(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113128225 B
(45)授权公告日 2024.06.21

(21)申请号 201911416526.7

(22)申请日 2019.12.31

(65)同一申请的已公布的文献号
申请公布号 CN 113128225 A

(43)申请公布日 2021.07.16

(73)专利权人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼资本大厦一座四层847号邮箱

(72)发明人 丁瑞雪 马春平 王潇斌 徐光伟 谢朋峻 黄非 司罗 龙定坤

(74)专利代理机构 北京博浩百睿知识产权代理有限责任公司 11134
专利代理师 谢湘宁

(51)Int.Cl.
G06F 40/295(2020.01)

(56)对比文件

张仲伟;曹雷;陈希亮;寇大磊;宋天挺.基于神经网络的知识推理研究综述.计算机工程与应用.2019,(第12期),全文.

顾凌云.基于多注意力的中文命名实体识别.信息与电脑(理论版).2019,(第09期),全文.

审查员 崔小利

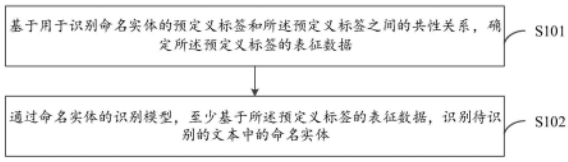
权利要求书4页 说明书23页 附图10页

(54)发明名称

命名实体的识别方法、装置、电子设备及计算机存储介质

(57)摘要

本申请实施例提供了一种命名实体的识别方法、命名实体的识别模型的训练方法、图神经网络模型的训练方法、装置、电子设备及计算机存储介质,涉及自然语言处理技术领域。其中,所述方法包括:基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系,确定所述预定义标签的表征数据,其中,所述预定义标签的表征数据包括所述预定义标签之间的共性特征数据;通过命名实体的识别模型,至少基于所述预定义标签的表征数据,识别待识别的文本中的命名实体。通过本申请实施例,能够有效提升属于预定义标签中的长尾类型的实体的识别效果。



CN 113128225 B

1.一种命名实体的识别方法,其特征在于,所述方法包括:

基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系,确定所述预定义标签的表征数据,其中,所述预定义标签的表征数据包括所述预定义标签之间的共性特征数据,所述预定义标签包括前缀和预定义实体类型;

通过命名实体的识别模型中的编码层,对待识别的文本中的文字进行编码,以获得所述待识别的文本中的文字的第一特征数据,其中,命名实体的识别模型为基于双向长短期记忆网络的条件随机场模型,其中编码层为双向长短期记忆网络,解码层为条件随机场;

基于所述预定义标签的表征数据与所述文字的第一特征数据,确定所述文字取得对应所述预定义标签的概率;

通过所述命名实体的识别模型中的解码层,基于所述文字取得对应所述预定义标签的概率,对所述待识别的文本中的文字进行解码,以获得所述待识别的文本中的命名实体。

2.根据权利要求1所述的方法,其特征在于,所述基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系,确定所述预定义标签的表征数据,包括:

基于所述预定义标签包括的预定义实体类型和所述预定义实体类型之间的共性关系,确定所述预定义实体类型的表征数据,其中,所述预定义实体类型的表征数据包括所述预定义实体类型之间的共性特征数据。

3.根据权利要求2所述的方法,其特征在于,所述基于所述预定义标签包括的预定义实体类型和所述预定义实体类型之间的共性关系,确定所述预定义实体类型的表征数据之前,所述方法还包括:

对所述预定义实体类型执行共性关系的提取操作,以获得所述预定义实体类型的共性结构;

确定所述共性结构表示的预定义实体类型之间的共性关系为所述预定义实体类型之间的共性关系。

4.根据权利要求2所述的方法,其特征在于,所述基于所述预定义标签包括的预定义实体类型和所述预定义实体类型之间的共性关系,确定所述预定义实体类型的表征数据,包括:

将所述预定义实体类型和所述预定义实体类型之间的共性关系分别作为图结构数据的节点和边;

通过图神经网络模型,对所述图结构数据执行编码操作,以获得所述图结构数据的节点的表征数据;

将所述图结构数据的节点的表征数据作为所述图结构数据的节点表示的预定义实体类型的表征数据。

5.根据权利要求1所述的方法,其特征在于,所述基于所述预定义标签的表征数据与所述文字的第一特征数据,确定所述文字取得对应所述预定义标签的概率,包括:

通过所述命名实体的识别模型中的全连接层,对所述文字的第一特征数据进行映射,以获得与所述预定义标签的表征数据的数据大小相同的所述文字的第二特征数据;

对所述预定义标签的表征数据与所述文字的第二特征数据进行点乘,以获得所述文字取得对应所述预定义标签的概率。

6.一种命名实体的识别装置,其特征在于,所述装置包括:

确定模块,用于基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系,确定所述预定义标签的表征数据,其中,所述预定义标签的表征数据包括所述预定义标签之间的共性特征数据,所述预定义标签包括前缀和预定义实体类型;

识别模块,用于:

通过命名实体的识别模型中的编码层,对待识别的文本中的文字进行编码,以获得所述待识别的文本中的文字的第一特征数据,其中,命名实体的识别模型为基于双向长短期记忆网络的条件随机场模型,其中编码层为双向长短期记忆网络,解码层为条件随机场;

基于所述预定义标签的表征数据与所述文字的第一特征数据,确定所述文字取得对应所述预定义标签的概率;

通过所述命名实体的识别模型中的解码层,基于所述文字取得对应所述预定义标签的概率,对所述待识别的文本中的文字进行解码,以获得所述待识别的文本中的命名实体。

7.一种命名实体的识别方法,其特征在于,所述方法包括:

确定预定义标签的数量是否超过预设数量,或者所述预定义标签中的实体类型是否存在长尾类型,其中,所述预定义标签用于识别电商平台的网页文本中的命名实体;

如果确定所述预定义标签的数量超过所述预设数量,或者所述预定义标签中的实体类型存在所述长尾类型,则基于所述预定义标签和所述预定义标签之间的共性关系,确定所述预定义标签的表征数据,其中,所述预定义标签的表征数据包括所述预定义标签之间的共性特征数据,所述预定义标签包括前缀和预定义实体类型;

通过命名实体的识别模型中的编码层,对待识别的文本中的文字进行编码,以获得所述待识别的文本中的文字的第一特征数据,其中,命名实体的识别模型为基于双向长短期记忆网络的条件随机场模型,其中编码层为双向长短期记忆网络,解码层为条件随机场,所述文本为网页文本;

基于所述预定义标签的表征数据与所述文字的第一特征数据,确定所述文字取得对应所述预定义标签的概率;

通过所述命名实体的识别模型中的解码层,基于所述文字取得对应所述预定义标签的概率,对所述待识别的文本中的文字进行解码,以获得所述待识别的文本中的命名实体。

8.一种命名实体的识别模型的训练方法,其特征在于,所述方法包括:

通过命名实体的识别模型中的编码层,对待识别的文本中的文字进行编码,以获得所述待识别的文本中的文字的第一特征数据,其中,命名实体的识别模型为基于双向长短期记忆网络的条件随机场模型,其中编码层为双向长短期记忆网络,解码层为条件随机场,所述文本为文本样本;

基于预定义标签的表征数据与所述文字的第一特征数据,确定所述文字取得对应所述预定义标签的概率;

通过所述命名实体的识别模型中的解码层,基于所述文字取得对应所述预定义标签的概率,对所述待识别的文本中的文字进行解码,以获得所述待识别的文本样本中的命名实体,以获得所述文本样本中的命名实体识别数据,其中,所述预定义标签的表征数据包括所述预定义标签之间的共性特征数据,所述预定义标签包括前缀和预定义实体类型;

基于所述文本样本中的命名实体识别数据和命名实体标注数据,对待训练的所述命名实体的识别模型进行训练。

9.根据权利要求8所述的方法,其特征在于,所述基于所述文本样本中的命名实体识别数据和命名实体标注数据,对待训练的所述命名实体的识别模型进行训练,包括:

通过目标损失函数,确定所述命名实体识别数据和所述命名实体标注数据之间的差异值;

基于所述差异值,调整待训练的所述命名实体的识别模型的模型参数。

10.一种命名实体的识别方法,其特征在于,所述方法包括:

基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系,确定所述预定义标签的表征数据,其中,所述预定义标签的表征数据包括所述预定义标签之间的共性特征数据,所述预定义标签包括前缀和预定义实体类型;

通过命名实体的识别模型中的编码层,对待识别的文本中的文字进行编码,以获得所述待识别的文本中的文字的第一特征数据,其中,命名实体的识别模型为基于双向长短期记忆网络的条件随机场模型,其中编码层为双向长短期记忆网络,解码层为条件随机场;

基于所述预定义标签的表征数据与所述文字的第一特征数据,确定所述文字取得对应所述预定义标签的概率;

通过所述命名实体的识别模型中的解码层,基于所述文字取得对应所述预定义标签的概率,对所述待识别的文本中的文字进行解码,以获得所述待识别的文本中属于长尾类型的命名实体。

11.一种命名实体的识别方法,其特征在于,所述方法包括:

通过图神经网络模型,对图结构数据执行编码操作,以获得所述图结构数据的节点的结构特征表征数据,其中,所述图结构数据的节点和边分别表示预定义实体类型和所述预定义实体类型之间的共性关系;

将所述图结构数据的节点的结构特征表征数据作为所述图结构数据的节点表示的预定义实体类型的表征数据,其中,所述预定义实体类型的表征数据包括所述预定义实体类型之间的共性特征数据;

通过命名实体的识别模型中的编码层,对待识别的文本中的文字进行编码,以获得所述待识别的文本中的文字的第一特征数据,其中,命名实体的识别模型为基于双向长短期记忆网络的条件随机场模型,其中编码层为双向长短期记忆网络,解码层为条件随机场;

基于预定义标签的表征数据与所述文字的第一特征数据,确定所述文字取得对应所述预定义标签的概率;

通过所述命名实体的识别模型中的解码层,基于所述文字取得对应所述预定义标签的概率,对所述待识别的文本中的文字进行解码,以获得所述待识别的文本中属于长尾类型的命名实体。

12.一种命名实体的识别方法,其特征在于,所述方法包括:

基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系,确定所述预定义标签的表征数据,其中,所述预定义标签的表征数据包括所述预定义标签之间的共性特征数据,所述预定义标签包括前缀和预定义实体类型;

通过命名实体的识别模型中的编码层,对待识别的文本中的文字进行编码,以获得所述待识别的文本中的文字的第一特征数据,其中,命名实体的识别模型为基于双向长短期记忆网络的条件随机场模型,其中编码层为双向长短期记忆网络,解码层为条件随机场,所

述文本为案件图谱；

基于所述预定义标签的表征数据与所述文字的第一特征数据，确定所述文字取得对应所述预定义标签的概率；

通过所述命名实体的识别模型中的解码层，基于所述文字取得对应所述预定义标签的概率，对所述待识别的文本中的文字进行解码，以获得所述待识别的文本中属于长尾类型的命名实体。

13.一种命名实体的识别方法，其特征在于，所述方法包括：

基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系，确定所述预定义标签的表征数据，其中，所述预定义标签的表征数据包括所述预定义标签之间的共性特征数据，所述预定义标签包括前缀和预定义实体类型；

通过命名实体的识别模型中的编码层，对待识别的文本中的文字进行编码，以获得所述待识别的文本中的文字的第一特征数据，其中，命名实体的识别模型为基于双向长短期记忆网络的条件随机场模型，其中编码层为双向长短期记忆网络，解码层为条件随机场，所述文本为起诉书；

基于所述预定义标签的表征数据与所述文字的第一特征数据，确定所述文字取得对应所述预定义标签的概率；

通过所述命名实体的识别模型中的解码层，基于所述文字取得对应所述预定义标签的概率，对所述待识别的文本中的文字进行解码，以获得所述待识别的文本中属于长尾类型的命名实体。

14.一种电子设备，其特征在于，所述设备包括：

一个或多个处理器；

计算机可读介质，配置为存储一个或多个程序，

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求1-5中任意一项权利要求所述的命名实体的识别方法，实现如权利要求7所述的命名实体的识别方法，实现如权利要求8或9所述的命名实体的识别模型的训练方法，实现如权利要求10所述的命名实体的识别方法，实现如权利要求11所述的命名实体的识别方法，实现如权利要求12所述的命名实体的识别方法，或者实现如权利要求13所述的命名实体的识别方法。

15.一种计算机可读介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现如权利要求1-5中任意一项权利要求所述的命名实体的识别方法，实现如权利要求7所述的命名实体的识别方法，实现如权利要求8或9所述的命名实体的识别模型的训练方法，实现如权利要求10所述的命名实体的识别方法，实现如权利要求11所述的命名实体的识别方法，实现如权利要求12所述的命名实体的识别方法，或者实现如权利要求13所述的命名实体的识别方法。

命名实体的识别方法、装置、电子设备及计算机存储介质

技术领域

[0001] 本申请实施例涉及自然语言处理技术领域,尤其涉及一种命名实体的识 别方法、命名实体的识别模型的训练方法、图神经网络模型的训练方法、装 置、电子设备及计算机存储介质。

背景技术

[0002] 命名实体识别(Named Entity Recognition,简称NER),又作为“专名识 别”,是指识别文本中具有特定意义的实体,主要包括人名、地名、机构名、 特定意义的网络词汇、其他专有名词等。命名实体识别在信息提取、问答系 统、句法分析、机器翻译等应用领域中发挥重要作用。因此,对文本进行命 名实体识别是很多信息处理顶层应用的基础。

[0003] 目前,在命名实体识别的应用中,经常会出现一些缺乏训练数据的实体 类型(长尾类型)。属于长尾类型的实体在语料中出现的次数较少,但十分 重要。由于长尾类型实体的训练数据的数量较少,导致长尾类型实体的最终 的识别结果接近随机初始化。由此可见,如何有效提升长尾类型的实体的识 别效果成为当前亟待解决的技术问题。

发明内容

[0004] 有鉴于此,本申请实施例提供一种命名实体的识别方法、命名实体的识 别模型的训练方法、图神经网络模型的训练方法、装置、电子设备及计算机 存储介质,以解决现有技术中存在的如何有效提升长尾类型的实体的识别效 果的技术问题。

[0005] 根据本发明实施例的第一方面,提供了一种命名实体的识别方法。所述 方法包括:基于用于识别命名实体的预定义标签和所述预定义标签之间的共 性关系,确定所述预定义标签的表征数据,其中,所述预定义标签的表征数 据包括所述预定义标签之间的共性特征数据;通过命名实体的识别模型,至 少基于所述预定义标签的表征数据,识别待识别的文本中的命名实体。

[0006] 根据本发明实施例的第二方面,提供了一种命名实体的识别装置。所述 装置包括:确定模块,用于基于用于识别命名实体的预定义标签和所述预定 义标签之间的共性关系,确定所述预定义标签的表征数据,其中,所述预定 义标签的表征数据包括所述预定义标签之间的共性特征数据;识别模块,用 于通过命名实体的识别模型,至少基于所述预定义标签的表征数据,识别待 识别的文本中的命名实体。

[0007] 根据本发明实施例的第三方面,提供了一种命名实体的识别方法。所述 方法包括:确定预定义标签的数量是否超过预设数量,或者所述预定义标签 中的实体类型是否存在长尾类型,其中,所述预定义标签用于识别电商平台 的网页文本中的命名实体;如果确定所述预定义标签的数量超过所述预设数 量,或者所述预定义标签中的实体类型存在所述长尾类型,则基于所述预定 义标签和所述预定义标签之间的共性关系,确定所述预定义标签的表征数据, 其中,所述预定义标签的表征数据包括所述预定义标签之间的共性特征数据, 并通过命名实体的识别模型,至少基于所述预定义标签的表征数据,识别待 识别的

所述网页文本中的命名实体。

[0008] 根据本发明实施例的第四方面,提供了一种命名实体的识别模型的训练 方法。所述方法包括:通过待训练的命名实体的识别模型,至少基于用于识 别命名实体的预定义标签的表征数据,对文本样本中的命名实体进行识别, 以获得所述文本样本中的命名实体识别数据,其中,所述预定义标签的表征 数据包括所述预定义标签之间的共性特征数据;基于所述文本样本中的命名 实体识别数据和命名实体标注数据,对待训练的所述命名实体的识别模型进 行训练。

[0009] 根据本发明实施例的第五方面,提供了一种图神经网络模型的训练方法。所述方法包括:通过待训练的图神经网络模型,对图结构数据样本执行编码 操作,以获得所述图结构数据样本的节点的结构特征表征数据,其中,所述 图结构数据样本的节点和边分别表示预定义实体类型和所述预定义实体类型 之间的共性关系;基于所述图结构数据样本的节点的结构特征表征数据和结 构特征标注数据,对待训练的所述图神经网络模型进行训练。

[0010] 根据本发明实施例的第六方面,提供了一种命名实体的识别方法。所述 方法包括:基于用于识别命名实体的预定义标签和所述预定义标签之间的共 性关系,确定所述预定义标签的表征数据,其中,所述预定义标签的表征数 据包括所述预定义标签之间的共性特征数据;通过命名实体的识别模型,至 少基于所述预定义标签的表征数据,识别待识别的文本中属于长尾类型的命 名实体。

[0011] 根据本发明实施例的第七方面,提供了一种命名实体的识别方法。所述 方法包括:通过图神经网络模型,对图结构数据执行编码操作,以获得所述 图结构数据的节点的结构特征表征数据,其中,所述图结构数据的节点和边 分别表示预定义实体类型和所述预定义实体类型之间的共性关系;将所述图 结构数据的节点的结构特征表征数据作为所述图结构数据的节点表示的预定 义实体类型的表征数据,其中,所述预定义实体类型的表征数据包括所述预 定义实体类型之间的共性特征数据;通过命名实体的识别模型,至少基于所 述预定义实体类型的表征数据,识别待识别的文本中属于长尾类型的命名实 体。

[0012] 根据本发明实施例的第八方面,提供了一种命名实体的识别方法。所述 方法包括:基于用于识别命名实体的预定义标签和所述预定义标签之间的共 性关系,确定所述预定义标签的表征数据,其中,所述预定义标签的表征数 据包括所述预定义标签之间的共性特征数据;通过命名实体的识别模型,至 少基于所述预定义标签的表征数据,识别待识别的案件图谱中属于长尾类型 的命名实体。

[0013] 根据本发明实施例的第九方面,提供了一种命名实体的识别方法。所述 方法包括:基于用于识别命名实体的预定义标签和所述预定义标签之间的共 性关系,确定所述预定义标签的表征数据,其中,所述预定义标签的表征数 据包括所述预定义标签之间的共性特征数据;通过命名实体的识别模型,至 少基于所述预定义标签的表征数据,识别待识别的起诉书中属于长尾类型的 命名实体。

[0014] 根据本申请实施例的第十方面,提供了一种电子设备,包括:一个或多 个处理器;计算机可读介质,配置为存储一个或多个程序,当所述一个或多 个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现如上 述实施例的第一方面、第三方面、第六方面、第七方面、第八方面,或者第 九方面所述的命名实体识别方法,实现如上述实施例的

第四方面所述的命名 实体的识别模型的训练方法,或者实现如上述实施例的第五方面所述的图神 经网络模型的训练方法。

[0015] 根据本申请实施例的第十一方面,提供了一种计算机可读介质,其上存 储有计算机程序,该程序被处理器执行时实现如上述实施例的第一方面、第 三方面、第六方面、第七方面、第八方面,或者第九方面所述的命名实体识 别方法,实现如上述实施例的第四方面所述的命名实体的识别模型的训练方 法,或者实现如上述实施例的第五方面所述的图神经网络模型的训练方法。

[0016] 根据本申请实施例提供的命名实体的识别方案,基于用于识别命名实体 的预定义标签和预定义标签之间的共性关系,确定预定义标签的表征数据, 其中,预定义标签的表征数据包括预定义标签之间的共性特征数据,并通过 命名实体的识别模型,至少基于预定义标签的表征数据,识别待识别的文本 中的命名实体,与现有的其它方式相比,充分利用不同的预定义标签之间的 共性关系,以获得包括不同的预定义标签之间的共性特征数据的预定义标签 表征数据,并通过预定义标签表征数据中的共性特征数据,使用属于预定义 标签中的非长尾类型的实体的训练效果,来有效提升属于预定义标签中的长 尾类型的实体的训练效果,从而能够有效提升属于预定义标签中的长尾类型 的实体的识别效果。

附图说明

[0017] 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实施例中记载的一些实施例,对于本领域普通技 术人员来讲,还可以根据这些附图获得其他的附图。

[0018] 图1A为本申请实施例提供的一种命名实体的识别系统的结构示意图;

[0019] 图1B为本申请实施例一中命名实体的识别方法的步骤流程图;

[0020] 图1C为根据本申请实施例一提供的层次聚类过程的示意图;

[0021] 图1D为根据本申请实施例一提供的图结构数据的编码过程的示意图;

[0022] 图2A为本申请实施例二中命名实体的识别方法的步骤流程图;

[0023] 图2B为根据本申请实施例二提供的命名实体的识别过程的示意图;

[0024] 图2C为根据本申请实施例二提供的命名实体的识别过程的示意图;

[0025] 图3为本申请实施例三中命名实体的识别模型的训练方法的步骤流程 图;

[0026] 图4为本申请实施例四中图神经网络模型的训练方法的步骤流程图;

[0027] 图5为本申请实施例五中命名实体的识别方法的步骤流程图;

[0028] 图6为本申请实施例六中命名实体的识别方法的步骤流程图;

[0029] 图7为本申请实施例七中命名实体的识别方法的步骤流程图;

[0030] 图8为本申请实施例八中命名实体的识别方法的步骤流程图;

[0031] 图9为本申请实施例九中命名实体的识别装置的结构示意图;

[0032] 图10为本申请实施例十中命名实体的识别装置的结构示意图;

[0033] 图11为本申请实施例十一中电子设备的结构示意图;

[0034] 图12为本申请实施例十二中电子设备的硬件结构。

具体实施方式

[0035] 为了使本领域的人员更好地理解本发明实施例中的技术方案,下面将结 合本发明实施例中的附图,对本发明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发明实施例一部分实施例,而不是全部 的实施例。基于本发明实施例中的实施例,本领域普通技术人员所获得的所 有其他实施例,都应当属于本发明实施例保护的范围。

[0036] 下面结合本发明实施例附图进一步说明本发明实施例的具体实现。

[0037] 目前,在命名实体识别的应用中,提出利用深度学习来识别语料中的命 名实体,即基于监督学习模型,从人工标注的训练数据中学习命名实体的识 别模型,然后将此模型用于对实际场景的文本(称为测试数据)进行命名实 体识别。在基于神经网络模型的命名实体识别中,当需要识别的预定义标签 中存在较多的实体类型时,经常会出现一些缺乏训练数据的实体类型(长尾 类型)。属于长尾类型的实体在语料中出现的次数较少,但十分重要。由于 长尾类型实体的训练数据的数量较少,导致长尾类型实体的最终的识别结果 接近随机初始化。基于此,本申请实施例提供了一种命名实体的识别方法, 充分利用不同的预定义标签之间的共性关系,以获得包括不同的预定义标签 之间的共性特征数据的预定义标签表征数据,并通过预定义标签表征数据中 的共性特征数据,使用属于预定义标签中的非长尾类型的实体的训练效果, 来有效提升属于预定义标签中的长尾类型的实体的训练效果,从而能够有效 提升属于预定义标签中的长尾类型的实体的识别效果。

[0038] 参照图1A,为实现本申请实施例提供的命名实体的识别方法的一种系统 结构示意图,该系统可以包括服务器以及终端设备A,应该理解,图1A所 呈现的服务器与终端设备A仅是示例性说明,并不会对两者的实现形式做限 定。

[0039] 在实际应用中,服务器与终端设备A之间可以是有线或无线网络连接, 具体可以通过GSM(Global System for Mobile Communications,全球移动通 信系统)、GPRS(General Packet Radio Service,通用分组无线业务)、LTE (Long Term Evolution,长期演进)等移动网络实现通信连接,或者是通过 蓝牙、WIFI、红外线等方式进行通信连接,本申请实施例对服务器与终端设 备A之间的具体通信连接方式不做限定。

[0040] 服务器可以是为用户提供服务的服务设备,具体可以是独立的应用服务 设备,也可以是由多个服务器构成的服务集群,实际应用中,其可以是云服 务器、云主机、虚拟中心等,本申请实施例对该服务器的结构及其实现形式 不作限定。

[0041] 终端设备A可以是面向用户,并能够与用户进行交互的终端,如手机、 笔记本、电脑、iPad、智能音响等,还可以各种自助终端,如医院、银行、 车站等场所中的自助服务机,此外,终端设备A还可以是支持交互的智能机 器,如聊天机器人、扫地机器人、点餐服务机器人等。本申请实施例对终端 设备的产品类型及其物理形态不做限定,本申请实施例需要其具有交互功能, 可以通过安装如新闻浏览等交互类应用程序实现。

[0042] 在进行命名实体识别时,终端设备A可通过网络向服务器发送针对待识 别文本的命名实体的识别请求。服务器接收终端设备A发送的针对待识别文 本的命名实体的识别请求,并基于命名实体的识别请求向所述终端设备A返 回针对所述命名实体的识别请求的响应结果。例如,当待识别的文本为“中 国”时,命名实体的识别请求携带待识别文本“中国”,并且针对命名实体 的识别请求的响应结果为“中(B-LOC)国(E-LOC)”,其中,预定义标 签

“B-LOC”可理解为“中”是位置实体类型的实体的开始,预定义标签 “E-LOC”可理解为“国”是位置实体类型的实体的结束。由此可见,本申 请实施例提供的命名实体的识别方法可以由服务器执行,具体的实现过程可 以参照下文方法实施例的描述。

[0043] 结合上图1A所示的系统结构示意图,参照图1B,为本申请实施例一的 命名实体的识别方法的流程示意图,可以应用于各种应用场景下的文本的识 别过程,具体可以由服务器执行,如图1B所示,该方法可以包括但并不局 限于以下步骤:

[0044] 在步骤S101中,基于用于识别命名实体的预定义标签和预定义标签之 间的共性关系,确定预定义标签的表征数据。

[0045] 在本申请实施例中,用于识别命名实体的每个预定义标签由前缀和预定 义实体类型组成。前缀B表示这个字是一个实体的开始,前缀I表示这个字 在一个实体内部,前缀E表示这个字是一个实体的结尾,前缀S表示这个实 体是一个单字词。预定义实体类型由不同的数据集定义,可以是时间(TIME)、地点(LOC)、人名(PER)、机构(ORG),也可以是其他自定义的实体 类型,实体类型的集合都必须包含其他类型(O)用以表示没有被划分到预 定义实体类型集合里面的字。由此,预定义标签可理解为用于指示文本中的 文字所属的实体的实体类型,以及文本中的文字在所属的实体中的位置的标 签,例如,“B-LOC”、“E-LOC”等。预定义标签之间的共性关系可理解为 预定义标签之间具有共同属性或者共性特征的关系。例如,预定义标签“B- 政治家”与预定义标签“B-军人”之间的共同属性为“人”,预定义标签“B- 歌手”与预定义标签“B-演员”之间的共同属性为“明星”等。预定义标签 的表征数据可理解为用于表征预定义标签的语义特征的数据,例如,预定义 标签的表征向量,并且预定义标签的表征数据包括预定义标签之间的共性特 征数据。共性特征数据可理解为用于表征预定义标签之间的共性特征的数据。 例如,表征预定义标签“B-政治家”与预定义标签“B-军人”之间的共同特 征“人”的数据,表征预定义标签“B-歌手”与预定义标签“B-演员”之间 的共同特征“明星”的数据。可以理解的是,以上描述仅为示例性的,本申 请实施例对此不做任何限定。

[0046] 在一些可选实施例中,预定义标签包括前缀和预定义实体类型。在基于 用于识别命名实体的预定义标签和预定义标签之间的共性关系,确定预定义 标签的表征数据时,基于预定义标签包括的预定义实体类型和预定义实体类 型之间的共性关系,确定预定义实体类型的表征数据,其中,预定义实体类 型的表征数据包括预定义实体类型之间的共性特征数据。在确定得到预定义 实体类型的表征数据之后,可通过全连接层,对预定义实体类型的表征数据 执行映射操作,以获得预定义标签的表征数据。籍此,通过基于预定义实体类型和预定义实体类型之间的共性关系确定得到的预定义实体类型的表征数 据,能够准确地确定预定义标签的表征数据。可以理解的是,以上描述仅为 示例性的,本申请实施例对此不做任何限定。

[0047] 在一个具体的例子中,预定义实体类型之间的共性关系可理解为预定义 实体类型之间具有共同属性或者共性特征的关系。例如,预定义实体类型“政 治家”与预定义实体类型“军人”之间的共同属性为“人”,预定义实体类 型“歌手”与预定义实体类型“演员”之间的共同属性为“明星”等。预定 义实体类型的表征数据可理解为用于表征预定义实体类型的语义特征的数 据,例如,预定义实体类型的表征向量,并且预定义实体类型的表征数据包 括预定义实体类型之间的共性特征数据。预定义实体类型之间的共性特征数 据可理

解为用于表征预定义实体类型之间的共性特征的数据。例如,表征预 定义实体类型“政治家”与预定义实体类型“军人”之间的共同特征“人” 的数据,表征预定义实体类型“歌手”与预定义实体类型“演员”之间的共 同特征“明星”的数据。此外,全连接层的每一个结点都与上一层的所有结 点相连,用来把前边提取到的特征综合起来。由于其全相连的特性,一般全 连接层的参数也是最多的。因此,全连接层可以理解为整合上一层中具有类 别区分性的局部信息的计算层。在通过全连接层,对预定义实体类型的表征 数据执行映射操作时,通过全连接层,以降维的方式对预定义实体类型的表 征向量执行映射操作,以获得预定义标签的表征向量。可以理解的是,以上 描述仅为示例性的,本申请实施例对此不做任何限定。

[0048] 在一些可选实施例中,在确定预定义实体类型的表征数据之前,该方法 还包括:对预定义实体类型执行共性关系的提取操作,以获得预定义实体类 型的共性结构;确定共性结构表示的预定义实体类型之间的共性关系为预定 义实体类型之间的共性关系。籍此,通过对预定义实体类型执行共性关系的 提取操作获得的预定义实体类型的共性结构,能够准确地确定预定义实体类 型之间的共性关系。可以理解的是,以上描述仅为示例性的,本申请实施例 对此不做任何限定。

[0049] 在一个具体的例子中,在对预定义实体类型执行共性关系的提取操作时, 对预定义实体类型执行层次聚类操作,以获得预定义实体类型的树结构。在 确定共性结构表示的预定义实体类型之间的共性关系为预定义实体类型之间 的共性关系时,确定树结构表示的预定义实体类型之间的层次结构关系为预 定义实体类型之间的共性关系。其中,共性结构具体为预定义实体类型的树 结构,共性关系具体为树结构表示的预定义实体类型之间的层次结构关系。 可以理解的是,以上描述仅为示例性的,例如,还可以通过人工,建立预定 义实体类型的树结构,本申请实施例对此不做任何限定。其中,层次聚类操 作是基于簇间的相似度在不同层次上分析数据,从而形成树形的聚类结构, 层次聚类操作一般有两种划分策略:自底向上的聚合策略和自顶向下的分拆 策略。在对预定义实体类型执行层次聚类操作时,可采用自底向上的聚合策 略。该策略假设每个样本点都是单独的簇类,然后在算法运行的每一次迭代 中找出相似度较高的簇类进行合并,该过程不断重复,直到达到预设的簇类 个数k或只有一个簇类。该策略的基本思想是:1)计算数据集的相似矩阵; 2)假设每个样本点为一个簇类;3)循环:合并相似度最高的两个簇类,然 后更新相似矩阵;4)当簇类个数为1时,循环终止。为了更好地理解,对预 定义实体类型执行的层次聚类操作进行如图1C的图示说明。假设有6种预 定义实体类型{A,B,C,D,E,F},且每种预定义实体类型都为一个簇类, 计算每个簇类之间的相似度,得到相似矩阵。如果B和C的相似度最高,合 并B和C为一个簇类。现在还有五个簇类,分别为A,BC,D,E,F。更新 簇类之间的相似矩阵,相似矩阵的大小为五行五列。如果簇类BC和D的相 似度最高,合并簇类BC和D为一个簇类。现在还有四个簇类,分别为A, BCD,E,F。更新簇类之间的相似矩阵,相似矩阵的大小为四行四列。如果 簇类E和F的相似度最高,合并簇类E和F为一个簇类。现在还有三个簇类, 分别为A,BCD,EF。更新簇类之间的相似矩阵,相似矩阵的大小为三行三 列。如果簇类BCD和EF的相似度最高,合并簇类BCD和EF为一个簇类。 现在还有两个簇类,分别为A,BCDEF。最后合并簇类A和BCDEF为一个 簇类,层次聚类操作结束。根据上面描述的步骤,可使用树结构对层次聚类 操作进行可视化,记录了簇类的聚合顺序。其中,可使用距离来评价簇间的 相似度,

即距离越小相似度越高,距离越大相似度越低。常用的簇间相似度 的计算方法包括:最小距离法、最大距离法、平均距离法、中心距离法、最 小方差法等。如图1C所示,树结构表示的预定义实体类型B和C之间的层 次结构关系为预定义实体类型B和C之间的共性关系,树结构表示的预定义 实体类型A和B之间的层次结构关系为预定义实体类型A和B之间的共性关系等。可以理解的是,以上描述仅为示例性的,本申请实施例对此不做任 何限定。

[0050] 在一些可选实施例中,在基于预定义标签包括的预定义实体类型和预定 义实体类型之间的共性关系,确定预定义实体类型的表征数据时,将预定义 实体类型和预定义实体类型之间的共性关系分别作为图结构数据的节点和 边;通过图神经网络模型,对图结构数据执行编码操作,以获得图结构数据 的节点的表征数据;将图结构数据的节点的表征数据作为图结构数据的节点 表示的预定义实体类型的表征数据。籍此,通过图神经网络模型,对预定义 实体类型和预定义实体类型之间的共性关系建模得到的图结构数据执行编码 操作,能够准确地确定预定义实体类型的表征数据。可以理解的是,以上描 述仅为示例性的,本申请实施例对此不做任何限定。

[0051] 在一个具体的例子中,在将预定义实体类型和预定义实体类型之间的共 性关系分别作为图结构数据的节点和边时,将树结构中的预定义实体类型和 树结构表示的预定义实体类型之间的层次结构关系分别作为图结构数据的节 点和边。通过将树结构中的预定义实体类型和树结构表示的预定义实体类型 之间的层次结构关系分别作为图结构数据的节点和边,来建模得到图结构数 据,并使用预定义实体类型的语义表征向量作为图结构数据的初始状态。然 后,通过图神经网络模型,对图结构数据执行编码操作,学习不同的预定义 实体类型之间的层次结构关系,以获得图结构数据的节点的表征向量,并将 图结构数据的节点的表征向量作为图结构数据的节点表示的预定义实体类型 的表征向量。在通过图神经网络模型,对图结构数据执行编码操作时,在图 结构数据中的节点上执行随机游走生成节点序列;运行skip-gram模型(连 续跳跃元语法模型),根据生成的图结构数据的节点序列学习图结构数据中 的每个节点的表征向量。其中,图结构数据可理解为一种非线性的数据结构, 图数据结构在实际生活中有很多例子,比如,交通运输网,地铁网络,社交网络,计算机中的状态执行(自动机)等等都可以抽象成图数据结构。在本 申请实施例中,图结构数据可为节点表示预定义实体类型,并且边表示预定 义实体类型之间的共性关系的图结构数据。层次结构关系可理解为预定义标 签之间具有层次属性或者层次特征的关系。可以理解的是,以上描述仅为示 例性的,本申请实施例对此不做任何限定。

[0052] 在一个具体的例子中,图神经网络模型(graph neural network,简称 GNN),可以指在广义的图结构上,利用一种递归聚合的方式,将图中节点的 信息进行传播,最终学习每个图节点的表征向量。更为具体的,深度学习理 论中的图神经网络模型可以是在拓扑空间(topological space)内按图(graph)结 构组织以进行关系推理(relational reasoning)的函数集合。该图神经网络模型 可以为图卷积网络模型、对于图中的节点使用向量建模的方法的图神经网络 模型、LINE(Large-scale Information Network Embedding,大规模信息网络嵌 入)模型、Node2vec(一种对于图中的节点使用向量建模的方法)模型、SDNE(Structural Deep Network Embedding,结构深层网络嵌入)模型和图自编 码器模型等。具体地,如图1D所示,将图结构数据输入至图神经网络模型 中,该图神经网络模型中包括多层神经网络,第一层神经网络对图结构数据 进行计算,得到计算结果,

作为下一层神经网络的输入,在下一层神经网络 中可以将每个节点和该节点的邻居节点的计算结果进行综合计算,得到下一 层神经网络的计算结果,以此类推,在该图神经网络模型中的最后一层神经 网络输出图结构数据,根据输出的图结构数据的节点的向量表示,从而确定 得到图结构数据的节点对应的预定义实体类型的向量表示。可以理解的是, 以上描述仅为示例性的,本申请实施例对此不做任何限定。

[0053] 在步骤S102中,通过命名实体的识别模型,至少基于预定义标签的表 征数据,识别待识别的文本中的命名实体。

[0054] 在本申请实施例中,命名实体的识别模型可为基于双向长短期记忆网络 的条件随机场模型。该模型在序列标注任务上具有很高的准确度。用于识别 命名实体的预定义标签的表征数据蕴含了很强的语义信息,例如,预定义标 签“B-政治家”的表征数据为[0.1,0.2,0.5,0.6,0.7],预定义标签“B-军人”的 表征数据为[0.1,0.2,0.3,0.5,0.8],那么预定义标签“B-政治家”的表征数据与 预定义标签“B-军人”的表征数据均包括共性特征数据[0.1,0.2]。当属于预定 义标签“B-政治家”中的预定义实体类型“政治家”的实体的训练数据较少, 且属于预定义标签“B-军人”中的预定义实体类型“军人”的实体的训练数 据较多时,可通过预定义标签“B-政治家”的表征数据包括的共性特征数据 [0.1,0.2],使用属于预定义标签“B-军人”中的预定义实体类型“军人”的实 体的训练效果,来有效提升属于预定义标签“B-政治家”中的预定义实体类 型“政治家”的实体的训练效果,从而能够有效提升属于预定义标签“B-政 治家”中的预定义实体类型“政治家”的实体的识别效果。因此,在识别文 本内容中的命名实体时,如果能充分利用预定义标签的表征数据,对文本内 容中的命名实体识别能够起到提升的作用。其中,双向长短期记忆网络由两 个普通的循环神经网络组成,一个正向的循环神经网络,利用过去的信息, 一个逆序的循环神经网络,利用未来的信息,这样在时刻t,既能够使用t-1 时刻的信息,又能够利用到t+1时刻的信息。一般来说,由于双向长短期记 忆网络能够同时利用过去时刻和未来时刻的信息,会比单向长短期记忆网络 最终的预测更加准确。可以理解的是,以上描述仅为示例性的,本申请实施 例对此不做任何限定。

[0055] 在一个具体的例子中,条件随机场模型是根据海量的特征工程提取足够 的不同维度的特征,然后根据这些特征做序列标注。实际应用中,条件随机 场模型是一种无向图模型,它是在给定需要标记的观察序列(词、句子数值等) 的条件下,计算整个标记序列的联合概率分布。条件随机场模型是一个端到 端的,所有特征提取的工作交给深度学习模型来做,根据双向长短时记忆网 络得到的X(如X_1、X_2…X_i…X_n),可以利用立足于局部适配解,算出可能的 序列Y(如Y_1、Y_2…Y_i…Y_n)的概率分布,也就是最终的标记,即命名实体识 别结果。具体地,对输入的文本内容进行时间序列建模,然后使用双向长短 时记忆网络计算每个字取每个标签的概率,最后利用条件随机场模型解码。 更具体地,通过图神经网络模型和全连接层,获得每个预定义标签i的表征 向量V_i。在得到每个预定义标签i的表征向量V_i之后,对表征向量V_i与双 向长短时记忆网络输出的文字的特征向量执行点乘操作,以获得每一个文字 取预定义标签i的概率。随后将每一个文字取预定义标签i的概率输入条件 随机场模型进行解码。可以理解的是,以上描述仅为示例性的,本申请实施 例对此不做任何限定。

[0056] 通过本申请实施例提供的命名实体的识别方法,基于用于识别命名实体 的预定

义标签和预定义标签之间的共性关系,确定预定义标签的表征数据, 其中,预定义标签的表征数据包括预定义标签之间的共性特征数据,并通过 命名实体的识别模型,至少基于预定义标签的表征数据,识别待识别的文本 中的命名实体,与现有的其它方式相比,充分利用不同的预定义标签之间的 共性关系,以获得包括不同的预定义标签之间的共性特征数据的预定义标签 表征数据,并通过预定义标签表征数据中的共性特征数据,使用属于预定义 标签中的非长尾类型的实体的训练效果,来有效提升属于预定义标签中的长 尾类型的实体的训练效果,从而能够来有效提升属于预定义标签中的长尾类 型的实体的识别效果。

[0057] 本实施例的命名实体的识别方法可以由任意适当的具有数据处理能力的 设备执行,包括但不限于:摄像头、终端、移动终端、PC机、服务器、车载 设备、娱乐设备、广告设备、个人数码助理(PDA)、平板电脑、笔记本电 脑、掌上游戏机、智能眼镜、智能手表、可穿戴设备、虚拟显示设备或显示 增强设备等。

[0058] 参照图2A,示出了本申请实施例二的命名实体的识别方法的步骤流程 图。

[0059] 具体地,本实施例提供的命名实体的识别方法包括以下步骤:

[0060] 在步骤S201中,基于用于识别命名实体的预定义标签和预定义标签之 间的共性关系,确定预定义标签的表征数据。

[0061] 由于步骤S201的具体实施方式与上述步骤S101的具体实施方式类似, 在此不再赘述。

[0062] 在步骤S202中,通过命名实体的识别模型中的编码层,对待识别的文 本中的文字进行编码,以获得待识别的文本中的文字的第一特征数据。

[0063] 在本申请实施例中,命名实体的识别模型可为基于双向长短期记忆网络 的条件随机场模型。编码层可为双向长短期记忆网络,是一种循环神经网络 (RNN)的变种,适合于处理和预测时间序列中间隔和延迟相对较长的重要 事件。待识别的文本可为电商平台的网页文本、案件图谱、法院的起诉书等。 第一特征数据可为待识别的文本中的文字的特征编码向量。此外,步骤S202 的执行顺序可在步骤S201的执行顺序之前,或者步骤S202与步骤S201并 行执行。可以理解的是,以上描述仅为示例性的,本申请实施例对此不做任 何限定。

[0064] 在步骤S203中,基于预定义标签的表征数据与文字的第一特征数据, 确定文字取得对应预定义标签的概率。

[0065] 在一些可选实施例中,在确定文字取得对应预定义标签的概率时,通过 命名实体的识别模型中的全连接层,对文字的第一特征数据进行映射,以获 得与预定义标签的表征数据的数据大小相同的文字的第二特征数据;对预定 义标签的表征数据与文字的第二特征数据进行点乘,以获得文字取得对应预 定义标签的概率。籍此,通过对预定义标签的表征数据与文字的第二特征数 据进行点乘,能够准确地确定文字取得对应预定义标签的概率。可以理解的 是,以上描述仅为示例性的,本申请实施例对此不做任何限定。

[0066] 在一个具体的例子中,第一特征数据可为第一特征编码向量,预定义标 签的表征数据可为预定义标签的表征向量,第二特征数据可为维度与预定义 标签的表征向量的维度相同的第二特征编码向量。其中,维度可理解为独立 的时空坐标的数目,例如,向量的时空坐标的数目。在确定文字取得对应预 定义标签的概率时,通过命名实体的识别模型中的全连接层,对文字的第一 特征编码向量执行映射操作,以获得维度与预定义标签的表征向

量的维度相 同的文字的第二特征编码向量;对预定义标签的表征向量与文字的第二特征编码向量进行点乘,以获得文字取得对应预定义标签的概率。可以理解的是, 以上描述仅为示例性的,本申请实施例对此不做任何限定。

[0067] 在步骤S204中,通过命名实体的识别模型中的解码层,基于文字取得 对应预定义标签的概率,对待识别的文本中的文字进行解码,以获得待识别 的文本中的命名实体。

[0068] 在本申请实施例中,命名实体的识别模型可为基于双向长短期记忆网络 的条件随机场模型。解码层可为条件随机场模型,是一种基于遵循马尔可夫 性的概率图模型,适用于解决序列标注、时序标注等问题。在通过命名实体 的识别模型中的解码层,基于文字取得对应预定义标签的概率,对待识别的 文本中的文字进行解码时,通过条件随机场模型,基于文字取得对应预定义 标签的概率,对待识别的文本中的文字进行序列标注,以获得待识别的文本 中的命名实体。可以理解的是,以上描述仅为示例性的,本申请实施例对此 不做任何限定。

[0069] 在一个具体的例子中,如图2B所示,以待识别的文本为“李先生访问 中国”为例,对命名实体的识别过程进行详细的说明。首先,对待识别文本 “李先生访问中国”进行时间序列建模,以获得“李先生访问中国”的时间 序列。然后,通过命名实体的识别模型的编码层,对时间序列“李先生访问 中国”中的每一个文字进行编码,以获得时间序列“李先生访问中国”中的 每一个文字的特征编码向量。在获得时间序列中的每一个文字的特征编码向量之后,可对用于识别命名实体的预定义标签中的预定义实体类型执行层次 聚类操作,以获得预定义实体类型之间的层次结构关系。在获得预定义实体 类型之间的层次结构关系之后,可通过图神经网络模型,对预定义实体类型 和预定义实体类型之间的层次结构关系执行编码操作,以获得预定义标签的 表征向量。在获得预定义标签的表征向量之后,可对时间序列中的每一个文 字的特征编码向量与预定义标签的表征向量进行点乘,以获得时间序列中的 每一个文字取得每一个预定义标签的概率,随后将时间序列中的每一个文字取得每一个预定义标签的概率输入命名实体的识别模型的解码层进行解码, 以获得时间序列中的每一个文字取得的预定义标签。具体地,待识别文本中 的“李”字取得预定义标签“B-PER”,待识别文本中的“先”字取得预定 义标签“I-PER”,待识别文本中的“生”字取得预定义标签“E-PER”,待 识别文本中的“访”字取得预定义标签“0”,待识别文本中的“问”字取得预定义标签“0”,待识别文本中的“中”字取得预定义标签“B-LOC”, 待识别文本中的“国”字取得预定义标签“E-LOC”。其中,时间序列建模 分为时域建模和频域建模两类,一般采用时域建模,需要分析系统的频率特 性时则采用频域建模。时域建模采用曲线拟合和参数估计的方法(如最小二 乘法等),频域建模采用谱分析的方法。时间序列建模主要决定于被观测序 列的性质、可用观测值的数目和模型的使用情况等三个因素。可以理解的是, 以上描述仅为示例性的,本申请实施例对此不做任何限定。

[0070] 在一个具体的例子中,如图2C所示,在图示应用场景中,包括终端设 备B和服务器。其中,终端设备B用于将自身获取的待识别的文本发送至服 务器,服务器用于执行本申请实施例提供的命名实体的识别方法,以识别终 端设备B发送的待识别的文本中的命名实体。

[0071] 当用户需要通过确定待识别的文本中包括的命名实体,来获取该待识别 的文本的相关信息时,用户可以在终端设备B上提供的待识别文本输入栏中 输入待识别的文本,

进而,在终端设备B获取到用户输入的待识别的文本之 后,将该待识别的文本发送至服务器。

[0072] 服务器获取到终端设备B发送的待识别的文本后,对该待识别的文本进 行时间序列建模,得到与之对应的时间序列。然后,服务器将该时间序列输 入至自身运行的命名实体的识别模型中,该命名实体的识别模型通过对输入 的时间序列进行相应的处理,输出各个文字对应的预定义标签。命名实体的 识别模型具体对输入的时间序列进行识别时,通过命名实体的识别模型的编 码层,对时间序列中的每一个文字进行编码,以获得时间序列中的每一个文 字的第一特征数据。在获得时间序列中的每一个文字的第一特征数据之后,通过命名实体的识别模型的全连接层,对时间序列中的每一个文字的第一特 征数据进行映射,以获得与预定义标签的表征数据的数据大小相同的第二特 征数据。与此同时,可对用于识别命名实体的预定义标签中的预定义实体类 型执行层次聚类操作,以获得预定义实体类型之间的层次结构关系。在获得 预定义实体类型之间的层次结构关系之后,可通过命名实体的识别模型的图 神经网络模型,对预定义实体类型和预定义实体类型之间的层次结构关系进 行编码,以获得预定义实体类型的表征数据。在获得预定义实体类型的表征数据之后,可通过命名实体的识别模型的全连接层,对预定义实体类型的表 征数据进行映射,以获得预定义标签的表征数据。在获得预定义标签的表征 数据之后,可对第二特征数据与预定义标签的表征数据进行点乘,以获得时 间序列中的每一个文字取得每一个预定义标签的概率,随后将时间序列中的 每一个文字取得每一个预定义标签的概率输入命名实体的识别模型的解码层 进行解码,以获得时间序列中的每一个文字取得的预定义标签。进而,服务 器根据命名实体的识别模型输出的时间序列中的每一个文字取得的预定义标签,确定待识别的文本中的命名实体。由此,服务器可以向终端设备B返回 命名实体的识别结果。

[0073] 需要说明的是,上述运行于服务器中的命名实体的识别模型采用了图神 经网络模型,对预定义实体类型和预定义实体类型之间的层次结构关系执行 编码操作,以获得预定义实体类型的表征数据。进而,根据预定义实体类型 的表征数据,确定预定义标签的表征数据,从而可通过预定义标签的表征数 据中的共性特征数据,使用属于预定义标签中的非长尾类型的实体的训练效 果,来有效提升属于预定义标签中的长尾类型的实体的训练效果,从而能够 有效提升属于预定义标签中的长尾类型的实体的识别效果。

[0074] 需要说明的是,上述图2C所示场景仅为一种示例,在实际应用中,本 申请实施例提供的命名实体的识别方法还可以应用于终端设备B,在此不对 该命名实体的识别方法的应用场景做任何具体限定。

[0075] 在训练命名实体的识别模型时,可采用编码层、全连接层、图神经网络 模型和解码层联合训练的方式,首先对文本内容进行时间序列建模,再通过 编码层对每个字的上下文信息进行编码,并输出一个向量作为该字的编码。与此同时,与图神经网络模型连接的全连接层输出预定义标签的表征向量。然后,将字的编码与预定义标签的表征向量的点乘结果作为解码层的输入, 通过解码层计算出该字的正确标签所在路径的分数与该文本内容的所有路径 的分数之和的比值作为整个神经网络模型的优化目标,每次针对输入的数据 通过梯度下降算法更新解码层的参数以最大化该目标。可以理解的是,以上 描述仅为示例性的,本申请实施例对此不做任何限定。

[0076] 在实际应用当中,本申请实施例提供的命名实体的识别方法可应用于电 商平台的网页文本的识别。具体地,确定用于电商平台的网页文本的命名实 体识别的预定义标签的数量是否超过预设数量,或者预定义标签中的实体类 型是否存在长尾类型;如果确定预定义标签的数量超过预设数量,或者预定 义标签中的实体类型存在长尾类型,则基于预定义标签和预定义标签之间的 共性关系,确定预定义标签的表征数据,其中,预定义标签的表征数据包括 预定义标签之间的共性特征数据,并通过命名实体的识别模型,至少基于预定义标签的表征数据,识别待识别的网页文本中的命名实体。籍此,在确定 预定义标签的数量超过预设数量,或者预定义标签中的实体类型存在长尾类 型时,充分利用不同的预定义标签之间的共性关系,以获得包括不同的预定 义标签之间的共性特征数据的预定义标签的表征数据,并通过预定义标签的 表征数据中的共性特征数据,使用属于预定义标签中的非长尾类型的实体的 训练效果,来有效提升属于预定义标签中的长尾类型的实体的训练效果,从 而能够有效提升属于预定义标签中的长尾类型的实体的识别效果。可以理解的是,以上描述仅为示例性的,本申请实施例对此不做任何限定。

[0077] 在一个具体的例子中,预设数量可由本领域技术人员根据实际需要进行 设定,本申请实施例对此不做任何限定。长尾类型可理解为预定义标签中的 训练数据小于预设数据量的实体类型。其中,预设数据量可由本领域技术人 员根据实际需要进行设定,本申请实施例对此不做任何限定。

[0078] 通过本申请实施例提供的命名实体的识别方法,基于用于识别命名实体 的预定义标签和预定义标签之间的共性关系,确定预定义标签的表征数据, 其中,预定义标签的表征数据包括预定义标签之间的共性特征数据,并通过 命名实体的识别模型中的编码层,对待识别的文本中的文字进行编码,以获 得待识别的文本中的文字的第一特征数据,再基于预定义标签的表征数据与 文字的第一特征数据,确定文字取得预定义标签的概率,再通过命名实体的 识别模型中的解码层,基于文字取得预定义标签的概率,对待识别的文本中的文字进行解码,以获得待识别的文本中的命名实体,与现有的其它方式相 比,充分利用不同的预定义标签之间的共性关系,以获得包括不同的预定义 标签之间的共性特征数据的预定义标签的表征数据,并通过预定义标签的表 征数据中的共性特征数据,使用属于预定义标签中的非长尾类型的实体的训 练效果,来有效提升属于预定义标签中的长尾类型的实体的训练效果,从而 能够有效提升属于预定义标签中的长尾类型的实体的识别效果。

[0079] 本实施例的命名实体的识别方法可以由任意适当的具有数据处理能力的 设备执行,包括但不限于:摄像头、终端、移动终端、PC机、服务器、车载 设备、娱乐设备、广告设备、个人数码助理(PDA)、平板电脑、笔记本电 脑、掌上游戏机、智能眼镜、智能手表、可穿戴设备、虚拟显示设备或显示 增强设备等。

[0080] 参照图3,示出了本申请实施例三的命名实体的识别模型的训练方法的 步骤流程图。

[0081] 具体地,本实施例提供的命名实体的识别模型的训练方法包括以下步骤:

[0082] 在步骤S301中,通过待训练的命名实体的识别模型,至少基于用于识 别命名实体的预定义标签的表征数据,对文本样本中的命名实体进行识别, 以获得所述文本样本中的命名实体识别数据。

[0083] 在本申请实施例中,预定义标签的表征数据包括预定义标签之间的共性 特征数

据。命名实体识别数据可理解为对文本样本中的命名实体进行识别所 获得的的标签数据，例如，“中”的标签数据为“B-LOC”，“国”的标签 数据为“E-LOC”。可以理解的是，以上描述仅为示例性的，本申请实施例 对此不做任何限定。

[0084] 在一些可选实施例中，在通过待训练的命名实体的识别模型，至少基于 用于识别命名实体的预定义标签的表征数据，对文本样本中的命名实体进行 识别时，通过命名实体的识别模型中的编码层，对文本样本中的文字进行编 码，以获得文本样本中的文字的第一特征数据；基于预定义标签的表征数据 与文字的第一特征数据，确定文字取得对应预定义标签的概率；通过命名实 体的识别模型中的解码层，基于文字取得对应预定义标签的概率，对文本样 本中的文字进行解码，以获得文本样本中的命名实体识别数据。可以理解的是，以上描述仅为示例性的，本申请实施例对此不做任何限定。

[0085] 在一些可选实施例中，在基于预定义标签的表征数据与文字的第一特征 数据，确定文字取得对应预定义标签的概率时，通过命名实体的识别模型中 的全连接层，对文字的第一特征数据进行映射，以获得与预定义标签的表征 数据的数据大小相同的文字的第二特征数据；对预定义标签的表征数据与文 字的第二特征数据进行点乘，以获得文字取得对应预定义标签的概率。可以 理解的是，以上描述仅为示例性的，本申请实施例对此不做任何限定。

[0086] 在步骤S302中，基于文本样本中的命名实体识别数据和命名实体标注 数据，对待训练的命名实体的识别模型进行训练。

[0087] 在本申请实施例中，命名实体标注数据可理解为对文本样本中的命名实 体进行标注所获得的的标签数据，例如，“奥”的标签数据为“B-PER”， “巴”的标签数据为“I-PER”。可以理解的是，以上描述仅为示例性的，本 申请实施例对此不做任何限定。

[0088] 在一些可选实施例中，在基于文本样本中的命名实体识别数据和命名实 体标注数据，对待训练的命名实体的识别模型进行训练时，通过目标损失函 数，确定命名实体识别数据和命名实体标注数据之间的差异值；基于差异值， 调整待训练的命名实体的识别模型的模型参数。其中，目标损失函数可为交 叉熵损失函数、softmax损失函数、L1损失函数、L2损失函数等任意损失函 数。在调整待训练的命名实体的识别模型的模型参数时，可采用反向传播算 法，或者随机梯度下降算法来调整命名实体的识别模型的模型参数。可以理解的是，以上描述仅为示例性的，本申请实施例对此不做任何限定。

[0089] 在一个具体的例子中，通过确定命名实体识别数据和命名实体标注数据 之间的差异值，对当前获得的命名实体识别数据进行评估，以作为后续训练 命名实体的识别模型的依据。具体地，可将差异值反向传输给命名实体的识 别模型，从而迭代地训练命名实体的识别模型。命名实体的识别模型的训练 是一个迭代的过程，本申请实施例仅对其中的一次训练过程进行了说明，但 本领域技术人员应当明了，对命名实体的识别模型的每次训练都可采用该训 练方式，直至完成命名实体的识别模型的训练。可以理解的是，以上描述仅为示例性的，本申请实施例对此不做任何限定。

[0090] 根据本申请实施例提供的命名实体的识别模型的训练方法，通过待训练 的命名实体的识别模型，至少基于用于识别命名实体的预定义标签的表征数 据，对文本样本中的命名实体进行识别，以获得文本样本中的命名实体识别 数据，其中，预定义标签的表征数据包括预定义标签之间的共性特征数据， 并基于文本样本中的命名实体识别数据和命名

实体标注数据,对待训练的命 名实体的识别模型进行训练,与现有的其它方式相比,通过待训练的命名实 体的识别模型,至少基于用于识别命名实体的预定义标签的表征数据,对文 本样本中的命名实体进行识别,以获得文本样本中的命名实体识别数据,并 基于文本样本中的命名实体识别数据和命名实体标注数据,对待训练的命名 实体的识别模型进行训练,能够使得训练得到的命名实体的识别模型针对命 名实体具有更强的识别性能。

[0091] 本实施例的命名实体的识别模型的训练方法可以由任意适当的具有数据 处理能力的设备执行,包括但不限于:摄像头、终端、移动终端、PC机、服 务器、车载设备、娱乐设备、广告设备、个人数码助理(PDA)、平板电脑、 笔记本电脑、掌上游戏机、智能眼镜、智能手表、可穿戴设备、虚拟显示设 备或显示增强设备等。

[0092] 参照图4,示出了本申请实施例四的图神经网络模型的训练方法的步骤 流程图。

[0093] 具体地,本实施例提供的图神经网络模型的训练方法包括以下步骤:

[0094] 在步骤S401中,通过待训练的图神经网络模型,对图结构数据样本执 行编码操作,以获得图结构数据样本的节点的结构特征表征数据。

[0095] 在本申请实施例中,图结构数据样本的节点和边分别表示预定义实体类 型和预定义实体类型之间的共性关系。结构特征表征数据可理解为用于表征 节点在图结构数据样本中的结构特征的数据,例如,结构特征表征向量。可 以理解的是,以上描述仅为示例性的,本申请实施例对此不做任何限定。

[0096] 在步骤S402中,基于所述图结构数据样本的节点的结构特征表征数据 和结构特征标注数据,对待训练的所述图神经网络模型进行训练。

[0097] 在本申请实施例中,结构特征标注数据可理解为用于标注节点在图结构 数据样本中的结构特征的数据,例如,结构特征标注向量。可以理解的是, 以上描述仅为示例性的,本申请实施例对此不做任何限定。

[0098] 在一些可选实施例中,在基于图结构数据样本的节点的结构特征表征数 据和结构特征标注数据,对待训练的所述图神经网络模型进行训练时,通过 目标损失函数,确定图结构数据样本的节点的结构特征表征数据和结构特征 标注数据之间的差异值;基于差异值,调整待训练的图神经网络模型的模型 参数。其中,目标损失函数可为交叉熵损失函数、softmax损失函数、L1损 失函数、L2损失函数等任意损失函数。在调整待训练的图神经网络模型的模 型参数时,可采用反向传播算法,或者随机梯度下降算法来调整图神经网络模型的模型参数。可以理解的是,以上描述仅为示例性的,本申请实施例对 此不做任何限定。

[0099] 在一个具体的例子中,通过确定结构特征表征数据和结构特征标注数据 之间的差异值,对当前获得的结构特征表征数据进行评估,以作为后续训练 图神经网络模型的依据。具体地,可将差异值反向传输给图神经网络模型, 从而迭代地训练图神经网络模型。图神经网络模型的训练是一个迭代的过程, 本申请实施例仅对其中的一次训练过程进行了说明,但本领域技术人员应当 明了,对图神经网络模型的每次训练都可采用该训练方式,直至完成图神经 网络模型的训练。可以理解的是,以上描述仅为示例性的,本申请实施例对 此不做任何限定。

[0100] 根据本申请实施例提供的图神经网络模型的训练方法,通过待训练的图 神经网络模型,对图结构数据样本执行编码操作,以获得图结构数据样本的 节点的结构特征表征

数据,其中,图结构数据样本的节点和边分别表示预定 义实体类型和预定义实体类型之间的共性关系,并基于图结构数据样本的节 点的结构特征表征数据和结构特征标注数据,对待训练的图神经网络模型进 行训练,与现有的其它方式相比,通过待训练的图神经网络模型,对图结构 数据样本执行编码操作,以获得图结构数据样本的节点的结构特征表征数据, 并基于图结构数据样本的节点的结构特征表征数据和结构特征标注数据,对 待训练的图神经网络模型进行训练,能够使得训练得到的图神经网络模型针 对图结构数据具有更强的编码性能。

[0101] 本实施例的图神经网络模型的训练方法可以由任意适当的具有数据处理 能力的设备执行,包括但不限于:摄像头、终端、移动终端、PC机、服务器、 车载设备、娱乐设备、广告设备、个人数码助理(PDA)、平板电脑、笔记 本电脑、掌上游戏机、智能眼镜、智能手表、可穿戴设备、虚拟显示设备或 显示增强设备等。

[0102] 参照图5,示出了本申请实施例五的命名实体的识别方法的步骤流程图。

[0103] 具体地,本实施例提供的命名实体的识别方法包括以下步骤:

[0104] 在步骤S501中,基于用于识别命名实体的预定义标签和预定义标签之 间的共性关系,确定预定义标签的表征数据。

[0105] 其中,预定义标签的表征数据包括预定义标签之间的共性特征数据。可 以理解的是,以上描述仅为示例性的,本申请实施例对此不做任何限定。

[0106] 在步骤S502中,通过命名实体的识别模型,至少基于预定义标签的表 征数据,识别待识别的文本中属于长尾类型的命名实体。

[0107] 在本申请实施例中,长尾类型可理解为预定义标签中的训练数据小于预 设数据量的实体类型。可以理解的是,以上描述仅为示例性的,本申请实施 例对此不做任何限定。

[0108] 通过本申请实施例提供的命名实体的识别方法,基于用于识别命名实体 的预定义标签和预定义标签之间的共性关系,确定预定义标签的表征数据, 其中,预定义标签的表征数据包括预定义标签之间的共性特征数据,并通过 命名实体的识别模型,至少基于预定义标签的表征数据,识别待识别的文本 中属于长尾类型的命名实体,与现有的其它方式相比,充分利用不同的预定 义标签之间的共性关系,以获得包括不同的预定义标签之间的共性特征数据 的预定义标签表征数据,并通过预定义标签表征数据中的共性特征数据,使用属于预定义标签中的非长尾类型的实体的训练效果,来有效提升属于预定 义标签中的长尾类型的实体的训练效果,从而能够来有效提升属于预定义标 签中的长尾类型的实体的识别效果。

[0109] 本实施例的命名实体的识别方法可以由任意适当的具有数据处理能力的 设备执行,包括但不限于:摄像头、终端、移动终端、PC机、服务器、车载 设备、娱乐设备、广告设备、个人数码助理(PDA)、平板电脑、笔记本电 脑、掌上游戏机、智能眼镜、智能手表、可穿戴设备、虚拟显示设备或显示 增强设备等。

[0110] 参照图6,示出了本申请实施例六的命名实体的识别方法的步骤流程图。

[0111] 具体地,本实施例提供的命名实体的识别方法包括以下步骤:

[0112] 在步骤S601中,通过图神经网络模型,对图结构数据执行编码操作, 以获得图结构数据的节点的结构特征表征数据。

[0113] 在本申请实施例中,图结构数据的节点和边分别表示预定义实体类型和 预定义

实体类型之间的共性关系。结构特征表征数据可理解为用于表征节点 在图结构数据样本中的结构特征的数据,例如,结构特征表征向量。可以理 解的是,以上描述仅为示例性的,本申请实施例对此不做任何限定。

[0114] 在步骤S602中,将图结构数据的节点的结构特征表征数据作为图结构 数据的节点表示的预定义实体类型的表征数据。

[0115] 在本申请实施例中,预定义实体类型的表征数据包括预定义实体类型之 间的共性特征数据。可以理解的是,以上描述仅为示例性的,本申请实施例 对此不做任何限定。

[0116] 在步骤S603中,通过命名实体的识别模型,至少基于预定义实体类型 的表征数据,识别待识别的文本中属于长尾类型的命名实体。

[0117] 在本申请实施例中,长尾类型可理解为预定义标签中的训练数据小于预 设数据量的实体类型。可以理解的是,以上描述仅为示例性的,本申请实施 例对此不做任何限定。

[0118] 根据本申请实施例提供的命名实体的识别方法,通过图神经网络模型, 对图结构数据执行编码操作,以获得图结构数据的节点的结构特征表征数据, 其中,图结构数据的节点和边分别表示预定义实体类型和预定义实体类型之 间的共性关系,并将图结构数据的节点的结构特征表征数据作为图结构数据 的节点表示的预定义实体类型的表征数据,其中,预定义实体类型的表征数 据包括预定义实体类型之间的共性特征数据,并通过命名实体的识别模型, 至少基于预定义实体类型的表征数据,识别待识别的文本中属于长尾类型的 命名实体,与现有的其它方式相比,充分利用不同的预定义标签之间的共性 关系,以获得包括不同的预定义标签之间的共性特征数据的预定义标签表征 数据,并通过预定义标签表征数据中的共性特征数据,使用属于预定义标签 中的非长尾类型的实体的训练效果,来有效提升属于预定义标签中的长尾类 型的实体的训练效果,从而能够来有效提升属于预定义标签中的长尾类型的 实体的识别效果。

[0119] 本实施例的命名实体的识别方法可以由任意适当的具有数据处理能力的 设备执行,包括但不限于:摄像头、终端、移动终端、PC机、服务器、车载 设备、娱乐设备、广告设备、个人数码助理(PDA)、平板电脑、笔记本电 脑、掌上游戏机、智能眼镜、智能手表、可穿戴设备、虚拟显示设备或显示 增强设备等。

[0120] 参照图7,示出了本申请实施例七的命名实体的识别方法的步骤流程图。

[0121] 具体地,本实施例提供的命名实体的识别方法包括以下步骤:

[0122] 在步骤S701中,基于用于识别命名实体的预定义标签和预定义标签之 间的共性关系,确定预定义标签的表征数据。

[0123] 在本申请实施例中,预定义标签的表征数据包括预定义标签之间的共性 特征数据。可以理解的是,以上描述仅为示例性的,本申请实施例对此不做 任何限定。

[0124] 在步骤S702中,通过命名实体的识别模型,至少基于预定义标签的表 征数据,识别待识别的案件图谱中属于长尾类型的命名实体。

[0125] 在本申请实施例中,长尾类型可理解为预定义标签中的训练数据小于预 设数据量的实体类型。案件图谱可为司法案件图谱、民法案件图谱、刑罚案 件图谱等。可以理解的是,以上描述仅为示例性的,本申请实施例对此不做 任何限定。

[0126] 通过本申请实施例提供的命名实体的识别方法,基于用于识别命名实体 的预定义标签和预定义标签之间的共性关系,确定预定义标签的表征数据, 其中,预定义标签的

表征数据包括预定义标签之间的共性特征数据,并通过 命名实体的识别模型,至少基于预定义标签的表征数据,识别待识别的案件 图谱中属于长尾类型的命名实体,与现有的其它方式相比,充分利用不同的 预定义标签之间的共性关系,以获得包括不同的预定义标签之间的共性特征 数据的预定义标签表征数据,并通过预定义标签表征数据中的共性特征数据, 使用案件图谱中属于预定义标签中的非长尾类型的实体的训练效果,来有效 提升案件图谱中属于预定义标签中的长尾类型的实体的训练效果,从而能够 来有效提升案件图谱中属于预定义标签中的长尾类型的实体的识别效果。

[0127] 本实施例的命名实体的识别方法可以由任意适当的具有数据处理能力的 设备执行,包括但不限于:摄像头、终端、移动终端、PC机、服务器、车载 设备、娱乐设备、广告设备、个人数码助理(PDA)、平板电脑、笔记本电 脑、掌上游戏机、智能眼镜、智能手表、可穿戴设备、虚拟显示设备或显示 增强设备等。

[0128] 参照图8,示出了本申请实施例八的命名实体的识别方法的步骤流程图。

[0129] 具体地,本实施例提供的命名实体的识别方法包括以下步骤:

[0130] 在步骤S801中,基于用于识别命名实体的预定义标签和预定义标签之 间的共性关系,确定预定义标签的表征数据。

[0131] 在本申请实施例中,预定义标签的表征数据包括预定义标签之间的共性 特征数据。可以理解的是,以上描述仅为示例性的,本申请实施例对此不做 任何限定。

[0132] 在步骤S802中,通过命名实体的识别模型,至少基于预定义标签的表 征数据,识别待识别的公诉书中属于长尾类型的命名实体。

[0133] 在本申请实施例中,长尾类型可理解为预定义标签中的训练数据小于预 设数据量的实体类型。公诉书可理解为人民检察院依照法定的诉讼程序代表 国家对被告人向人民法院提起诉讼的文书。可以理解的是,以上描述仅为示 例性的,本申请实施例对此不做任何限定。

[0134] 通过本申请实施例提供的命名实体的识别方法,基于用于识别命名实体 的预定义标签和预定义标签之间的共性关系,确定预定义标签的表征数据, 其中,预定义标签的表征数据包括预定义标签之间的共性特征数据,并通过 命名实体的识别模型,至少基于预定义标签的表征数据,识别待识别的公诉 书中属于长尾类型的命名实体,与现有的其它方式相比,充分利用不同的预 定义标签之间的共性关系,以获得包括不同的预定义标签之间的共性特征数 据的预定义标签表征数据,并通过预定义标签表征数据中的共性特征数据, 使用公诉书中属于预定义标签中的非长尾类型的实体的训练效果,来有效提 升公诉书中属于预定义标签中的长尾类型的实体的训练效果,从而能够有效 提升公诉书中属于预定义标签中的长尾类型的实体的识别效果。

[0135] 本实施例的命名实体的识别方法可以由任意适当的具有数据处理能力的 设备执行,包括但不限于:摄像头、终端、移动终端、PC机、服务器、车载 设备、娱乐设备、广告设备、个人数码助理(PDA)、平板电脑、笔记本电 脑、掌上游戏机、智能眼镜、智能手表、可穿戴设备、虚拟显示设备或显示 增强设备等。

[0136] 参照图9,示出了本申请实施例九中命名实体的识别装置的结构示意图。

[0137] 本实施例的命名实体的识别装置包括:确定模块901,用于基于用于识 别命名实体的预定义标签和所述预定义标签之间的共性关系,确定所述预定 义标签的表征数据,其

中,所述预定义标签的表征数据包括所述预定义标签 之间的共性特征数据;识别模块902,用于通过命名实体的识别模型,至少 基于所述预定义标签的表征数据,识别待识别的文本中的命名实体。

[0138] 本实施例的命名实体的识别装置用于实现前述多个方法实施例中相应的 命名实体的识别方法,并具有相应的方法实施例的有益效果,在此不再赘述。

[0139] 参照图10,示出了本申请实施例十中命名实体的识别装置的结构示意 图。

[0140] 本实施例的命名实体的识别装置包括:确定模块1001,用于基于用于识 别命名实体的预定义标签和所述预定义标签之间的共性关系,确定所述预定 义标签的表征数据,其中,所述预定义标签的表征数据包括所述预定义标签 之间的共性特征数据;识别模块1002,用于通过命名实体的识别模型,至少 基于所述预定义标签的表征数据,识别待识别的文本中的命名实体。

[0141] 可选地,所述确定模块1001,包括:第一确定子模块10013,用于基于 所述预定义标签包括的预定义实体类型和所述预定义实体类型之间的共性关 系,确定所述预定义实体类型的表征数据,其中,所述预定义实体类型的表 征数据包括所述预定义实体类型之间的共性特征数据。

[0142] 可选地,所述第一确定子模块10013之前,所述确定模块1001还包括: 提取子模块10011,用于对所述预定义实体类型执行共性关系的提取操作, 以获得所述预定义实体类型的共性结构;第二确定子模块10012,用于确定 所述共性结构表示的预定义实体类型之间的共性关系为所述预定义实体类型 之间的共性关系。

[0143] 可选地,所述第一确定子模块10013,具体用于:将所述预定义实体类 型和所述预定义实体类型之间的共性关系分别作为图结构数据的节点和边; 通过图神经网络模型,对所述图结构数据执行编码操作,以获得所述图结构 数据的节点的表征数据;将所述图结构数据的节点的表征数据作为所述图结 构数据的节点表示的预定义实体类型的表征数据。

[0144] 可选地,所述识别模块1002,包括:编码子模块10021,用于通过所述 命名实体的识别模型中的编码层,对所述待识别的文本中的文字进行编码, 以获得所述待识别的文本中的文字的第一特征数据;第三确定子模块10022, 用于基于所述预定义标签的表征数据与所述文字的第一特征数据,确定所述 文字取得对应所述预定义标签的概率;解码子模块10023,用于通过所述命 名实体的识别模型中的解码层,基于所述文字取得对应所述预定义标签的概 率,对所述待识别的文本中的文字进行解码,以获得所述待识别的文本中的命名实体。

[0145] 可选地,所述第三确定子模块10022,具体用于:通过所述命名实体的 识别模型中的全连接层,对所述文字的第一特征数据进行映射,以获得与所 述预定义标签的表征数据的数据大小相同的所述文字的第二特征数据;对所 述预定义标签的表征数据与所述文字的第二特征数据进行点乘,以获得所述 文字取得对应所述预定义标签的概率。

[0146] 本实施例的命名实体的识别装置用于实现前述多个方法实施例中相应的 命名实体的识别方法,并具有相应的方法实施例的有益效果,在此不再赘述。

[0147] 图11为本申请实施例十一中电子设备的结构示意图;该电子设备可以包 括:

[0148] 一个或多个处理器1101;

[0149] 计算机可读介质1102,可以配置为存储一个或多个程序,

[0150] 当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或 多个处理器实现如上述实施例一、实施例二、实施例五、实施例六、实施例 七,或者实施例八所述的命名实体的识别方法,实现如上述实施例三所述的 命名实体的识别模型的训练方法,或者实现如上述实施例四所述的图神经网 络模型的训练方法。

[0151] 图12为本申请实施例十二中电子设备的硬件结构;如图12所示,该电 子设备的硬件结构可以包括:处理器1201,通信接口1202,计算机可读介质 1203和通信总线1204;

[0152] 其中处理器1201、通信接口1202、计算机可读介质1203通过通信总线 1204完成相互间的通信;

[0153] 可选地,通信接口1202可以为通信模块的接口,如GSM模块的接口;

[0154] 其中,处理器1201具体可以配置为:基于用于识别命名实体的预定义标 签和所述预定义标签之间的共性关系,确定所述预定义标签的表征数据,其 中,所述预定义标签的表征数据包括所述预定义标签之间的共性特征数据; 通过命名实体的识别模型,至少基于所述预定义标签的表征数据,识别待识 别的文本中的命名实体。此外,处理器1201还可以配置为:确定预定义标签 的数量是否超过预设数量,或者所述预定义标签中的实体类型是否存在长尾 类型,其中,所述预定义标签用于识别电商平台的网页文本中的命名实体; 如果确定所述预定义标签的数量超过所述预设数量,或者所述预定义标签中 的实体类型存在所述长尾类型,则基于所述预定义标签和所述预定义标签之 间的共性关系,确定所述预定义标签的表征数据,其中,所述预定义标签的 表征数据包括所述预定义标签之间的共性特征数据,并通过命名实体的识别 模型,至少基于所述预定义标签的表征数据,识别待识别的所述网页文本中 的命名实体。此外,处理器1201还可以配置为:通过待训练的命名实体的识 别模型,至少基于用于识别命名实体的预定义标签的表征数据,对文本样本 中的命名实体进行识别,以获得所述文本样本中的命名实体识别数据,其中, 所述预定义标签的表征数据包括所述预定义标签之间的共性特征数据;基于 所述文本样本中的命名实体识别数据和命名实体标注数据,对待训练的所述 命名实体的识别模型进行训练。此外,处理器1201还可以配置为:通过待训 练的图神经网络模型,对图结构数据样本执行编码操作,以获得所述图结构 数据样本的节点的结构特征表征数据,其中,所述图结构数据样本的节点和 边分别表示预定义实体类型和所述预定义实体类型之间的共性关系;基于所 述图结构数据样本的节点的结构特征表征数据和结构特征标注数据,对待训 练的所述图神经网络模型进行训练。此外,处理器1201还可以配置为:基于 用于识别命名实体的预定义标签和所述预定义标签之间的共性关系,确定所 述预定义标签的表征数据,其中,所述预定义标签的表征数据包括所述预定 义标签之间的共性特征数据;通过命名实体的识别模型,至少基于所述预定 义标签的表征数据,识别待识别的文本中属于长尾类型的命名实体。此外, 处理器1201还可以配置为:通过图神经网络模型,对图结构数据执行编码操 作,以获得所述图结构数据的节点的结构特征表征数据,其中,所述图结构 数据的节点和边分别表示预定义实体类型和所述预定义实体类型之间的共性 关系;将所述图结构数据的节点的结构特征表征数据作为所述图结构数据的 节点表示的预定义实体类型的表征数据,其中,所述预定义实体类型的表征 数据包括所述预定义实体类型之间的共性特征数据;通过命名实体的识别模 型,至少基于所述预定义实体类型的表征数据,识别待识别的文本中属于长 尾类型的命名实体。此外,处理器1201还可以配置为:基于用于识别命名实 体的预

定义标签和所述预定义标签之间的共性关系,确定所述预定义标签的 表征数据,其中,所述预定义标签的表征数据包括所述预定义标签之间的共 性特征数据;通过命名实体的识别模型,至少基于所述预定义标签的表征数 据,识别待识别的案件图谱中属于长尾类型的命名实体。此外,处理器1201 还可以配置为:基于用于识别命名实体的预定义标签和所述预定义标签之间 的共性关系,确定所述预定义标签的表征数据,其中,所述预定义标签的表 征数据包括所述预定义标签之间的共性特征数据;通过命名实体的识别模型, 至少基于所述预定义标签的表征数据,识别待识别的起诉书中属于长尾类型 的命名实体。

[0155] 处理器1201可以是通用处理器,包括中央处理器(Central Processing Unit,简称CPU)、网络处理器(Network Processor,简称NP)等;还可以是数 字信号处理器(DSP)、专用集成电路(ASIC)、现成可编程门阵列(FPGA)或者 其它可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。可以实 现或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。通用处理器 可以是微处理器或者该处理器也可以是任何常规的处理器等。

[0156] 计算机可读介质1203可以是,但不限于,随机存取存储介质(Random Access Memory,RAM),只读存储介质(Read Only Memory,ROM),可 编程只读存储介质(Programmable Read-Only Memory,PROM),可擦除只 读存储介质(Erasable Programmable Read-Only Memory,EPROM),电可擦 除只读存储介质(Electric Erasable Programmable Read-Only Memory, EEPROM)等。

[0157] 特别地,根据本公开的实施例,上文参考流程图描述的过程可以被实现 为计算机软件程序。例如,本公开的实施例包括一种计算机程序产品,其包 括承载在计算机可读介质上的计算机程序,该计算机程序包含配置为执行流 程图所示的方法的程序代码。在这样的实施例中,该计算机程序可以通过通 信部分从网络上被下载和安装,和/或从可拆卸介质被安装。在该计算机程序 被中央处理单元(CPU)执行时,执行本申请的方法中限定的上述功能。需 要说明的是,本申请所述的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读介质例如可以 但不限于是电、磁、光、电磁、红外线、或半导体的系统、装置或器件,或 者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于: 具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储介 质(RAM)、只读存储介质(ROM)、可擦式可编程只读存储介质(EPROM 或闪存)、光纤、便携式紧凑磁盘只读存储介质(CD-ROM)、光存储介质 件、磁存储介质件、或者上述的任意合适的组合。在本申请中,计算机可读 存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系 统、装置或者器件使用或者与其结合使用。而在本申请中,计算机可读的信 号介质可以包括在基带中或者作为载波一部分传播的数据信号,其中承载了 计算机可读的程序代码。这种传播的数据信号可以采用多种形式,包括但不 限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还 可以是计算机可读存储介质以外的任何计算机可读介质,该计算机可读介质 可以发送、传播或者传输配置为由指令执行系统、装置或者器件使用或者与 其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介 质传输,包括但不限于:无线、电线、光缆、RF等等,或者上述的任意合适 的组合。

[0158] 可以以一种或多种程序设计语言或其组合来编写配置为执行本申请的操 作的计算机程序代码,所述程序设计语言包括面向对象的程序设计语言—诸 如Java、Smalltalk、

C++,还包括常规的过程式程序设计语言—诸如"C"语言 或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地 在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部 分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远 程计算机的情形中,远程计算机可以通过任意种类的网络:包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如 利用因特网服务提供商来通过因特网连接)。

[0159] 附图中的流程图和框图,图示了按照本申请各种实施例的系统、方法和 计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或 框图中的每个方框可以代表一个模块、程序段、或代码的一部分,该模块、 程序段、或代码的一部分包含一个或多个配置为实现规定的逻辑功能的可执 行指令。上述具体实施例中有特定先后关系,但这些先后关系只是示例性的, 在具体实现的时候,这些步骤可能会更少、更多或执行顺序有调整。即在有 些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际上可以基本并行地执行,它们 有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框 图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以 用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用 硬件与计算机指令的组合来实现。

[0160] 描述于本申请实施例中所涉及到的模块可以通过软件的方式实现,也可 以通过硬件的方式来实现。所描述的模块也可以设置在处理器中,例如,可 以描述为:一种处理器包括确定模块和识别模块。其中,这些模块的名称在 某种情况下并不构成对该模块本身的限定,例如,确定模块还可以被描述为 "基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系,确 定所述预定义标签的表征数据的模块"。

[0161] 作为另一方面,本申请还提供了一种计算机可读介质,其上存储有计算 机程序,该程序被处理器执行时实现如上述实施例一、实施例二、实施例五、 实施例六、实施例七,或者实施例八所描述的命名实体的识别方法,实现如 上述实施例三所述的命名实体的识别模型的训练方法,或者实现如上述实施 例四所述的图神经网络模型的训练方法。

[0162] 作为另一方面,本申请还提供了一种计算机可读介质,该计算机可读介 质可以是上述实施例中描述的装置中所包含的;也可以是单独存在,而未装 配入该装置中。上述计算机可读介质承载有一个或者多个程序,当上述一个 或者多个程序被该装置执行时,使得该装置:基于用于识别命名实体的预定 义标签和所述预定义标签之间的共性关系,确定所述预定义标签的表征数据, 其中,所述预定义标签的表征数据包括所述预定义标签之间的共性特征数据; 通过命名实体的识别模型,至少基于所述预定义标签的表征数据,识别待识 别的文本中的命名实体。此外,还使得该装置:确定预定义标签的数量是否 超过预设数量,或者所述预定义标签中的实体类型是否存在长尾类型,其中, 所述预定义标签用于识别电商平台的网页文本中的命名实体;如果确定所述 预定义标签的数量超过所述预设数量,或者所述预定义标签中的实体类型存 在所述长尾类型,则基于所述预定义标签和所述预定义标签之间的共性关系, 确定所述预定义标签的表征数据,其中,所述预定义标签的表征数据包括所 述预定义标签之间的共性特征数据,并通过命名实体的识别模型,至少基于 所述预定义标签的表征数据,识别待识别的所述网页文本中的命名实体。此 外,还使得该装置:通过待训练的命名实体的识别模型,至少基于用于识别 命名实体的预定义标签的

表征数据,对文本样本中的命名实体进行识别,以 获得所述文本样本中的命名实体识别数据,其中,所述预定义标签的表征数 据包括所述预定义标签之间的共性特征数据;基于所述文本样本中的命名实 体识别数据和命名实体标注数据,对待训练的所述命名实体的识别模型进行 训练。此外,还使得该装置:通过待训练的图神经网络模型,对图结构数据 样本执行编码操作,以获得所述图结构数据样本的节点的结构特征表征数据, 其中,所述图结构数据样本的节点和边分别表示预定义实体类型和所述预定 义实体类型之间的共性关系;基于所述图结构数据样本的节点的结构特征表 征数据和结构特征标注数据,对待训练的所述图神经网络模型进行训练。此 外,还使得该装置:基于用于识别命名实体的预定义标签和所述预定义标签 之间的共性关系,确定所述预定义标签的表征数据,其中,所述预定义标签 的表征数据包括所述预定义标签之间的共性特征数据;通过命名实体的识别 模型,至少基于所述预定义标签的表征数据,识别待识别的文本中属于长尾 类型的命名实体。此外,还使得该装置:通过图神经网络模型,对图结构数 据执行编码操作,以获得所述图结构数据的节点的结构特征表征数据,其中, 所述图结构数据的节点和边分别表示预定义实体类型和所述预定义实体类型 之间的共性关系;将所述图结构数据的节点的结构特征表征数据作为所述图 结构数据的节点表示的预定义实体类型的表征数据,其中,所述预定义实体 类型的表征数据包括所述预定义实体类型之间的共性特征数据;通过命名实 体的识别模型,至少基于所述预定义实体类型的表征数据,识别待识别的文 本中属于长尾类型的命名实体。此外,还使得该装置:基于用于识别命名实 体的预定义标签和所述预定义标签之间的共性关系,确定所述预定义标签的 表征数据,其中,所述预定义标签的表征数据包括所述预定义标签之间的共 性特征数据;通过命名实体的识别模型,至少基于所述预定义标签的表征数 据,识别待识别的案件图谱中属于长尾类型的命名实体。此外,还使得该装 置:基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系, 确定所述预定义标签的表征数据,其中,所述预定义标签的表征数据包括所 述预定义标签之间的共性特征数据;通过命名实体的识别模型,至少基于所 述预定义标签的表征数据,识别待识别的起诉书中属于长尾类型的命名实体。

[0163] 在本公开的各种实施方式中所使用的表述“第一”、“第二”、“所述第一” 或“所述第二”可修饰各种部件而与顺序和/或重要性无关,但是这些表述不限 制相应部件。以上表述仅配置为将元件与其它元件区分开的目的。例如,第 一用户设备和第二用户设备表示不同的用户设备,虽然两者均是用户设备。 例如,在不背离本公开的范围的前提下,第一元件可称作第二元件,类似地, 第二元件可称作第一元件。

[0164] 当一个元件(例如,第一元件)称为与另一元件(例如,第二元件)“(可操作 地或可通信地)联接”或“(可操作地或可通信地)联接至”另一元件(例如,第二 元件)或“连接至”另一元件(例如,第二元件)时,应理解为该一个元件直接连 接至该另一元件或者该一个元件经由又一个元件(例如,第三元件)间接连接 至该另一个元件。相反,可理解,当元件(例如,第一元件)称为“直接连接” 或“直接联接”至另一元件(第二元件)时,则没有元件(例如,第三元件)插入在 这两者之间。

[0165] 以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领 域技术人员应当理解,本申请中所涉及的发明范围,并不限于上述技术特征 的特定组合而成的技术方案,同时也应涵盖在不脱离上述发明构思的情况下, 由上述技术特征或其等同特征进

行任意组合而形成的其它技术方案。例如上 述特征与本申请中公开的(但不限于)具有类似功能的技术特征进行互相替 换而形成的技术方案。

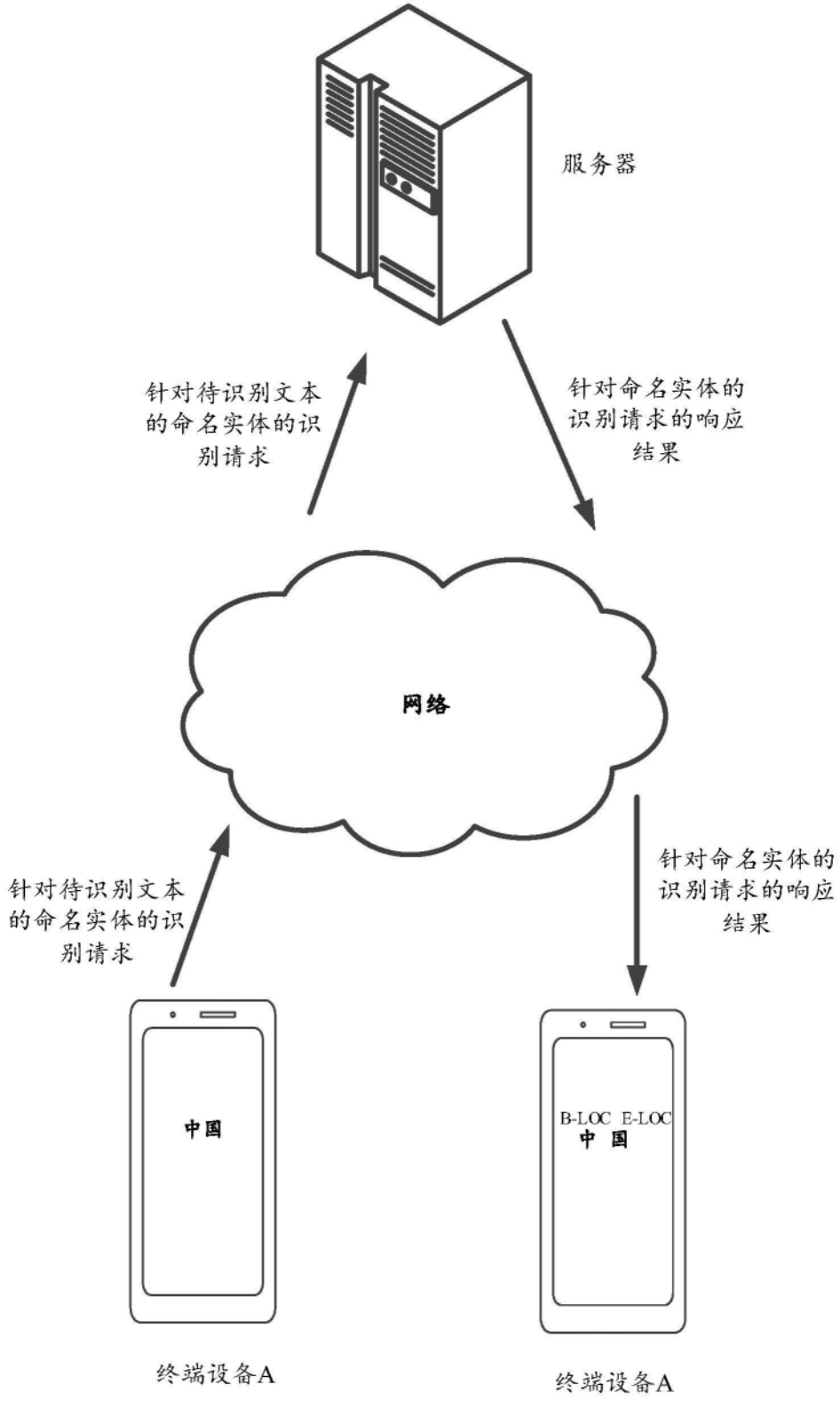

服务器
针对待识别文本的命名实体的识别请求
针对命名实体的识别请求的响应结果
网络
针对待识别文本的命名实体的识别请求
针对命名实体的识别请求的响应结果
中国
B-LOC E-LOC
中 国
终端设备A
终端设备A

图1A

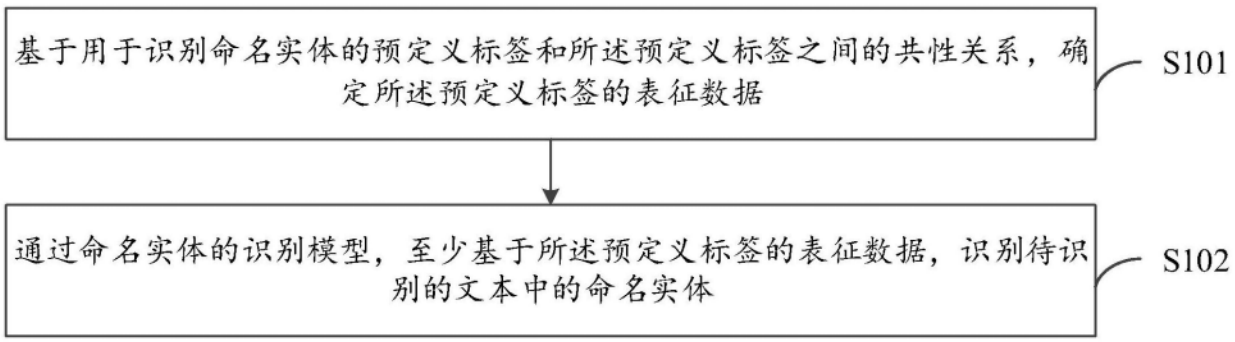

基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系，确定所述预定义标签的表征数据
S101
通过命名实体的识别模型，至少基于所述预定义标签的表征数据，识别待识别的文本中的命名实体
S102

图1B

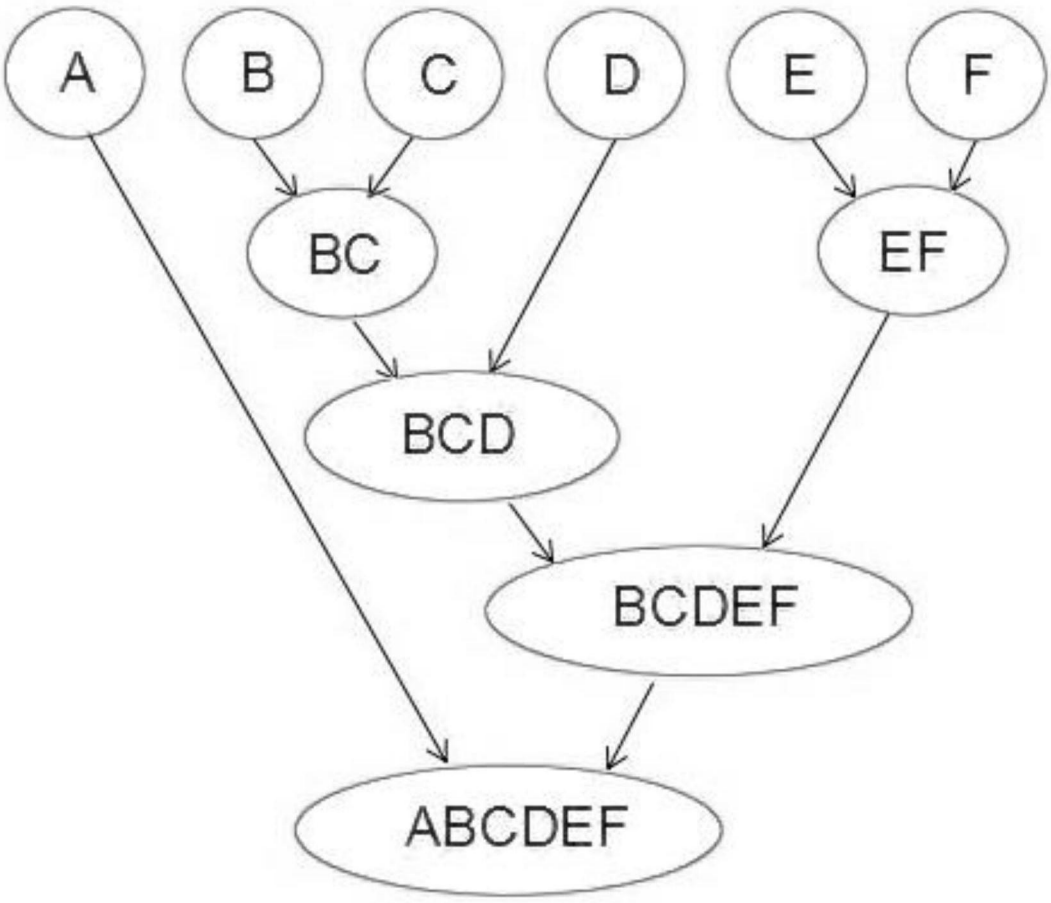

A
B
C
D
E
F
BC
EF
BCD
BCDEF
ABCDEF

图1C

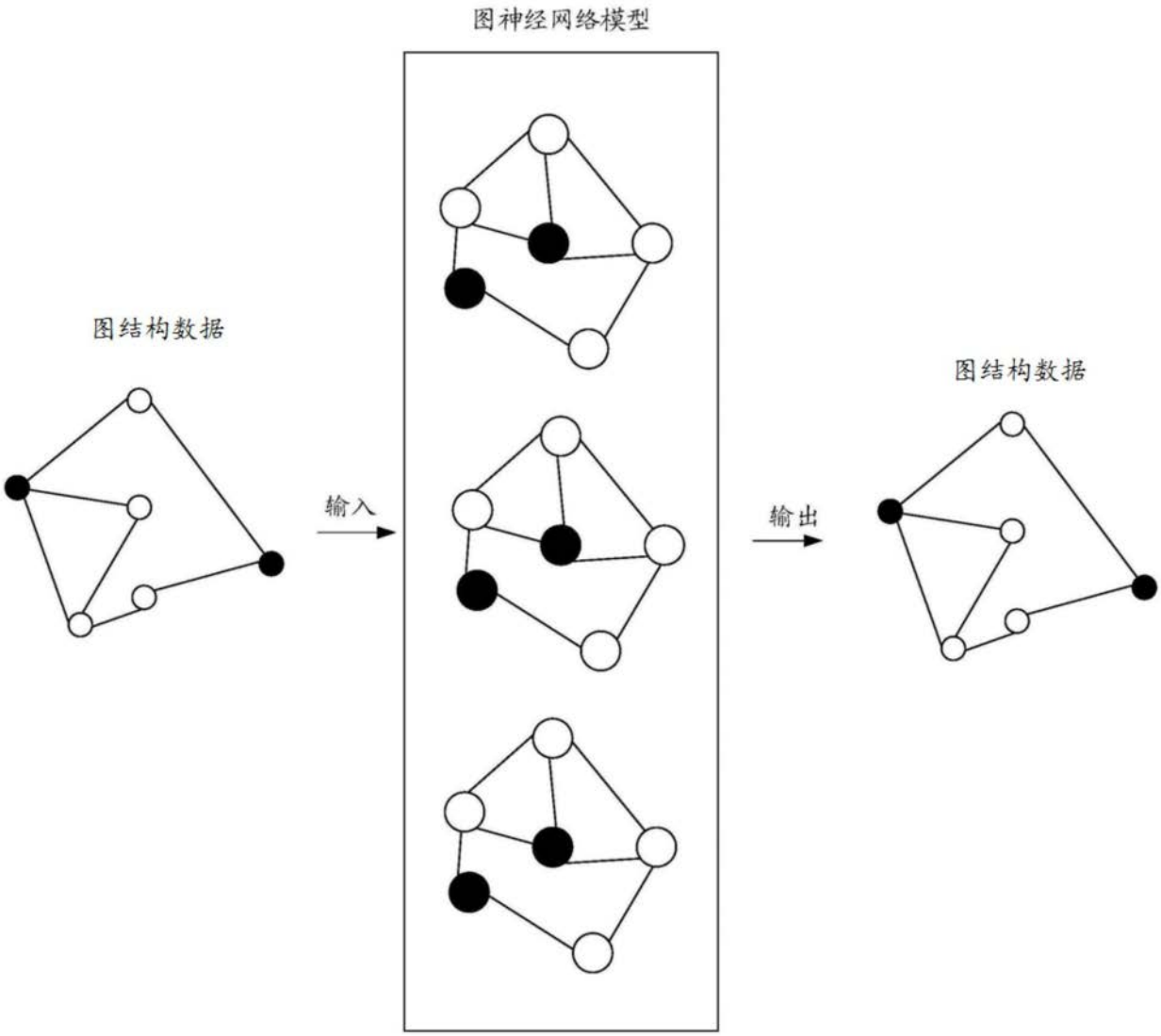

图神经网络模型
图结构数据
输入
输出
图结构数据

图1D

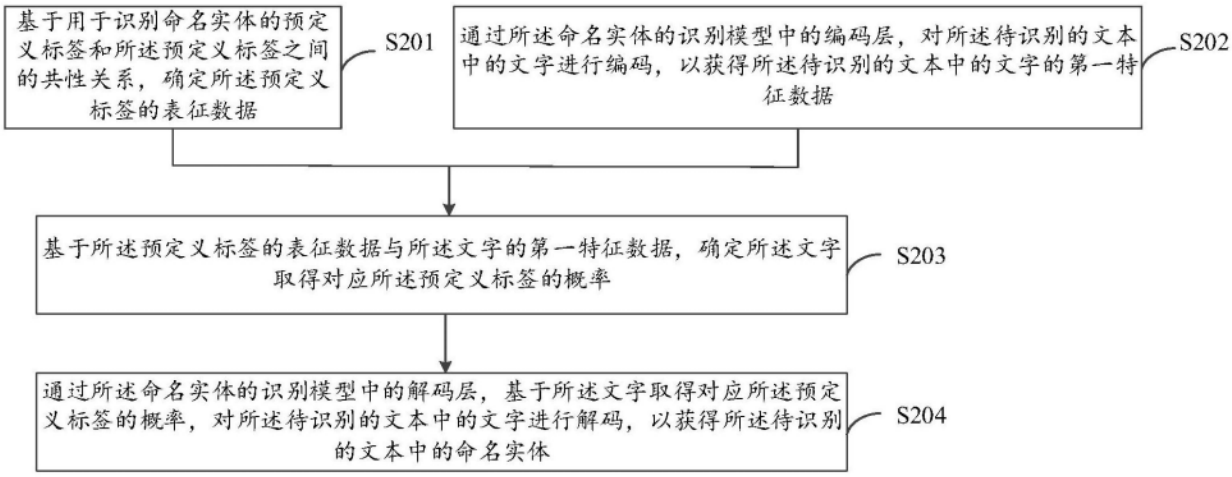

基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系，确定所述预定义标签的表征数据
S201
通过所述命名实体的识别模型中的编码层，对所述待识别的文本中的文字进行编码，以获得所述待识别的文本中的文字的第一特征数据
S202
基于所述预定义标签的表征数据与所述文字的第一特征数据，确定所述文字取得对应所述预定义标签的概率
S203
通过所述命名实体的识别模型中的解码层，基于所述文字取得对应所述预定义标签的概率，对所述待识别的文本中的文字进行解码，以获得所述待识别的文本中的命名实体
S204

图2A

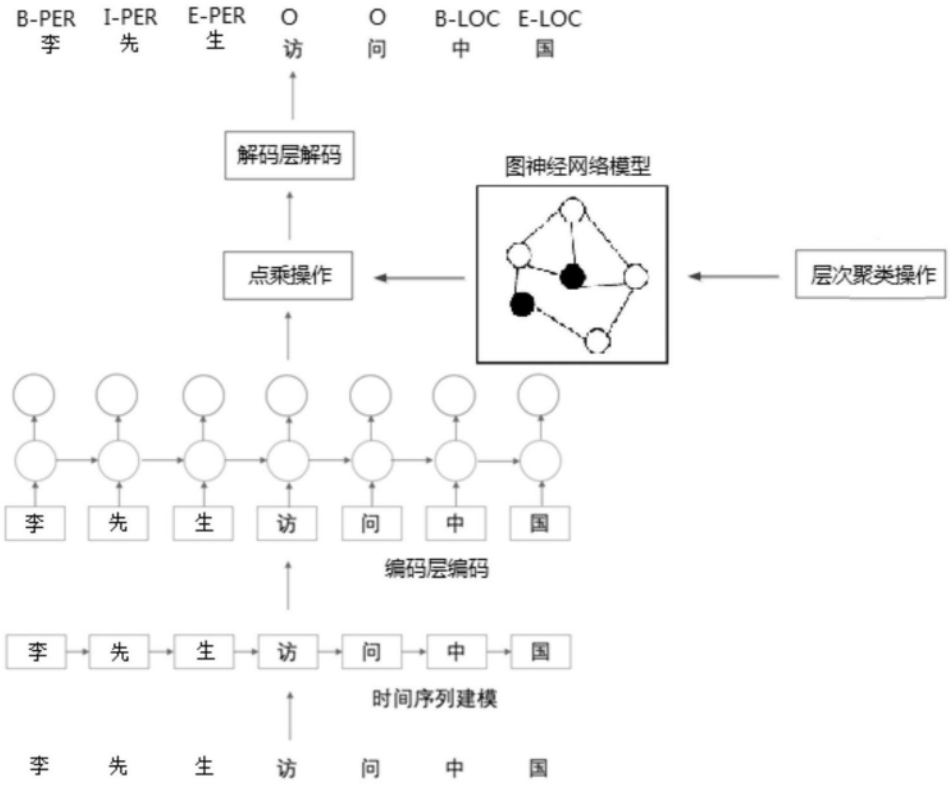

B-PER
李
I-PER
先
E-PER
生
O
访
O
问
B-LOC
中
E-LOC
国
解码层解码
图神经网络模型
点乘操作
层次聚类操作
李
先
生
访
问
中
国
编码层编码
李
先
生
访
问
中
国
时间序列建模
李
先
生
访
问
中
国

图2B

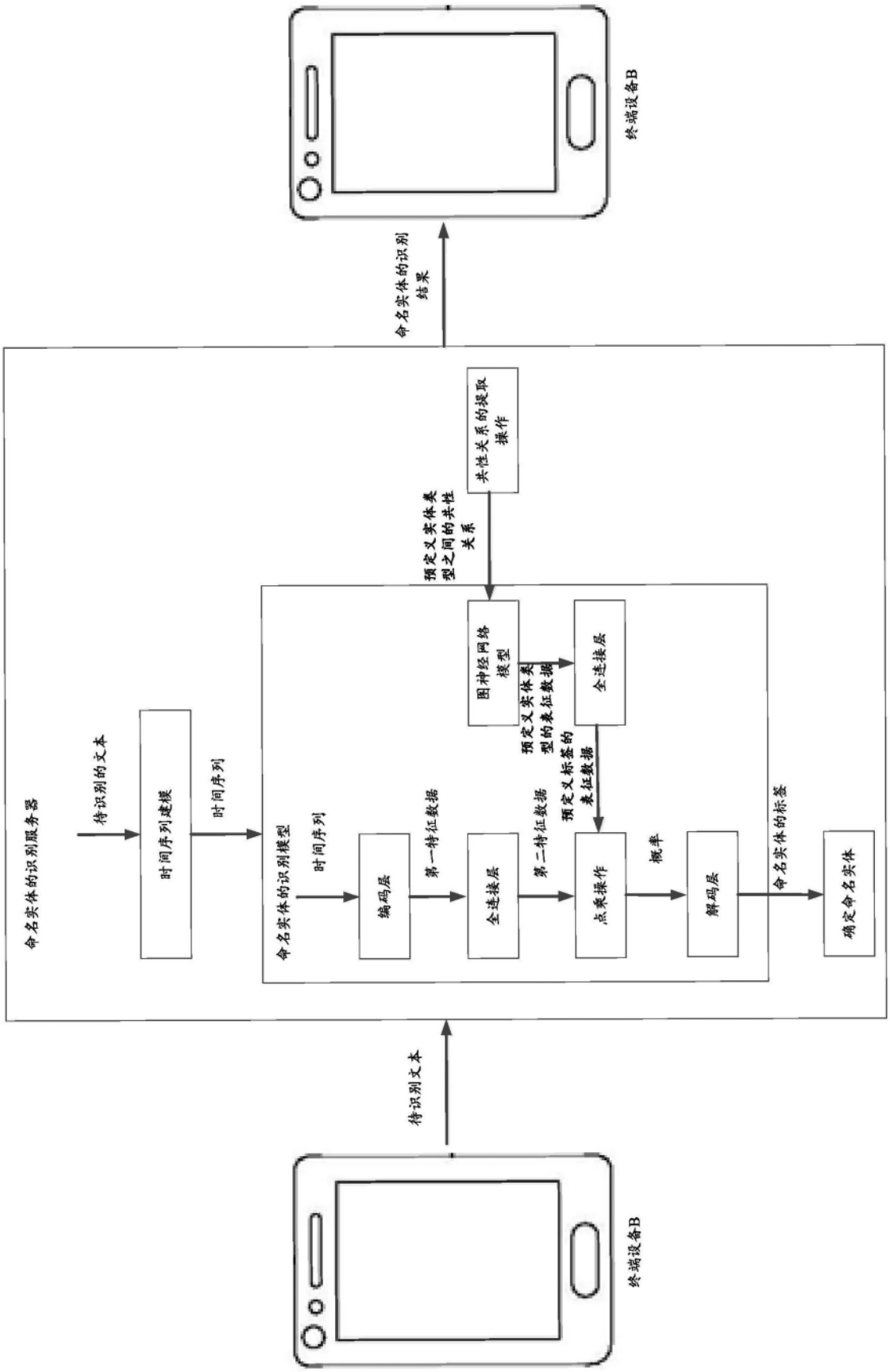

终端设备B
命名实体的识别结果
命名实体的识别服务器
待识别的文本
时间序列建模
时间序列
预定义实体类型之间的共性关系
共性关系的提取操作
命名实体的识别模型
时间序列
编码层
第一特征数据
全连接层
第二特征数据
图神经网络模型
预定义实体类型的表征数据
全连接层
预定义标签的表征数据
点乘操作
概率
解码层
命名实体的标签
确定命名实体
待识别文本
终端设备B

图2C

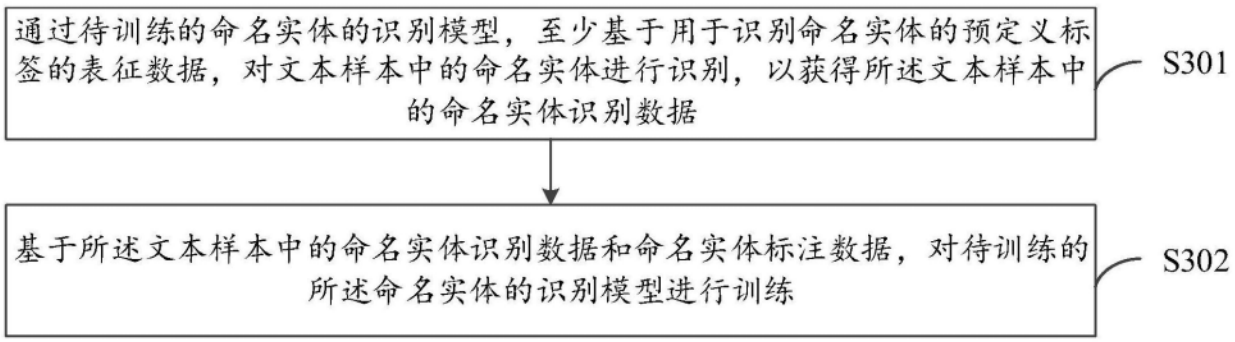

通过待训练的命名实体的识别模型，至少基于用于识别命名实体的预定义标签的表征数据，对文本样本中的命名实体进行识别，以获得所述文本样本中的命名实体识别数据
S301
基于所述文本样本中的命名实体识别数据和命名实体标注数据，对待训练的所述命名实体的识别模型进行训练
S302

图3

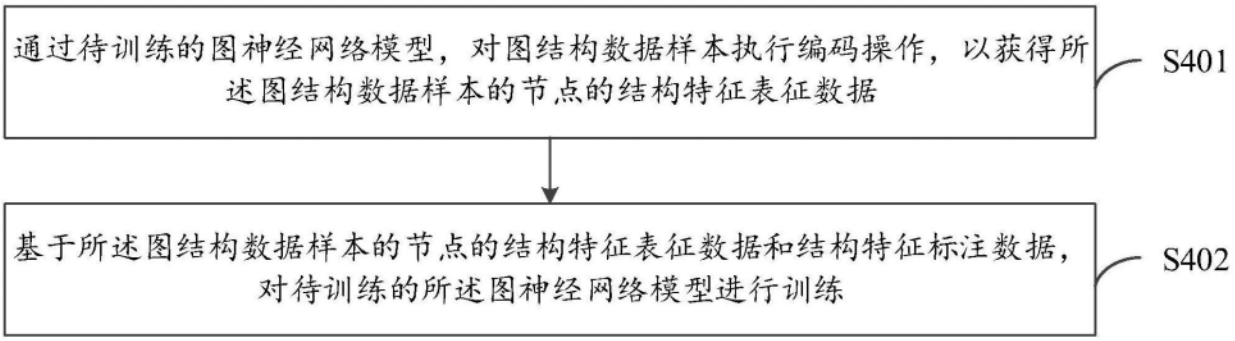

通过待训练的图神经网络模型，对图结构数据样本执行编码操作，以获得所述图结构数据样本的节点的结构特征表征数据
S401
基于所述图结构数据样本的节点的结构特征表征数据和结构特征标注数据，对待训练的所述图神经网络模型进行训练
S402

图4

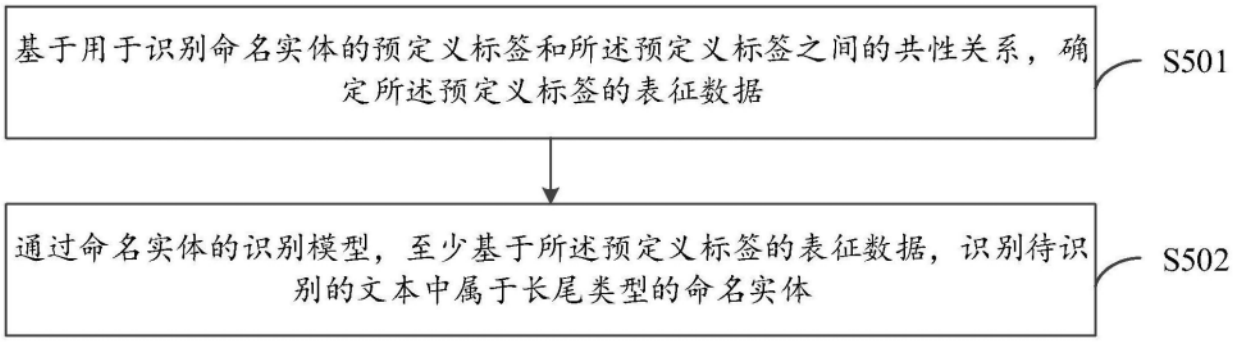

基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系，确定所述预定义标签的表征数据
S501
通过命名实体的识别模型，至少基于所述预定义标签的表征数据，识别待识别的文本中属于长尾类型的命名实体
S502

图5

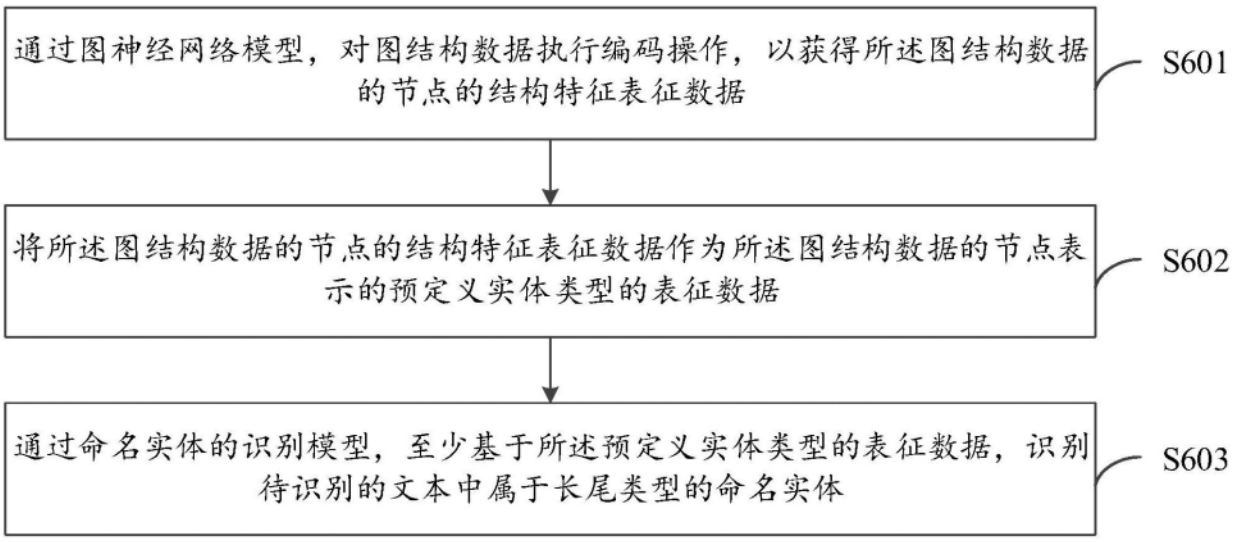

通过图神经网络模型，对图结构数据执行编码操作，以获得所述图结构数据的节点的结构特征表征数据
S601
将所述图结构数据的节点的结构特征表征数据作为所述图结构数据的节点表示的预定义实体类型的表征数据
S602
通过命名实体的识别模型，至少基于所述预定义实体类型的表征数据，识别待识别的文本中属于长尾类型的命名实体
S603

图6

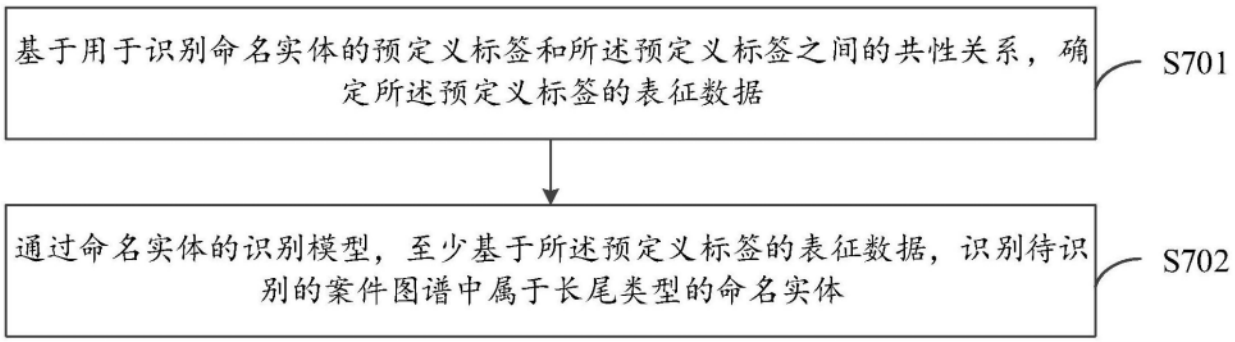

基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系，确定所述预定义标签的表征数据
S701
通过命名实体的识别模型，至少基于所述预定义标签的表征数据，识别待识别的案件图谱中属于长尾类型的命名实体
S702

图7

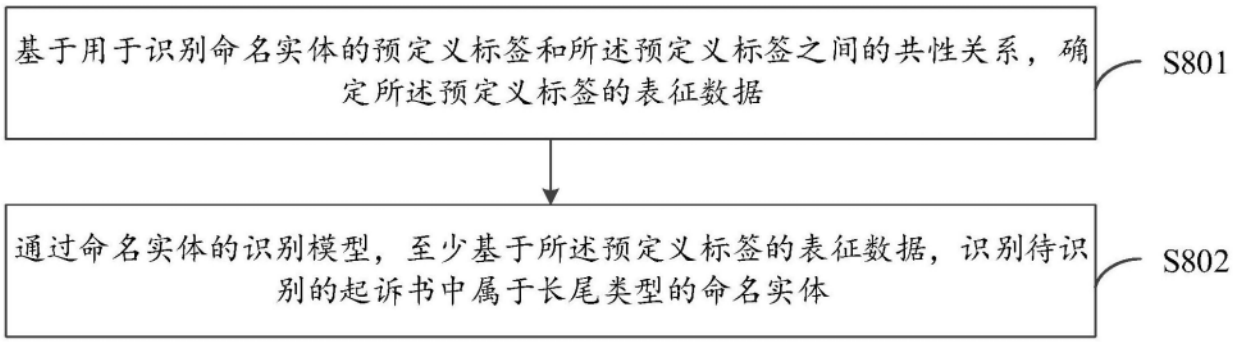

基于用于识别命名实体的预定义标签和所述预定义标签之间的共性关系，确定所述预定义标签的表征数据
S801
通过命名实体的识别模型，至少基于所述预定义标签的表征数据，识别待识别的起诉书中属于长尾类型的命名实体
S802

图8

图9

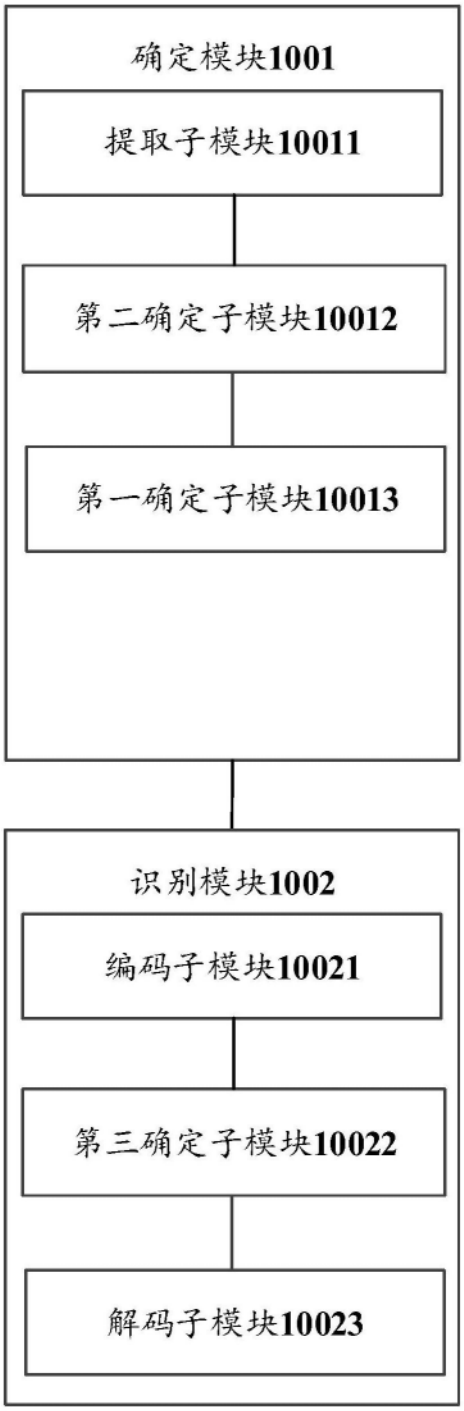

确定模块1001
提取子模块10011
第二确定子模块10012
第一确定子模块10013
识别模块1002
编码子模块10021
第三确定子模块10022
解码子模块10023

图10

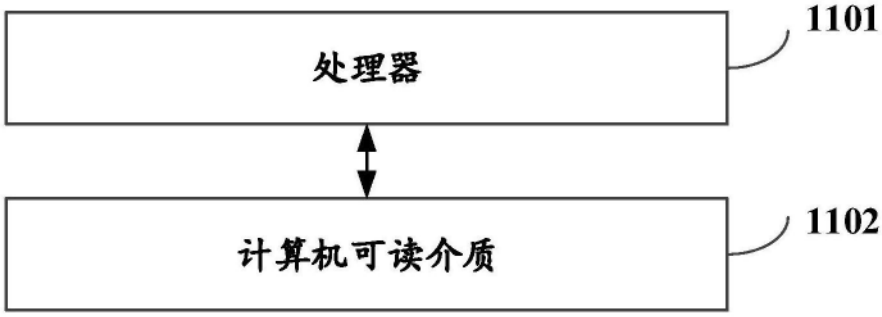

处理器
1101
计算机可读介质
1102

图11

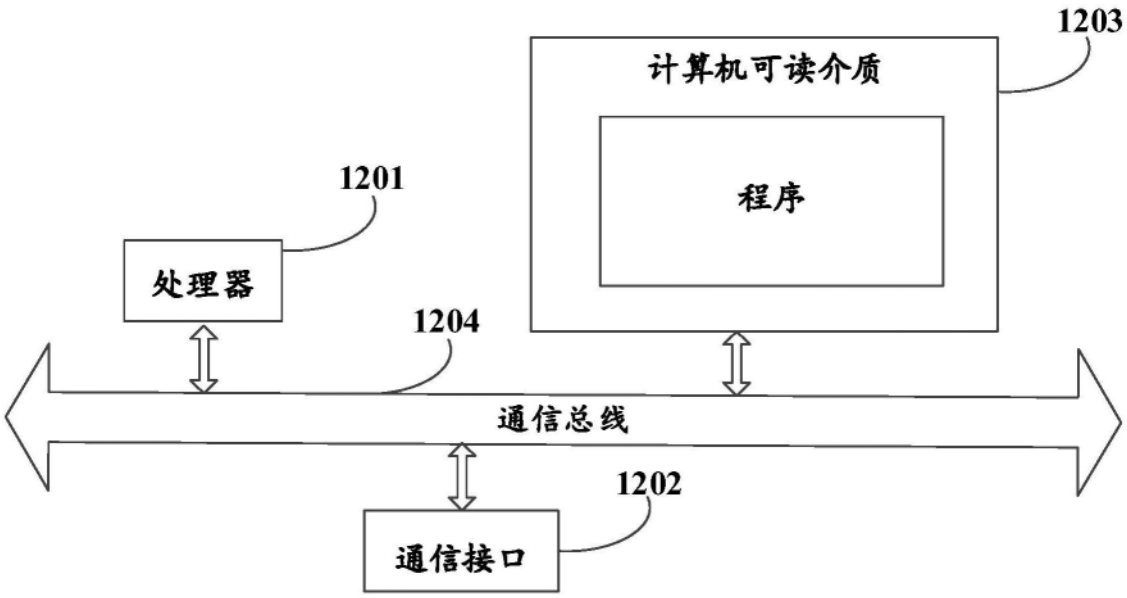

1203
计算机可读介质
程序
1201
处理器
1204
通信总线
1202
通信接口

图12