(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2022 年 7 月 7 日 (07.07.2022)

WIPO | PCT

(10) 国际公布号
WO 2022/142123 A1

(51) 国际专利分类号:
G06N 3/08 (2006.01)

(21) 国际申请号: PCT/CN2021/097545

(22) 国际申请日: 2021 年 5 月 31 日 (31.05.2021)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202011626618.0 2020年12月31日 (31.12.2020) CN

(71) 申请人: 平安科技（深圳）有限公司 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) [CN/CN]; 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

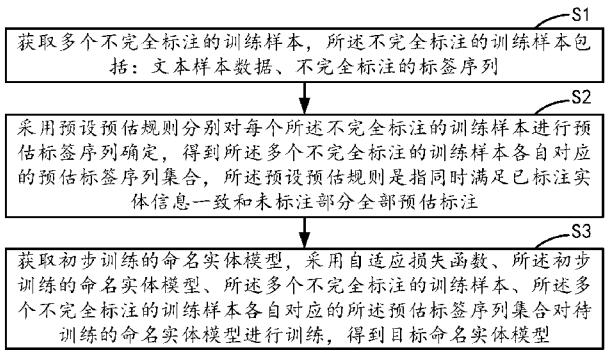
(72) 发明人: 阮鸿涛 (RUAN, Hongtao); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 郑立颖 (ZHENG, Liying); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 胡沛弦 (HU, Peixian); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。 徐亮 (XU, Liang); 中国广东省深圳市福田区福田街道福安社区益田路5033号平安金融中心23楼, Guangdong 518000 (CN)。

(74) 代理人: 深圳市明日今典知识产权代理事务所（普通合伙） (SHENZHEN MINGRIJINDIAN INTELLECTUAL PROPERTY AGENCY FIRM(GENERAL)); 中国广东省深圳市南山区南头街道安乐社区关口二路15号智恒产业园30栋4层406室, Guangdong 518000 (CN)。

(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

(54) **Title:** TRAINING METHOD AND APPARATUS FOR NAMED ENTITY MODEL, DEVICE, AND MEDIUM

(54) **发明名称:** 命名实体模型的训练方法、装置、设备及介质



S1 Acquire a plurality of incompletely labeled training samples, the incompletely labeled training samples comprising: text sample data and an incompletely labeled label sequence

S2 Determine an estimated label sequence of each incompletely labeled training sample by means of preset estimation rules, and obtain an estimated label sequence set respectively corresponding to the plurality of incompletely labeled training samples, the preset estimation rules referring to simultaneously meeting requirements that labeled entity information is consistent and there are estimation labels for the entire unlabeled entity portion

S3 Acquire a named entity model for preliminary training, and use an adaptive loss function, the named entity model for preliminary training, the plurality of incompletely labeled training samples and the estimated label sequence set respectively corresponding to the plurality of incompletely labeled training samples to train a named entity model to be trained so as to obtain a target named entity model

图 1

(57) **Abstract:** The present application relates to the technical field of artificial intelligence, and discloses a training method and apparatus for a named entity model, a device, and a medium. The method comprises: acquiring a plurality of incompletely labeled training samples, the incompletely labeled training samples comprising: text sample data and an incompletely labeled label sequence; determining an estimated label sequence of each incompletely labeled training sample by means of preset estimation rules, and obtaining an estimated label sequence set respectively corresponding to the plurality of incompletely labeled training samples; and acquiring a named entity model for preliminary training, and using an adaptive loss function, the named entity model for preliminary training, the plurality of incompletely labeled training samples and the estimated label sequence set respectively corresponding to the plurality of incompletely labeled training samples to train a named entity model to be trained so as to obtain a target named entity model. Hence, dependence on labeling quality is reduced, and an adaptive loss function is used to avoid redirecting attention towards a large number of label sequences during training.

WO 2022/142123 A1

[见续页]

LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要：本申请涉及人工智能技术领域，揭示了一种命名实体模型的训练方法、装置、设备及介质，其中方法包括：获取多个不完全标注的训练样本，所述不完全标注的训练样本包括：文本样本数据、不完全标注的标签序列；采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合；获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型。从而减少了对标注质量的依赖，采用自适应损失函数避免在训练时将注意力分散到大量的标签序列上。

命名实体模型的训练方法、装置、设备及介质

本申请要求于2020年12月31日提交中国专利局、申请号为2020116266180，发明名称为“命名实体模型的训练方法、装置、设备及介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本申请涉及到人工智能技术领域，特别是涉及到一种命名实体模型的训练方法、装置、设备及介质。

背景技术

目前命名实体识别模型训练都依赖于大量的完全标注的数据，然而现实情况是大量的、高质量的完全标注的数据的获取极其昂贵且困难。为了解决这个问题，很多情况下由标注人员输出的命名实体数据都是标注不完全的，即仅标注了一部分实体，然后利用不完全标注的数据训练命名实体识别模型。发明人意识到不完全标注的数据的未被标注为实体的内容可以为任何标签，而文本中命名实体一般是稀疏的，导致可能的标签序列的数量随着未标注文本内容的长度增加而呈现指数增加，因为现有利用不完全标注的数据训练命名实体识别模型时将注意力分散到大量的标签序列上，使得模型在搜索真实标签序列的时候遇到较大的困难。

技术问题

旨在解决现有技术利用不完全标注的数据训练命名实体识别模型时将注意力分散到大量的标签序列上，使得模型在搜索真实标签序列的时候遇到较大的困难的技术问题。

技术解决方案

本申请的主要目的为提供一种命名实体模型的训练方法、装置、设备及介质，旨在解决现有技术利用不完全标注的数据训练命名实体识别模型时将注意力分散到大量的标签序列上，使得模型在搜索真实标签序列的时候遇到较大的困难的技术问题。

为了实现上述发明目的，本申请提出一种命名实体模型的训练方法，所述方法包括：

获取多个不完全标注的训练样本，所述不完全标注的训练样本包括：文本样本数据、不完全标注的标签序列；

采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合，所述预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注；

获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型。

本申请还提出了一种命名实体模型的训练装置，所述装置包括：

训练样本获取模块，用于获取多个不完全标注的训练样本，所述不完全标注的训练样本包括：文本样本数据、不完全标注的标签序列；

预估标签序列集合确定模块，用于采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各

自对应的预估标签序列集合，所述预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注；

模型训练模块，用于获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型。

本申请还提出了一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现如下方法步骤：

获取多个不完全标注的训练样本，所述不完全标注的训练样本包括：文本样本数据、不完全标注的标签序列；

采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合，所述预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注；

获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型。本申请还提出了一种计算机可读存储介质，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如下方法步骤：

获取多个不完全标注的训练样本，所述不完全标注的训练样本包括：文本样本数据、不完全标注的标签序列；

采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合，所述预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注；

获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型。

有益效果

本申请的命名实体模型的训练方法、装置、设备及介质，首先采用预设预估规则分别对每个不完全标注的训练样本进行预估标签序列确定，得到多个不完全标注的训练样本各自对应的预估标签序列集合，预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注，然后采用自适应损失函数、初步训练的命名实体模型、多个不完全标注的训练样本、多个不完全标注的训练样本各自对应的预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型，通过采用不完全标注的训练样本，减少了对标注质量的依赖，采用自适应损失函数避免在训练时将注意力分散到大量的标签序列上，从而使采用多个不完全标注的训练样本训练模型也可以取得较好的效果。

附图说明

图1 为本申请一实施例的命名实体模型的训练方法的流程示意图；

图2 为本申请一实施例的命名实体模型的训练装置的结构示意框图；

图3 为本申请一实施例的计算机设备的结构示意框图。

本申请目的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

本发明的实施方式

为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

为了解决现有技术利用不完全标注的数据训练命名实体识别模型时将注意力分散到大量的标签序列上，使得模型在搜索真实标签序列的时候遇到较大的困难的技术问题，本申请提出了一种命名实体模型的训练方法，所述方法应用于人工智能技术领域。所述命名实体模型的训练方法通过采用不完全标注的训练样本，减少了对标注质量的依赖，采用自适应损失函数避免在训练时将注意力分散到大量的标签序列上，从而使采用多个不完全标注的训练样本训练模型也可以取得较好的效果。

参照图 1，本申请实施例中提供一种命名实体模型的训练方法，所述方法包括:

S1: 获取多个不完全标注的训练样本，所述不完全标注的训练样本包括: 文本样本数据、不完全标注的标签序列;

S2: 采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合，所述预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注;

S3: 获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型。

本实施例首先通过采用预设预估规则分别对每个不完全标注的训练样本进行预估标签序列确定，得到多个不完全标注的训练样本各自对应的预估标签序列集合，预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注，然后采用自适应损失函数、初步训练的命名实体模型、多个不完全标注的训练样本、多个不完全标注的训练样本各自对应的预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型，通过采用不完全标注的训练样本，减少了对标注质量的依赖，采用自适应损失函数避免在训练时将注意力分散到大量的标签序列上，从而使采用多个不完全标注的训练样本训练模型也可以取得较好的效果。

对于S1，可以从数据库中获取多个不完全标注的训练样本，也可以是用户输入的多个不完全标注的训练样本，还可以是第三方应用系统发送的多个不完全标注的训练样本。

文本样本数据中包括多个文字。

在每个不完全标注的训练样本中，不完全标注的标签序列是对文本样本数据的实体进行不完全标注的结果。

比如，不完全标注的训练样本的文本样本数据$x=(x_1,x_2,\cdots,x_n)$，其中每个$x_i(i=1,2,\cdots,n)$代表文本样本数据中的一个字，x对应的不完全标注的标签序列为 $y_u=(-,y_2,-,\ldots y_i,\ldots,-)$，其中$y_i$代表标注人员标注的字$x_i$对应的标签，“-”代表未被标注，即“-”对应文本样本数据中的字可能为实体或非实体，在此举例不做具体限定。

对于S2，分别对每个所述不完全标注的训练样本进行所有可能的预估标签序列确定，预估标签序列中包含了预估标签序列对应的所述不完全标注的训练样本的不完全标注的标签序列的已标注实体信息，并且预估标签序列是完全标注（包括：可能标注和不完全标注的标签序列的已标注实体信息）的。

比如，不完全标注的训练样本x的不完全标注的标签序列为$y_u=(-,y_2,-,\dots y_i,\dots,-)$，一条与$y_u$中已标注实体信息一致未标注部分全部预估标注的预估标签序列为$y_c=(y_{c1},y_2,y_{c3},\dots y_i,\dots,y_{cn})$，其中$y_{c2}$表示未标注的位置2上一种可能的标签，将所有符合$y_u$的已标注实体信息未标注部分全部预估标注的预估标签序列y_c组成的集合为$C(y_u)$，$C(y_u)$即为所述不完全标注的训练样本x对应的预估标签序列集合，在此举例不做具体限定。

对于S3，可以从数据库中获取初步训练的命名实体模型，也可以是用户输入的初步训练的命名实体模型，还可以是第三方应用系统发送的初步训练的命名实体模型；采用所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，训练时采用自适应损失函数作为损失函数，将训练结束的待训练的命名实体模型作为目标命名实体模型。

初步训练的命名实体模型，是指采用少量的完全标注的训练样本对预训练模型和条件随机场模型训练得到的模型。

待训练的命名实体模型包括：预训练模型、条件随机场模型。预训练模型是基于Bert网络训练得到的模型。

自适应损失函数，是根据训练的进程可以调整的损失函数，从而避免在训练时将注意力分散到大量的标签序列上。

在一个实施例中，上述所述采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合的步骤，包括的步骤，包括：

S21：从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本，作为目标不完全标注的训练样本；

S22：从所述目标不完全标注的训练样本的所述不完全标注的标签序列中提取出已标注实体信息，得到所述目标不完全标注的训练样本对应的已标注实体信息；

S23：采用所述目标不完全标注的训练样本对应的所述已标注实体信息从所述目标不完全标注的训练样本的所述文本样本数据中找出未标注的文字，得到所述目标不完全标注的训练样本对应的未标注文本数据；

S24：分别对所述目标不完全标注的训练样本对应的所述未标注文本数据中每个文字进行所有可能的标签预估，得到所述目标不完全标注的训练样本对应的所述未标注文本数据的各个文字各自对应的预估标签集合；

S25：分别将所述目标不完全标注的训练样本对应的所述未标注文本数据的各个文字各自对应的所述预估标签集合和所述目标不完全标注的训练样本对应的所述已标注实体信息进行所有可能的标签序列组合，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合；

S26：重复执行所述从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本作为目标不完全标注的训练样本的步骤，直至确定所述多个不完全标注的训练样本各自对应的所述预估标签序列集合。

本实施例实现了确定预估标签序列，为后续进行模型训练提供了数据基础。

对于S21，从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本，将获取的所述不完全标注的训练样本作为目标不完全标注的训练样本。

对于S22，从所述目标不完全标注的训练样本对应的所述不完全标注的标签序列中提取出所有的已标注实体的信息，将提取得到的已标注实体的信息作为所述目标不完全标注的训练样本对应的已标注实体信息。

已标注实体信息包括：被标注为实体的文字在文本样本数据中的位置数据。

对于S23，采用所述目标不完全标注的训练样本对应的所述已标注实体信息从所述目标不完全标注的训练样本的所述文本样本数据中找出未标注的文字，将找出的未标注的文字作为所述目标不完全标注的训练样本对应的未标注文本数据。

未标注文本数据包括：在文本样本数据中的位置数据、未标注的文字，在未标注文本数据中每个在文本样本数据中的位置数据对应一个未标注的文字。

对于S24，分别对所述目标不完全标注的训练样本对应的所述未标注文本数据中每个文字（也就是未标注的文字）进行所有可能的标签预估，将一个文字（也就是未标注的文字）对应的所有可能的标签预估结果作为一个预估标签集合。

对于S25，从所述目标不完全标注的训练样本对应的所述未标注文本数据的各个文字各自对应的所述预估标签集合中的每个所述预估标签集合中挑出一个预估标签，将挑选出的预估标签作为可能的待组合预估标签集合；将所述目标不完全标注的训练样本对应的所述已标注实体信息和所述多个可能的待组合预估标签集合中每个可能的待组合预估标签集合按位置数据的先后顺序进行组合，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合。

对于S26，重复执行步骤S21至步骤S26，直至确定所述多个不完全标注的训练样本各自对应的所述预估标签序列集合。

在一个实施例中，上述获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型的步骤，包括：

S31：从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本，作为目标不完全标注的训练样本；

S32：采用所述初步训练的命名实体模型分别对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个预估标签序列进行概率分布计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析概率分布数据；

S33：采用所述待训练的命名实体模型分别对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个所述预估标签序列进行条件概率计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析条件概率数据；

S34：采用所述待训练的命名实体模型，对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的所有所述预估标签序列进行最有可能的标签序列解析，得到所述目标不完全标注的训练样本对应的最有可能标签序列集合；

S35：将所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析概率分布数据、所述目标不完全标注

的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析条件概率数据和所述目标不完全标注的训练样本对应的所述最有可能标签序列集合输入所述自适应损失函数进行计算，得到所述待训练的命名实体模型的损失值，根据所述损失值更新所述待训练的命名实体模型的参数，更新后的所述待训练的命名实体模型被用于下一次计算所述待分析条件概率数据、所述最有可能标签序列集合；

S36：重复执行上述方法步骤直至所述损失值达到第一收敛条件或迭代次数达到第二收敛条件，将所述损失值达到所述第一收敛条件或迭代次数达到所述第二收敛条件的所述待训练的命名实体模型，确定为所述目标命名实体模型。

本实施例实现了采用自适应损失函数避免在训练时将注意力分散到大量的标签序列上，从而使采用多个不完全标注的训练样本训练模型也可以取得较好的效果。

对于S31，从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本，将获取的所述不完全标注的训练样本作为目标不完全标注的训练样本。

对于S32，分别将所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个预估标签序列输入所述初步训练的命名实体模型进行概率分布预测，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析概率分布数据。也就是说，待分析概率分布数据的数量和目标不完全标注的训练样本对应的所述预估标签序列集合的预估标签序列的数量相同。

对于S33，分别将所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个所述预估标签序列输入所述待训练的命名实体模型，获取所述待训练的命名实体模型的条件随机场模型输出的条件概率，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析条件概率数据。也就是说，待分析条件概率数据的数量与目标不完全标注的训练样本对应的所述预估标签序列集合的预估标签序列的数量相同。

对于S34，采用当前的所述待训练的命名实体模型的条件随机场模型的状态转移矩阵和当前的所述待训练的命名实体模型的预训练模型的输出进行最有可能的标签序列解析，将解析得到的所有所述最有可能标签序列作为所述目标不完全标注的训练样本对应的最有可能标签序列集合。

对于S35，根据所述损失值更新所述待训练的命名实体模型的参数的方法可以从现有技术中选择，在此不做赘述。

对于S36，重复执行步骤S31至S36，直至所述损失值达到第一收敛条件或迭代次数达到第二收敛条件。

所述第一收敛条件是指相邻两次计算的损失的大小满足lipschitz条件（利普希茨连续条件）。

所述迭代次数达到第二收敛条件是指所述待训练的命名实体模型被用于计算所述待分析条件概率数据、所述最有可能标签序列集合的次数，也就是说，计算一次所述待分析条件概率数据、所述最有可能标签序列集合，迭代次数增加1。

在一个实施例中，上述采用所述初步训练的命名实体模型分别对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个预估标签序列进行概率分布计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析概率分布数据的步骤，包括：

S321：基于前向-后向算法和所述初步训练的命名实体模型，分别对所述目标不完全标注的训练样本的所述文本样本数据的每个字进行各个标签的边缘概率计算，得到所述目标不完全标注的训练样本的所述文本样本数据的各个字各自对应的各个标签的边缘概率数据；

S322：分别根据所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个预估标签序列、所述目标不完全标注的训练样本的所述文本样本数据的各个字各自对应的各个标签的所述边缘概率数据进行各个字各自对应的所述边缘概率数据的相乘计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析概率分布数据。

本实施例实现了计算所有所述预估标签序列各自对应的待分析概率分布数据，为后续进行模型训练提供了数据基础。

对于S321，将所述目标不完全标注的训练样本的所述文本样本数据输入所述初步训练的命名实体模型，并且采用前向-后向算法计算出所述目标不完全标注的训练样本的所述文本样本数据的每个字对应各个标签的边缘概率（边缘分布）。也就是说，每个字对应的边缘概率和标签总数量相同。比如，标签包括：实体、非实体，则标签总数量为2，在此举例不做具体限定。

对于S322，从所述目标不完全标注的训练样本对应的所述预估标签序列集合中提取出一个预估标签序列作为目标预估标签序列；将目标预估标签序列的每个标签依次在所述目标不完全标注的训练样本的所述文本样本数据的各个字各自对应的各个标签的所述边缘概率数据中进行查找，将查找到的所述边缘概率数据相乘，将相乘结果作为所述目标预估标签序列对应的所述待分析概率分布数据；重复执行所述从所述目标不完全标注的训练样本对应的所述预估标签序列集合中提取出一个预估标签序列作为目标预估标签序列的步骤，直至确定所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析概率分布数据。

比如，所述目标不完全标注的训练样本的目标预估标签序列有10个预估标签，目标预估标签序列第二个位置的预估标签是实体，将所述目标不完全标注的训练样本的所述文本样本数据的第二个字（与目标预估标签序列第二个位置对应）对应的标签为实体（与目标预估标签序列第二个位置的预估标签对应）的所述边缘概率数据作为目标预估标签序列第二个位置的预估标签对应的边缘概率数据，然后将目标预估标签序列10个位置（与10个预估标签对应）的预估标签对应的边缘概率数据进行相乘（也就是10个边缘概率数据相乘），将相乘结果作为所述目标不完全标注的训练样本的目标预估标签序列对应的所述待分析概率分布数据，在此举例不做具体限定。

在一个实施例中，上述采用所述待训练的命名实体模型，对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的所有所述预估标签序列进行最有可能的标签序列解析，得到所述目标不完全标注的训练样本对应的最有可能标签序列集合的步骤，包括：

S341：将所述目标不完全标注的训练样本的所述文本样本数据输入所述待训练的命名实体模型进行计算，获取所述目标不完全标注的训练样本的预训练模型输出的所述目标不完全标注的训练样本对应的概率预测结果；

S342：采用k-best Viterbi Decoding算法根据所述待训练的命名实体模型的条件随机场模型的状态转移矩阵、所述目标不完全标注的训练样本的所述概率预测

结果进行最有可能的标签序列的解码计算，得到所述目标不完全标注的训练样本对应的所述最有可能标签序列集合。

本实施例实现了采用k-best Viterbi Decoding算法进行最有可能的标签序列的解码计算，从而为后续进行模型训练提供了数据基础。

对于S341，将所述目标不完全标注的训练样本的所述文本样本数据输入所述待训练的命名实体模型进行计算，将所述目标不完全标注的训练样本的预训练模型输出的概率作为所述目标不完全标注的训练样本对应的概率预测结果。

对于S342，从当前的所述待训练的命名实体模型中提取出条件随机场模型的状态转移矩阵；采用k-best Viterbi Decoding算法将提取出的状态转移矩阵和所述目标不完全标注的训练样本的所述概率预测结果进行最有可能的标签序列的解码计算，得到所述目标不完全标注的训练样本对应的所述最有可能标签序列集合。

采用k-best Viterbi Decoding算法进行最有可能的标签序列的解码计算的方法可以从现有技术中选择，在此不做赘述。

在一个实施例中，上述自适应损失函数的计算公式L(w,x)为：

$$L(w,x)=(1-\lambda)L_1(w,x)+\lambda L_2(w,x)$$

$$L_1(w,x)=-\log\sum_{y'\in C(y_u)}q(y'|x)p_w(y'|x)$$

$$L_2(w,x)=-\log\sum_{y'\in K_w(x)}p_w(y'|x)$$

其中，$q(y'|x)$是所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析概率分布数据，$p_w(y'|x)$是所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析条件概率数据，$C(y_u)$是所述目标不完全标注的训练样本对应的所述预估标签序列集合，$K_w(x)$所述目标不完全标注的训练样本对应的所述最有可能标签序列集合，log()是对数函数，λ是自适应参数，λ从0逐步增加到1。

本实施例实现了采用自适应损失函数避免在训练时将注意力分散到大量的标签序列上，从而使采用多个不完全标注的训练样本训练模型也可以取得较好的效果。

在训练初期，模型能够利用所有所述预估标签序列各自对应的所述待分析概率分布数据和标签信息进行训练，此时使自适应损失函数中的$L_1(w,x)$权重偏大，$L_2(w,x)$的权重偏小；在训练过程中，通过λ逐步调整损失函数，增加最有可能的所述预估标签序列的权重，此时使自适应损失函数中的$L_1(w,x)$权重偏小，$L_2(w,x)$的权重偏大，使模型更易把握住真实的标签序列，从而使采用多个不完全标注的训练样本训练模型也可以取得较好的效果。

在一个实施例中，上述自适应参数λ的计算公式为：

$$\lambda=\exp\left[\gamma\times\left(\frac{b}{B}-1\right)\right]$$

其中，exp[]是以自然常数e为底的指数函数，b是所述目标不完全标注的训练样本训练所述待训练的命名实体模型时对应的训练步数，B是预设的总训练步数，γ为控制λ增长速度的常数。

本实施例通过训练步数、总训练步数、控制λ增长速度的常数来控制自适应参数的值从0逐步增加到1，实现了根据训练进度从小到大调整自适应参数。

参照图 2，本申请提出了一种命名实体模型的训练装置，所述装置包括:

训练样本获取模块100，用于获取多个不完全标注的训练样本，所述不完全标注的训练样本包括：文本样本数据、不完全标注的标签序列;

预估标签序列集合确定模块200，用于采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合，所述预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注;

模型训练模块300，用于获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型。

本实施例首先通过采用预设预估规则分别对每个不完全标注的训练样本进行预估标签序列确定，得到多个不完全标注的训练样本各自对应的预估标签序列集合，预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注，然后采用自适应损失函数、初步训练的命名实体模型、多个不完全标注的训练样本、多个不完全标注的训练样本各自对应的预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型，通过采用不完全标注的训练样本，减少了对标注质量的依赖，采用自适应损失函数避免在训练时将注意力分散到大量的标签序列上，从而使采用多个不完全标注的训练样本训练模型也可以取得较好的效果。

参照图3，本申请实施例中还提供一种计算机设备，该计算机设备可以是服务器，其内部结构可以如图3所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该计算机设计的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的数据库用于储存命名实体模型的训练方法等数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种命名实体模型的训练方法。所述命名实体模型的训练方法，包括：获取多个不完全标注的训练样本，所述不完全标注的训练样本包括：文本样本数据、不完全标注的标签序列；采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合，所述预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注；获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型。

本实施例首先通过采用预设预估规则分别对每个不完全标注的训练样本进行预估标签序列确定，得到多个不完全标注的训练样本各自对应的预估标签序列集合，预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注，然后采用自适应损失函数、初步训练的命名实体模型、多个不完全标注的训

练样本、多个不完全标注的训练样本各自对应的预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型，通过采用不完全标注的训练样本，减少了对标注质量的依赖，采用自适应损失函数避免在训练时将注意力分散到大量的标签序列上，从而使采用多个不完全标注的训练样本训练模型也可以取得较好的效果。

本申请一实施例还提供一种计算机可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现一种命名实体模型的训练方法，包括步骤：获取多个不完全标注的训练样本，所述不完全标注的训练样本包括：文本样本数据、不完全标注的标签序列；采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合，所述预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注；获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型。

上述执行的命名实体模型的训练方法，首先通过采用预设预估规则分别对每个不完全标注的训练样本进行预估标签序列确定，得到多个不完全标注的训练样本各自对应的预估标签序列集合，预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注，然后采用自适应损失函数、初步训练的命名实体模型、多个不完全标注的训练样本、多个不完全标注的训练样本各自对应的预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型，通过采用不完全标注的训练样本，减少了对标注质量的依赖，采用自适应损失函数避免在训练时将注意力分散到大量的标签序列上，从而使采用多个不完全标注的训练样本训练模型也可以取得较好的效果。

所述计算机可读存储介质可以是非易失性，也可以是易失性。

本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一非易失性计算机可读取存储介质中，该计算机程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的和实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可以包括只读存储器（ROM）、可编程ROM（PROM）、电可编程ROM（EPROM）、电可擦除可编程ROM（EEPROM）或闪存。易失性存储器可包括随机存取存储器（RAM）或者外部高速缓冲存储器。作为说明而非局限，RAM以多种形式可得，诸如静态RAM（SRAM）、动态RAM（DRAM）、同步DRAM（SDRAM）、双速据率SDRAM（SSRSDRAM）、增强型SDRAM（ESDRAM）、同步链路（Synchlink）DRAM（SLDRAM）、存储器总线（Rambus）直接RAM（RDRAM）、直接存储器总线动态RAM（DRDRAM）、以及存储器总线动态RAM（RDRAM）等。

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、装置、物品或者方法不仅包括那些要素，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、装置、物品或者方法所固有的要素。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、装置、物品或者方法中还存在另外的相同要素。

以上所述仅为本申请的优选实施例，并非因此限制本申请的专利范围，凡是利用本申请说明书及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，均同理包括在本申请的专利保护范围内。

权利要求书

1.一种命名实体模型的训练方法，其中，所述方法包括:

获取多个不完全标注的训练样本，所述不完全标注的训练样本包括：文本样本数据、不完全标注的标签序列;

采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合，所述预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注;

获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型。

2.根据权利要求1所述的命名实体模型的训练方法，其中，所述采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合的步骤，包括:

从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本，作为目标不完全标注的训练样本;

从所述目标不完全标注的训练样本的所述不完全标注的标签序列中提取出已标注实体信息，得到所述目标不完全标注的训练样本对应的已标注实体信息;

采用所述目标不完全标注的训练样本对应的所述已标注实体信息从所述目标不完全标注的训练样本的所述文本样本数据中找出未标注的文字，得到所述目标不完全标注的训练样本对应的未标注文本数据;

分别对所述目标不完全标注的训练样本对应的所述未标注文本数据中每个文字进行所有可能的标签预估，得到所述目标不完全标注的训练样本对应的所述未标注文本数据的各个文字各自对应的预估标签集合;

分别将所述目标不完全标注的训练样本对应的所述未标注文本数据的各个文字各自对应的所述预估标签集合和所述目标不完全标注的训练样本对应的所述已标注实体信息进行所有可能的标签序列组合，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合;

重复执行所述从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本作为目标不完全标注的训练样本的步骤，直至确定所述多个不完全标注的训练样本各自对应的所述预估标签序列集合。

3.根据权利要求1所述的命名实体模型的训练方法，其中，所述获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型的步骤，包括:

从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本，作为目标不完全标注的训练样本;

采用所述初步训练的命名实体模型分别对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个预估标签序列进行概率分布计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析概率分布数据;

采用所述待训练的命名实体模型分别对所述目标不完全标注的训练样本对

应的所述预估标签序列集合中的每个所述预估标签序列进行条件概率计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析条件概率数据；

采用所述待训练的命名实体模型，对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的所有所述预估标签序列进行最有可能的标签序列解析，得到所述目标不完全标注的训练样本对应的最有可能标签序列集合；

将所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析概率分布数据、所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析条件概率数据和所述目标不完全标注的训练样本对应的所述最有可能标签序列集合输入所述自适应损失函数进行计算，得到所述待训练的命名实体模型的损失值，根据所述损失值更新所述待训练的命名实体模型的参数，更新后的所述待训练的命名实体模型被用于下一次计算所述待分析条件概率数据、所述最有可能标签序列集合；

重复执行上述方法步骤直至所述损失值达到第一收敛条件或迭代次数达到第二收敛条件，将所述损失值达到所述第一收敛条件或迭代次数达到所述第二收敛条件的所述待训练的命名实体模型，确定为所述目标命名实体模型。

4.根据权利要求3所述的命名实体模型的训练方法，其中，所述采用所述初步训练的命名实体模型分别对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个预估标签序列进行概率分布计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析概率分布数据的步骤，包括：

基于前向-后向算法和所述初步训练的命名实体模型，分别对所述目标不完全标注的训练样本的所述文本样本数据的每个字进行各个标签的边缘概率计算，得到所述目标不完全标注的训练样本的所述文本样本数据的各个字各自对应的各个标签的边缘概率数据；

分别根据所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个预估标签序列、所述目标不完全标注的训练样本的所述文本样本数据的各个字各自对应的各个标签的所述边缘概率数据进行各个字各自对应的所述边缘概率数据的相乘计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析概率分布数据。

5.根据权利要求3所述的命名实体模型的训练方法，其中，所述采用所述待训练的命名实体模型，对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的所有所述预估标签序列进行最有可能的标签序列解析，得到所述目标不完全标注的训练样本对应的最有可能标签序列集合的步骤，包括：

将所述目标不完全标注的训练样本的所述文本样本数据输入所述待训练的命名实体模型进行计算，获取所述目标不完全标注的训练样本的预训练模型输出的所述目标不完全标注的训练样本对应的概率预测结果；

采用k-best Viterbi Decoding算法根据所述待训练的命名实体模型的条件随机场模型的状态转移矩阵、所述目标不完全标注的训练样本的所述概率预测结果进行最有可能的标签序列的解码计算，得到所述目标不完全标注的训练样本对应的所述最有可能标签序列集合。

6.根据权利要求3所述的命名实体模型的训练方法，其中，所述自适应损失函数的计算公式L(w,x)为：

$$L(w,x)=(1-\lambda)L_1(w,x)+\lambda L_2(w,x)$$

$$L_1(w,x)=-\log\sum_{y'\in C(y_u)}q(y'|x)p_w(y'|x)$$

$$L_2(w,x)=-\log\sum_{y'\in K_w(x)}p_w(y'|x)$$

其中，$q(y'|x)$是所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析概率分布数据，$p_w(y'|x)$是所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析条件概率数据，$C(y_u)$是所述目标不完全标注的训练样本对应的所述预估标签序列集合，$K_w(x)$所述目标不完全标注的训练样本对应的所述最有可能标签序列集合，log()是对数函数，λ是自适应参数，λ从0逐步增加到1。

7.根据权利要求6所述的命名实体模型的训练方法，其中，所述自适应参数λ的计算公式为：

$$\lambda=\exp\left[\gamma\times\left(\frac{b}{B}-1\right)\right]$$

其中，exp[]是以自然常数e为底的指数函数，b是所述目标不完全标注的训练样本训练所述待训练的命名实体模型时对应的训练步数，B是预设的总训练步数，γ为控制λ增长速度的常数。

8.一种命名实体模型的训练装置，其中，所述装置包括：

训练样本获取模块，用于获取多个不完全标注的训练样本，所述不完全标注的训练样本包括：文本样本数据、不完全标注的标签序列；

预估标签序列集合确定模块，用于采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合，所述预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注；

模型训练模块，用于获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型。

9.一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，其中，所述处理器执行所述计算机程序时实现如下方法步骤：

获取多个不完全标注的训练样本，所述不完全标注的训练样本包括：文本样本数据、不完全标注的标签序列；

采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合，所述预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注；

获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型。

10.根据权利要求9所述的计算机设备，其中，所述采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标

注的训练样本各自对应的预估标签序列集合的步骤，包括：

从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本，作为目标不完全标注的训练样本；

从所述目标不完全标注的训练样本的所述不完全标注的标签序列中提取出已标注实体信息，得到所述目标不完全标注的训练样本对应的已标注实体信息；

采用所述目标不完全标注的训练样本对应的所述已标注实体信息从所述目标不完全标注的训练样本的所述文本样本数据中找出未标注的文字，得到所述目标不完全标注的训练样本对应的未标注文本数据；

分别对所述目标不完全标注的训练样本对应的所述未标注文本数据中每个文字进行所有可能的标签预估，得到所述目标不完全标注的训练样本对应的所述未标注文本数据的各个文字各自对应的预估标签集合；

分别将所述目标不完全标注的训练样本对应的所述未标注文本数据的各个文字各自对应的所述预估标签集合和所述目标不完全标注的训练样本对应的所述已标注实体信息进行所有可能的标签序列组合，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合；

重复执行所述从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本作为目标不完全标注的训练样本的步骤，直至确定所述多个不完全标注的训练样本各自对应的所述预估标签序列集合。

11.根据权利要求9所述的计算机设备，其中，所述获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型的步骤，包括：

从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本，作为目标不完全标注的训练样本；

采用所述初步训练的命名实体模型分别对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个预估标签序列进行概率分布计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析概率分布数据；

采用所述待训练的命名实体模型分别对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个所述预估标签序列进行条件概率计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析条件概率数据；

采用所述待训练的命名实体模型，对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的所有所述预估标签序列进行最有可能的标签序列解析，得到所述目标不完全标注的训练样本对应的最有可能标签序列集合；

将所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析概率分布数据、所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析条件概率数据和所述目标不完全标注的训练样本对应的所述最有可能标签序列集合输入所述自适应损失函数进行计算，得到所述待训练的命名实体模型的损失值，根据所述损失值更新所述待训练的命名实体模型的参数，更新后的所述待训练的命名实体模型被用于下一次计算所述待分析条件概率数据、所述最有可能标签序列集合；

重复执行上述方法步骤直至所述损失值达到第一收敛条件或迭代次数达到

第二收敛条件，将所述损失值达到所述第一收敛条件或迭代次数达到所述第二收敛条件的所述待训练的命名实体模型，确定为所述目标命名实体模型。

12.根据权利要求11所述的计算机设备，其中，所述采用所述初步训练的命名实体模型分别对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个预估标签序列进行概率分布计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析概率分布数据的步骤，包括：

基于前向-后向算法和所述初步训练的命名实体模型，分别对所述目标不完全标注的训练样本的所述文本样本数据的每个字进行各个标签的边缘概率计算，得到所述目标不完全标注的训练样本的所述文本样本数据的各个字各自对应的各个标签的边缘概率数据；

分别根据所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个预估标签序列、所述目标不完全标注的训练样本的所述文本样本数据的各个字各自对应的各个标签的所述边缘概率数据进行各个字各自对应的所述边缘概率数据的相乘计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析概率分布数据。

13.根据权利要求11所述的计算机设备，其中，所述采用所述待训练的命名实体模型，对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的所有所述预估标签序列进行最有可能的标签序列解析，得到所述目标不完全标注的训练样本对应的最有可能标签序列集合的步骤，包括：

将所述目标不完全标注的训练样本的所述文本样本数据输入所述待训练的命名实体模型进行计算，获取所述目标不完全标注的训练样本的预训练模型输出的所述目标不完全标注的训练样本对应的概率预测结果；

采用k-best Viterbi Decoding算法根据所述待训练的命名实体模型的条件随机场模型的状态转移矩阵、所述目标不完全标注的训练样本的所述概率预测结果进行最有可能的标签序列的解码计算，得到所述目标不完全标注的训练样本对应的所述最有可能标签序列集合。

14.根据权利要求11所述的计算机设备，其中，所述自适应损失函数的计算公式L(w,x)为：

$$L(w,x)=(1-\lambda)L_1(w,x)+\lambda L_2(w,x)$$

$$L_1(w,x)=-\log\sum_{y'\in C(y_u)} q(y'|x)p_w(y'|x)$$

$$L_2(w,x)=-\log\sum_{y'\in K_w(x)} p_w(y'|x)$$

其中，$q(y'|x)$是所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析概率分布数据，$p_w(y'|x)$是所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析条件概率数据，$C(y_u)$是所述目标不完全标注的训练样本对应的所述预估标签序列集合，$K_w(x)$所述目标不完全标注的训练样本对应的所述最有可能标签序列集合，log()是对数函数，λ是自适应参数，λ从0逐步增加到1。

15.一种计算机可读存储介质，其上存储有计算机程序，其中，所述计算机程序被处理器执行时实现如下方法步骤：

获取多个不完全标注的训练样本，所述不完全标注的训练样本包括：文本样本数据、不完全标注的标签序列；

采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合，所述预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注；

获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型。

16.根据权利要求15所述的计算机可读存储介质，其中，所述采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合的步骤，包括：

从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本，作为目标不完全标注的训练样本；

从所述目标不完全标注的训练样本的所述不完全标注的标签序列中提取出已标注实体信息，得到所述目标不完全标注的训练样本对应的已标注实体信息；

采用所述目标不完全标注的训练样本对应的所述已标注实体信息从所述目标不完全标注的训练样本的所述文本样本数据中找出未标注的文字，得到所述目标不完全标注的训练样本对应的未标注文本数据；

分别对所述目标不完全标注的训练样本对应的所述未标注文本数据中每个文字进行所有可能的标签预估，得到所述目标不完全标注的训练样本对应的所述未标注文本数据的各个文字各自对应的预估标签集合；

分别将所述目标不完全标注的训练样本对应的所述未标注文本数据的各个文字各自对应的所述预估标签集合和所述目标不完全标注的训练样本对应的所述已标注实体信息进行所有可能的标签序列组合，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合；

重复执行所述从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本作为目标不完全标注的训练样本的步骤，直至确定所述多个不完全标注的训练样本各自对应的所述预估标签序列集合。

17.根据权利要求15所述的计算机可读存储介质，其中，所述获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型的步骤，包括：

从所述多个不完全标注的训练样本中获取一个所述不完全标注的训练样本，作为目标不完全标注的训练样本；

采用所述初步训练的命名实体模型分别对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个预估标签序列进行概率分布计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析概率分布数据；

采用所述待训练的命名实体模型分别对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个所述预估标签序列进行条件概率计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析条件概率数据；

采用所述待训练的命名实体模型，对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的所有所述预估标签序列进行最有可能的标签序列解析，得到所述目标不完全标注的训练样本对应的最有可能标签序列集合；

将所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析概率分布数据、所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析条件概率数据和所述目标不完全标注的训练样本对应的所述最有可能标签序列集合输入所述自适应损失函数进行计算，得到所述待训练的命名实体模型的损失值，根据所述损失值更新所述待训练的命名实体模型的参数，更新后的所述待训练的命名实体模型被用于下一次计算所述待分析条件概率数据、所述最有可能标签序列集合；

重复执行上述方法步骤直至所述损失值达到第一收敛条件或迭代次数达到第二收敛条件，将所述损失值达到所述第一收敛条件或迭代次数达到所述第二收敛条件的所述待训练的命名实体模型，确定为所述目标命名实体模型。

18.根据权利要求17所述的计算机可读存储介质，其中，所述采用所述初步训练的命名实体模型分别对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个预估标签序列进行概率分布计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的待分析概率分布数据的步骤，包括：

基于前向-后向算法和所述初步训练的命名实体模型，分别对所述目标不完全标注的训练样本的所述文本样本数据的每个字进行各个标签的边缘概率计算，得到所述目标不完全标注的训练样本的所述文本样本数据的各个字各自对应的各个标签的边缘概率数据；

分别根据所述目标不完全标注的训练样本对应的所述预估标签序列集合中的每个预估标签序列、所述目标不完全标注的训练样本的所述文本样本数据的各个字各自对应的各个标签的所述边缘概率数据进行各个字各自对应的所述边缘概率数据的相乘计算，得到所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析概率分布数据。

19.根据权利要求17所述的计算机可读存储介质，其中，所述采用所述待训练的命名实体模型，对所述目标不完全标注的训练样本对应的所述预估标签序列集合中的所有所述预估标签序列进行最有可能的标签序列解析，得到所述目标不完全标注的训练样本对应的最有可能标签序列集合的步骤，包括：

将所述目标不完全标注的训练样本的所述文本样本数据输入所述待训练的命名实体模型进行计算，获取所述目标不完全标注的训练样本的预训练模型输出的所述目标不完全标注的训练样本对应的概率预测结果；

采用k-best Viterbi Decoding算法根据所述待训练的命名实体模型的条件随机场模型的状态转移矩阵、所述目标不完全标注的训练样本的所述概率预测结果进行最有可能的标签序列的解码计算，得到所述目标不完全标注的训练样本对应的所述最有可能标签序列集合。

20.根据权利要求17所述的计算机可读存储介质，其中，所述自适应损失函数的计算公式L(w,x)为：

$$L(w,x) = (1-\lambda)L_1(w,x) + \lambda L_2(w,x)$$

$$L_1(w,x) = -\log \sum_{y' \in C(y_u)} q(y'|x)p_w(y'|x)$$

$$L_2(w,x) = -\log \sum_{y' \in K_w(x)} p_w(y'|x)$$

其中，$q(y'|x)$是所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析概率分布数据，$p_w(y'|x)$是所述目标不完全标注的训练样本对应的所述预估标签序列集合中所有所述预估标签序列各自对应的所述待分析条件概率数据，$C(y_u)$是所述目标不完全标注的训练样本对应的所述预估标签序列集合，$K_w(x)$所述目标不完全标注的训练样本对应的所述最有可能标签序列集合，log()是对数函数，λ是自适应参数，λ从0逐步增加到1。

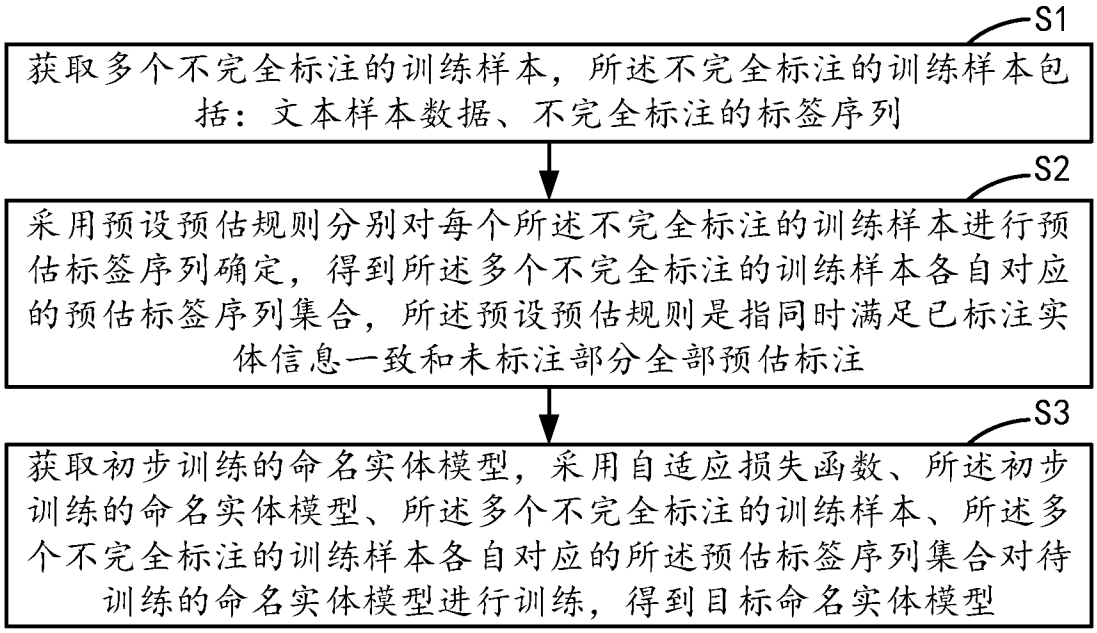

S1
获取多个不完全标注的训练样本，所述不完全标注的训练样本包括：文本样本数据、不完全标注的标签序列
S2
采用预设预估规则分别对每个所述不完全标注的训练样本进行预估标签序列确定，得到所述多个不完全标注的训练样本各自对应的预估标签序列集合，所述预设预估规则是指同时满足已标注实体信息一致和未标注部分全部预估标注
S3
获取初步训练的命名实体模型，采用自适应损失函数、所述初步训练的命名实体模型、所述多个不完全标注的训练样本、所述多个不完全标注的训练样本各自对应的所述预估标签序列集合对待训练的命名实体模型进行训练，得到目标命名实体模型

图 1

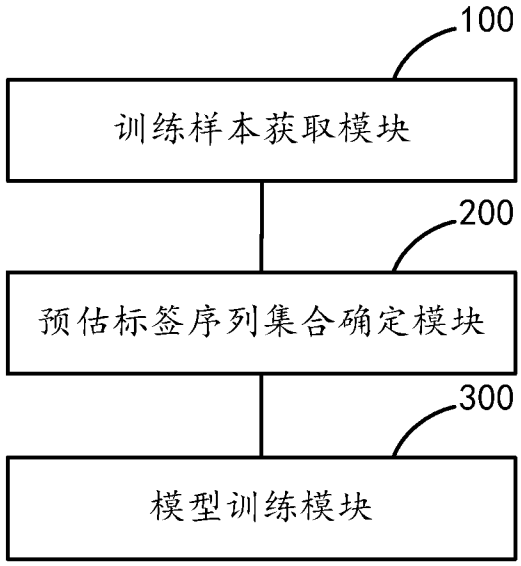

100
训练样本获取模块
200
预估标签序列集合确定模块
300
模型训练模块

图 2

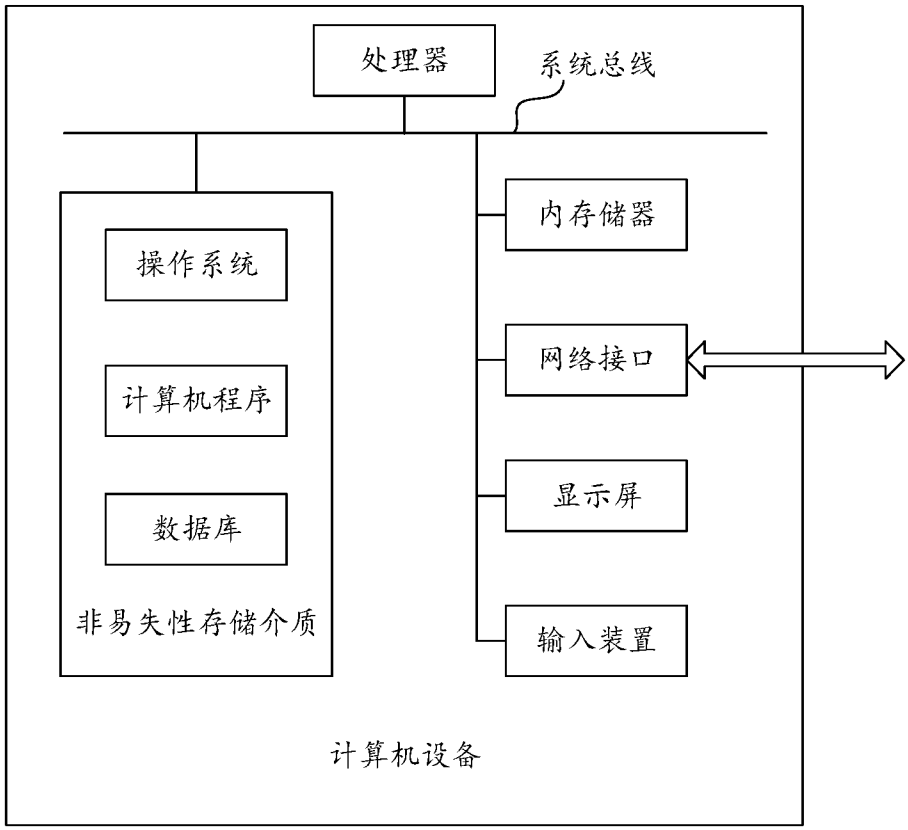


图 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/097545

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/08(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N;G06K;G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS; CNTXT; TWTXT; TWABS; DWPI; VEN; USTXT; EPTXT; WOTXT; CNKI: 命名实体, 标签, 标注, 批注, 注释, 训练, 样本, 模型, 损失函数, 规则, 概率, named-entity, named entity, tag, label, annotation, train, sample, model, loss function, rule, probability

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 112766485 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 07 May 2021 (2021-05-07) claims 1-10, description paragraphs 48-131	1-20
X	CN 111382572 A (BEIJING XIANGNONG HUIYU TECHNOLOGY CO., LTD.) 07 July 2020 (2020-07-07) description paragraphs 46-117	1-20
X	CN 110287480 A (GUANGZHOU DUOYI NETWORK CO., LTD. et al.) 27 September 2019 (2019-09-27) description, paragraphs 46-123	1-20
A	CN 110032649 A (UNIVERSITY OF SCIENCE AND TECHNOLOGY BEIJING) 19 July 2019 (2019-07-19) entire document	1-20
A	US 2020050662 A1 (ORACLE INTERNATIONAL CORPORATION) 13 February 2020 (2020-02-13) entire document	1-20

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
29 August 2021	**09 September 2021**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing 100088, China** Facsimile No. **(86-10)62019451**	Authorized officer Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2021/097545

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 112766485 A	07 May 2021	None	
CN 111382572 A	07 July 2020	None	
CN 110287480 A	27 September 2019	None	
CN 110032649 A	19 July 2019	None	
US 2020050662 A1	13 February 2020	None	

国际检索报告

国际申请号
PCT/CN2021/097545

A. 主题的分类

G06N 3/08(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06N;G06K;G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

CNABS;CNTXT;TWTXT;TWABS;DWPI;VEN;USTXT;EPTXT;WOTXT;CNKI:命名实体，标签，标注，批注，注释，训练，样本，模型，损失函数，规则，概率，named-entity, named entity, tag, label, annotation, train, sample, model, loss function, rule, probability

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
PX	CN 112766485 A (平安科技深圳有限公司) 2021年 5月 7日 (2021-05-07) 权利要求1-10，说明书第48-131段	1-20
X	CN 111382572 A (北京香侬慧语科技有限责任公司) 2020年 7月 7日 (2020-07-07) 说明书第46-117段	1-20
X	CN 110287480 A (广州多益网络股份有限公司 等) 2019年 9月 27日 (2019-09-27) 说明书第46-123段	1-20
A	CN 110032649 A (北京科技大学) 2019年 7月 19日 (2019-07-19) 全文	1-20
A	US 2020050662 A1 (ORACLE INTERNATIONAL CORPORATION) 2020年 2月 13日 (2020-02-13) 全文	1-20

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2021年 8月 29日	2021年 9月 9日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国 北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 房琦 电话号码 86-(20)-28950714

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2021/097545

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 112766485 A	2021年 5月 7日	无	
CN 111382572 A	2020年 7月 7日	无	
CN 110287480 A	2019年 9月 27日	无	
CN 110032649 A	2019年 7月 19日	无	
US 2020050662 A1	2020年 2月 13日	无	