有中文字符“美”、“丽”、“国”、“宝”及“莲”，可以知道的是“美”和“丽”(美丽)、“美”和“国”(美国)、以及“美”“宝”及“莲”(美宝莲)之间都存在语义关系，因此“美”的特征向量中某些维度上的特征与“丽”的特征向量中某些维度上的特征之间具有关联关系，“美”的特征向量中某些维度上的特征与“国”的特征向量中某些维度上的特征之间具有关联关系，以及“美”的特征向量中某些维度上的特征、与“宝”及“莲”的特征向量中某些维度上的特征之间具有关联关系。

[0057] 字向量词典是由双向神经网络模型训练算法对大量的文本样本数据训练得到的，特征向量中每个维度的特征所表示的含义，与具体的双向神经网络模型训练算法有关，本申请并不具体限定。

[0058] 需要说明的是，字向量词典中记录的是单个字符的特征向量，而非由字符组成的词语的特征向量，由于同一字符可以和不同的字符组合为词语，相较于词语而言字符数量少很多，因此字向量词典的训练速度较快且训练效果也较好。

[0059] 另外，字向量词典中还可以包含有每个字符的索引(index)，索引表示的字符在字向量词典中的位置。例如，字符['奔','驰','兰','博','法','拉','利','宾','利','超','经','典','款']的索引为[60,101,48,50,78,74,12,63,12,93,91,49,19]。索引的存在可以提高字向量词典的训练效率。

[0060] 获得预先训练生成的字向量词典后，在字向量词典中查找目标识别短句中每个字符所对应的特征向量。例如，目标识别短句为“基0本0跟0 zara在0国0内的地位一样”，则查找“基”“本”“跟”“z”“a”“r”“a”“在”“国”“内”“的”“地”“位”“一”“样”分别所对应的特征向量。假设特性向量为200维度，则可以得到每个字符所对应的200维度的特征向量。

[0061] 由于目标识别短句中包含有多个字符，每个字符都会有对应的特征向量，从而可以得到目标识别短句所对应的一个特征向量集合，可以理解的是，目标识别短句中包含有多个字符，则特征向量集合中便包含有多少个特性向量。

[0062] A3:将多个目标识别短句的特征向量集合输入至预先训练完成的双向神经网络模型中，以使双向神经网络模型基于特征向量集合识别出各个目标识别短句中包含的实体。

[0063] 其中，双向神经网络模型是一种实体识别模型，其可以使用文本样本预先训练完成，训练完成的双向神经网络模型可以对任何一个待识别文本进行实体识别。

[0064] 在一个具体示例中，双向神经网络模型可以包括两层，一层为Bi-lstm(Long Short-Term Memory，LSTM)层，即双向长短期记忆网络层，一层为条件随机场层(Conditional Random Field，CRF)。其中第一层自动提取输入短句的特征，将一个短句的各个字符的向量序列作为双向LSTM各个时间步的输入，再将正向LSTM输出的隐状态序列与反向LSTM在各个位置输出的隐状态进行按位置拼接得到完整的隐状态序列，在设置损失函数后(随即丢弃一些训练数据，防止过拟合)后接入一个线性层，得到自动提取的句子特征，这样把每一维都视作将字分类到第j个标签的打分值接下来再接入一个CRF层来进行标注。第二层进行句子级的序列标注，具体地，CRF层的参数是一个矩阵A，Aij代表从第i个标签到第j个标签的转移得分，进而在为一个位置进行标注的时候可以利用此前已经标注过的标签，可以看出整个序列的打分等于各个位置的打分之和，而每个位置的打分由两部分得到，一部分是由LSTM输出的，另一部分则由CRF的转移矩阵A决定。转移矩阵A代表的是每个实体标注互相转移的概率值，即前一个字符是某种实体标注，后一个字符是某种实体标注的概率。例如，前一个字符是商品开头词，后一个字符是商品中间词的概率；又如前一个字符是商品中间词，后一个字符不是实体的概率。

[0065] 经过双向神经模型的计算之后，便可以标注出目标识别短句中所包含的实体。标注内容不仅包含哪些字符表示了实体，且可以标注实体的具体类型。也可以是说，从标注中不仅可以确定目标识别短句中是否包含有实体，还可以确定实体的具体类型。

[0066] 例如，输入的目标识别短句为“基0本0跟0 zara在0国0内的地位一样”，输出识别

结果为:基本跟z B_COM_a I_COM_r I_COM_a E_COM_在国内的地位一样。其中,zara四个字符被添加了标注信息,标注信息中的“B”表示实体的开始字符,即从字符“z”开始;标注信息中的“I”表示实体的中间字符,即包括字符“a”及字符“r”;标注信息中的“E”表示实体的结束字符,即在字符“a”结束。标注信息中的“COM”表示实体的类型为商品。可见,从标注信息中可以看出该目标识别短句中包含有zara这个商品类型的实体。

[0067] 需要说明的是,一个待识别文本可以被切分为待识别短句,待识别短句可以划分为多个批次,每个批次中的待识别短句都按照上述步骤A1-A3的方式进行实体识别,所有批次均识别完成后便可以得到待识别文本中所包含的实体。

[0068] 由以上技术方案可知,本申请提供了一种实体识别方法,该方法将待识别文本切分为多个待识别短句,将多个待识别短句进行批次划分,对同一批次内的待识别短句进行长度补齐处理,并从字向量词典中查找待识别短句中每个字符对应的特征向量,将特征向量输入至双向神经网络模型中以识别出待识别短句中包含的实体。由于双神经网络模型是由文本样本训练而成的,其考虑到文本样本内字符之间的语义关系,因此使用双向神经网络模型所识别的实体符合语义规则,识别准确度更高。

[0069] 以下对双向神经网络模型的训练过程进行具体说明。如图2所示,双向神经网络模型的训练过程可以具体包括步骤S201-S204。

[0070] S201:获得具有实体类型标注的文本样本,并将文本样本切分为多个短句样本。

[0071] 其中,预先获得大量的文本样本,文本样本需要由人工进行标注,标注出文本样本中所包含的实体以及实体类型是什么。

[0072] 例如,将每个字符标注为“B-X”、“I-X”或者“O”。其中,“B-X”表示此字符所在的实体属于X类型并且此字符在此实体的开头位置,“I-X”表示此字符所在的实体属于X类型并且此字符在此实体的中间位置,“O”表示此字符不属于任何实体类型。X可以是预先设置的任何类型的实体类型,比如,将X可以为商品(COM),则标记可以包括:B-COM(商品实体的开头)、I-COM(商品实体的中间)、E-COM(商品实体的结束)。实体类型可以根据实际应用场景的业务需求而设置。

[0073] 需要说明的是,想要训练完成的双向神经网络模型能够识别出哪些类型的实体,则可以收集包含相应实体类型的文本样本对双向神经网络模型进行训练,可以理解的是,在训练之前需要在文本样本中对这些实体类型进行标注。在实际应用中,所标注的实体类型可以是书名、电影名、明星人物等等任何想要识别的类型。

[0074] 将文本样本进行切分,从而得到多个短句样本。切分方式可以参照上述关于待识别短句的切分方式,此处并不赘述。

[0075] S202:对多个短句样本进行批次划分,得到每个批次包含的短句样本。

[0076] S203:针对各个批次的短句样本进行长度补齐处理得到目标短句样本,查找目标短句样本中每个字符的特征向量以得到特征向量集合,使用双向神经网络训练算法,对多个目标短句样本的特征向量集合进行训练。

[0077] 具体地,针对各个批次的短句样本执行以下处理步骤B1～B3。

[0078] B1:将批次包含的多个短句样本处理为相同长度的短句,以得到多个目标短句样本。B2:确定目标样本短句中每个字符的特征向量,以得到目标短句样本的特征向量集合。B3:使用双向神经网络训练算法,对多个目标短句样本的特征向量集合进行训练。

[0079] 其中，关于步骤S202及步骤S203的说明可以参照步骤S102及步骤S103中的说明，此处并不赘述。不同的是，首个批次的短句样本中每个字符的特征向量是随机初始化的，其后每个批次的短句样本中每个字符的特征向量是根据前一批次的短句样本的训练结果调整的特征向量。

[0080] S204：若训练结果满足预设损失函数的要求，则停止训练过程并得到训练完成的双向神经网络模型；若训练结果不满足预设损失函数的要求，则执行下一批次的短句样本的处理步骤。

[0081] 其中，在每个批次的目标短句样本的训练结束后，需要使用预设损失函数对训练结果进行判断，如果训练结果不满足预设损失函数的要求，则说明训练未完成，需要调整训练模型中的参数，并根据调整后的参数对下一批次的目标短句样本进行重复训练，如果某个批次的目标短句样本的训练结果满足预设损失函数的要求，则说明训练完成，从而停止训练过程，并将训练模型中的参数记录下来，作为最终训练完成的双向神经网络模型中的参数。

[0082] 另外，若训练结果不满足预设损失函数的要求，在执行下一批次的短句样本的处理步骤之前，还包括：对目标样本短句中每个字符的特征向量进行调整。训练结果满足预设损失函数的要求，则将训练结果对应的调整后的特征向量作为字符的特征向量，并保存文本样本中每个字符对应的特征向量，以生成字向量词典。

[0083] 可见，通过上述方式可以训练得到双向神经网络模型，以及可以得到字向量词典。

[0084] 见图3，本发明实施例提供了一种实体识别装置，具体包括：待识别短句获取模块301、待识别短句划分模块302、待识别短句处理模块303。

[0085] 待识别短句获取模块301，用于获得待识别文本，并将待识别文本切分为多个待识别短句。

[0086] 待识别短句划分模块302，用于对多个待识别短句进行批次划分，得到每个批次包含的待识别短句。

[0087] 待识别短句处理模块303，用于针对每个批次的待识别短句，执行以下步骤：

[0088] 将批次包含的多个待识别短句处理为相同长度的短句，以得到多个目标识别短句。

[0089] 在预先生成的字向量词典中，查找目标识别短句中每个字符所对应的特征向量，以得到目标识别短句的特征向量集合；其中字向量词典中保存有字符对应的特征向量，具有语义关系的字符所对应的特征向量之间存在关联关系。

[0090] 将多个目标识别短句的特征向量集合输入至预先训练完成的双向神经网络模型中，以使双向神经网络模型基于特征向量集合识别出各个目标识别短句中包含的实体。

[0091] 在一个示例中，实体识别装置还可以包括训练模块，用于训练双向神经网络模型。见图4，训练模块的一种具体结构可以包括：样本短句获取子模块401、样本短句划分子模块402、样本短句处理子模块403及训练结果判断子模块404。

[0092] 样本短句获取子模块401，用于获得具有实体类型标注的文本样本，并将文本样本切分为多个短句样本；

[0093] 样本短句划分子模块402，用于对多个短句样本进行批次划分，得到每个批次包含的短句样本；

[0094] 样本短句处理子模块403,用于针对各个批次的短句样本,执行以下处理步骤:

[0095] 将批次包含的多个短句样本处理为相同长度的短句,以得到多个目标短句样本。

[0096] 确定目标样本短句中每个字符的特征向量,以得到目标短句样本的特征向量集合。

[0097] 使用双向神经网络训练算法,对多个目标短句样本的特征向量集合进行训练。

[0098] 训练结果判断子模块404,用于若训练结果满足预设损失函数的要求,则停止训练过程并得到训练完成的双向神经网络模型;若训练结果不满足预设损失函数的要求,则执行下一批次的短句样本的处理步骤。

[0099] 在一个示例中,训练结果判断子模块还包括:样本特征向量调整单元以及字向量词典生成单元。

[0100] 样本特征向量调整单元,用于若训练结果不满足预设损失函数的要求,在执行下一批次的短句样本的处理步骤之前,对所述目标样本短句中每个字符的特征向量进行调整;

[0101] 字向量词典生成单元,用于若训练结果满足预设损失函数的要求,则将所述训练结果对应的调整后的特征向量作为字符的特征向量,保存所述文本样本中每个字符对应的特征向量,以生成字向量词典。

[0102] 在一个示例中,待识别短句处理模块用于将所述批次包含的多个待识别短句处理为相同长度的短句,具体包括:

[0103] 待识别短句处理模块,用于确定所述批次包含的多个待识别短句中的最长短句,并计算所述最长短句的长度;以及确定所述批次包含的待识别短句中不满足所述长度的短句,并在不满足所述长度的短句中添加预设的无意义字符,以将所述多个待识别短句处理为相同长度的短句。

[0104] 在一个示例中,上述双向神经网络模型包括:双向长短期记忆网络层以及条件随机场层。

[0105] 由以上技术方案可知,本申请提供了一种实体识别装置,该装置包含待识别短句获取模块、待识别短句划分模块及待识别短句处理模块。待识别短句获取模块将待识别文本切分为多个待识别短句,待识别短句划分模块将多个待识别短句进行批次划分,待识别短句处理模块对同一批次内的待识别短句进行长度补齐处理,并从字向量词典中查找待识别短句中每个字符对应的特征向量,将特征向量输入至双向神经网络模型中以识别出待识别短句中包含的实体。由于双神经网络模型是由文本样本训练而成的,其考虑到文本样本内字符之间的语义关系,因此使用双向神经网络模型所识别的实体符合语义规则,识别准确度更高。

[0106] 见图5,其示出了本申请提供的一种实体识别设备的具体结构,包括:存储器501、处理器502及通信总线503。

[0107] 其中,存储器501、处理器502通过通信总线503完成相互间的通信。

[0108] 存储器501,用于存放程序;存储器501可能包含高速RAM存储器,也可能还包括非易失性存储器(non-volatile memory),例如至少一个磁盘存储器。

[0109] 处理器502,用于执行程序,程序可以包括程序代码,所述程序代码包括处理器的操作指令。其中,程序可具体用于:

[0110] 获得待识别文本，并将所述待识别文本切分为多个待识别短句；

[0111] 对多个所述待识别短句进行批次划分，得到每个批次包含的待识别短句；

[0112] 针对每个批次的待识别短句，执行以下步骤：

[0113] 将所述批次包含的多个待识别短句处理为相同长度的短句，以得到多个目标识别短句；

[0114] 在预先生成的字向量词典中，查找所述目标识别短句中每个字符所对应的特征向量，以得到所述目标识别短句的特征向量集合；其中所述字向量词典中保存有字符对应的特征向量，具有语义关系的字符所对应的特征向量之间存在关联关系；

[0115] 将多个所述目标识别短句的特征向量集合输入至预先训练完成的双向神经网络模型中，以使所述双向神经网络模型基于所述特征向量集合识别出各个所述目标识别短句中包含的实体。

[0116] 处理器502可能是一个中央处理器CPU，或者是特定集成电路ASIC(Application Specific Integrated Circuit)，或者是被配置成实施本申请实施例的一个或多个集成电路。

[0117] 需要说明的是，所述处理器可以执行与上述实体识别方法相关的各个步骤，此处并不赘述。

[0118] 本申请还提供了一种可读存储介质，其上存储有计算机程序，所述计算机程序可以被处理器执行，以实现以上各个实体识别方法实施例中的各个步骤。

[0119] 需要说明的是，本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

[0120] 还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，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、物品或者设备不仅包括那些要素，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、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括上述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0121] 对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本申请。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下，在其它实施例中实现。因此，本申请将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

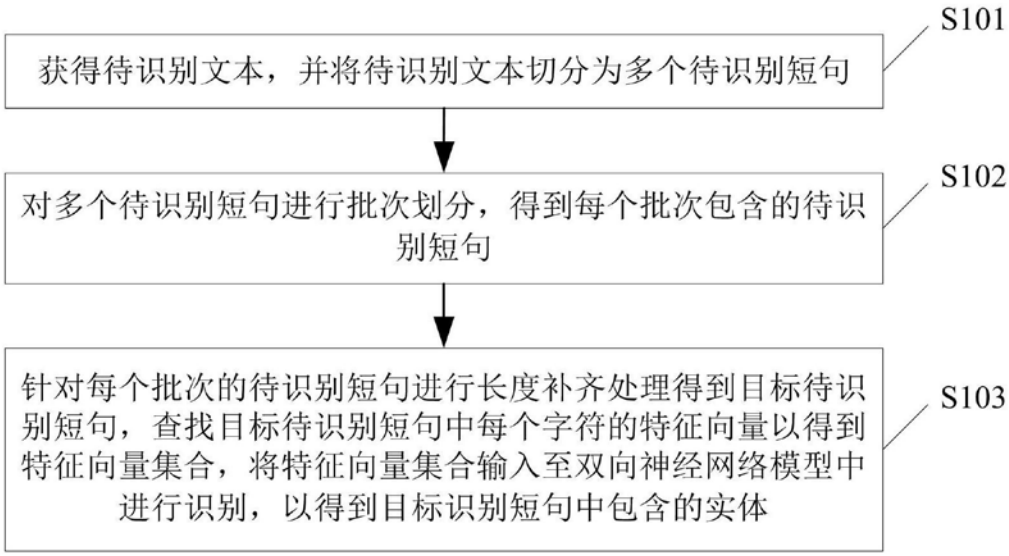

S101
获得待识别文本，并将待识别文本切分为多个待识别短句
S102
对多个待识别短句进行批次划分，得到每个批次包含的待识别短句
S103
针对每个批次的待识别短句进行长度补齐处理得到目标待识别短句，查找目标待识别短句中每个字符的特征向量以得到特征向量集合，将特征向量集合输入至双向神经网络模型中进行识别，以得到目标识别短句中包含的实体

图1

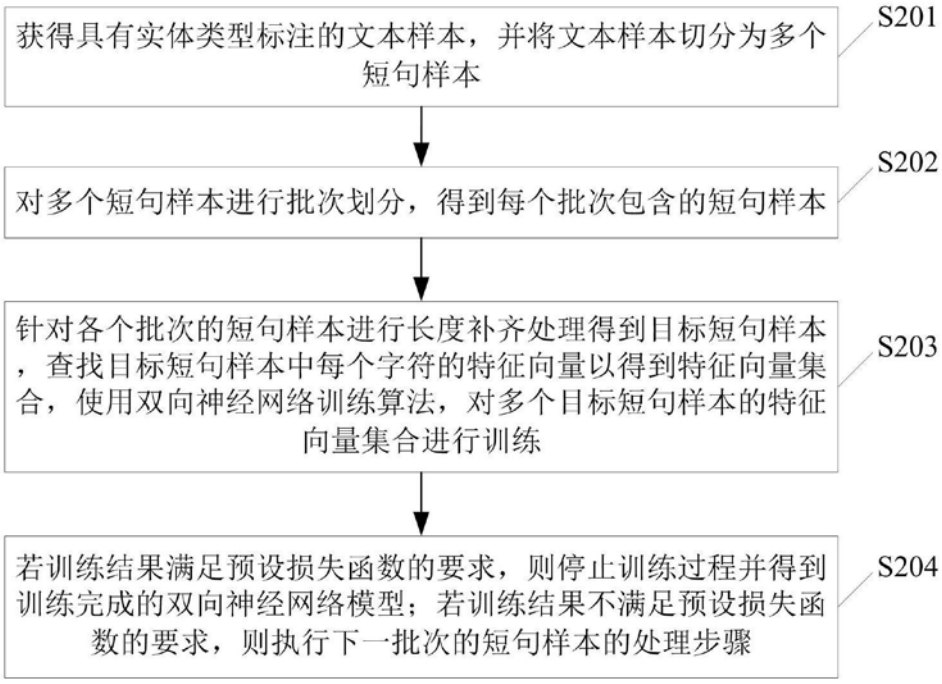

S201
获得具有实体类型标注的文本样本，并将文本样本切分为多个短句样本
S202
对多个短句样本进行批次划分，得到每个批次包含的短句样本
S203
针对各个批次的短句样本进行长度补齐处理得到目标短句样本，查找目标短句样本中每个字符的特征向量以得到特征向量集合，使用双向神经网络训练算法，对多个目标短句样本的特征向量集合进行训练
S204
若训练结果满足预设损失函数的要求，则停止训练过程并得到训练完成的双向神经网络模型；若训练结果不满足预设损失函数的要求，则执行下一批次的短句样本的处理步骤

图2

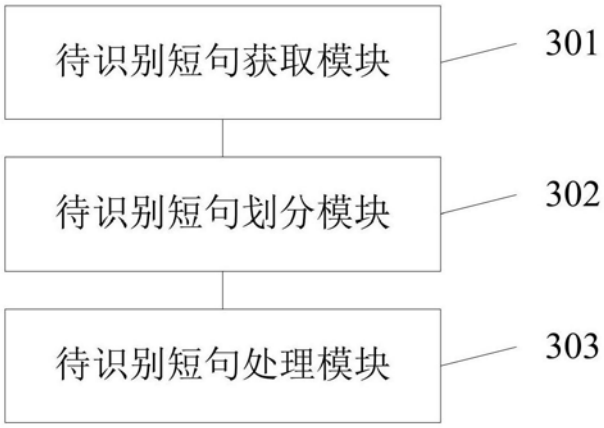

待识别短句获取模块
301
待识别短句划分模块
302
待识别短句处理模块
303

图3

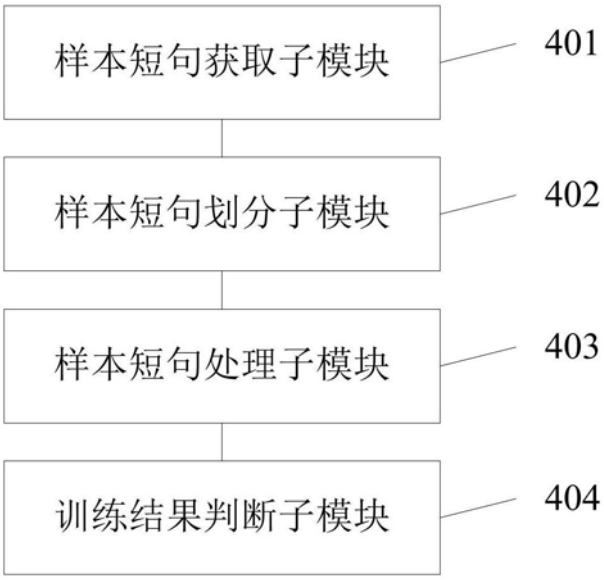

样本短句获取子模块
401
样本短句划分子模块
402
样本短句处理子模块
403
训练结果判断子模块
404

图4

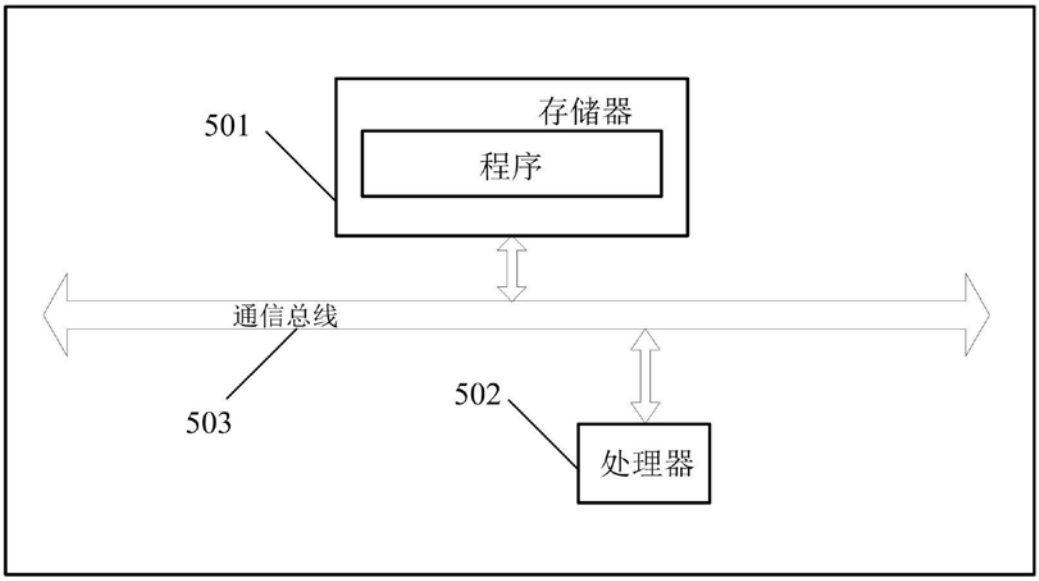


图5