

(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)



(10) 国际公布号
WO 2022/142122 A1

(51) 国际专利分类号:
G06K 9/62 (2006.01)

(21) 国际申请号: PCT/CN2021/097543

(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:
202011633046.9 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

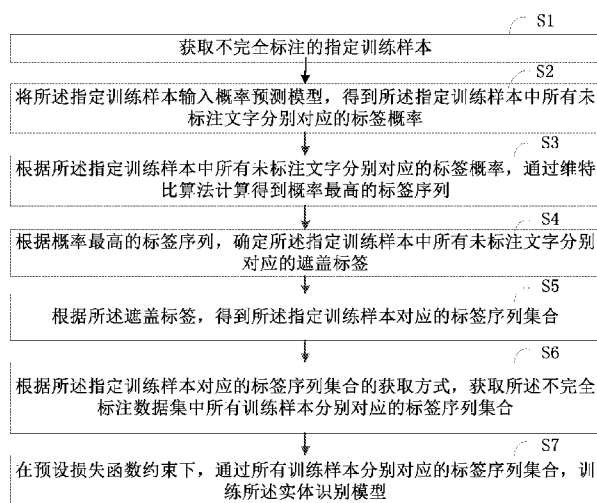
安金融中心 23 楼, Guangdong 518000 (CN)。 郑立颖(ZHENG, Liying); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 胡沛弦(HU, Peixian); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。 徐亮(XU, Liang); 中国广东省深圳市福田区福田街道福安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(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,

(54) Title: METHOD AND APPARATUS FOR TRAINING ENTITY RECOGNITION MODEL, AND DEVICE AND STORAGE MEDIUM

(54) 发明名称: 实体识别模型的训练方法、装置、设备和存储介质



S1 Acquire a designated training sample that is not completely labeled
S2 Input the designated training sample into a probability prediction model, so as to obtain label probabilities corresponding to all unlabeled characters in the designated training sample
S3 Calculate a label sequence with the highest probability according to the label probabilities respectively corresponding to all the unlabeled characters in the designated training sample and by means of a Viterbi algorithm
S4 Determine, according to the label sequence with the highest probability, covering labels respectively corresponding to all the unlabeled characters in the designated training sample
S5 Obtain, according to the covering labels, a label sequence set corresponding to the designated training sample
S6 Acquire, according to an acquisition method for the label sequence set corresponding to the designated training sample, label sequence sets respectively corresponding to all training samples in an incompletely labeled data set
S7 Under the constraint of a preset loss function, train an entity recognition model by means of the label sequence sets respectively corresponding to all the training samples

图 1

(57) Abstract: The present application relates to the field of natural language processing of artificial intelligence. Disclosed is a method for training an entity recognition model, the method comprising: acquiring a designated training sample that is not completely labeled; inputting the designated training sample into a probability prediction model, so as to obtain label probabilities corresponding to all unlabeled characters in the designated training sample; calculating a label sequence with the highest probability according to the label probabilities respectively corresponding to all the unlabeled characters and by means of a Viterbi algorithm; determining, according to the label sequence with the highest probability, covering labels respectively corresponding to all the unlabeled characters in the designated training sample; obtaining, according to the covering labels, a label sequence set corresponding to the designated training sample; acquiring, according to an acquisition method for the label sequence set corresponding to the designated training sample, label sequence sets respectively corresponding to all training samples in an incompletely labeled data set; and under the constraint of a preset

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, 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))。

loss function, training an entity recognition model by means of the label sequence sets respectively corresponding to all the training samples. An actual label sequence can be more easily recognized.

(57) 摘要: 本申请涉及人工智能的自然语言处理领域, 揭示了实体识别模型的训练方法, 包括: 获取不完全标注的指定训练样本; 将指定训练样本输入概率预测模型, 得到指定训练样本中所有未标注文字分别对应的标签概率; 根据所有未标注文字分别对应的标签概率, 通过维特比算法计算得到概率最高的标签序列; 根据概率最高的标签序列, 确定指定训练样本中所有未标注文字分别对应的遮盖标签; 根据遮盖标签, 得到指定训练样本对应的标签序列集合; 根据指定训练样本对应的标签序列集合的获取方式, 获取不完全标注数据集中所有训练样本分别对应的标签序列集合; 在预设损失函数约束下, 通过所有训练样本分别对应的标签序列集合, 训练实体识别模型。更易识别真实标签序列。

实体识别模型的训练方法、装置、设备和存储介质

本申请要求于 2020 年 12 月 31 日提交中国专利局、申请号为 202011633046.9，发明名称为“实体识别模型的训练方法、装置、设备和存储介质”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

本申请涉及人工智能的自然语言处理领域，本申请特别是涉及到实体识别模型的训练方法、装置、设备和存储介质。

背景技术

实体识别模型的训练依赖大量标注完全的数据，但是高质量的标注数据通常需要非常专业的标注人员，导致训练数据获取困难且成本昂贵。现为节约成本，可利用不完全标注数据训练实体识别模型，不完全标注数据即文本中仅有部分实体被标注，而其他未标注的内容可能是非实体，也可能是实体。为提高利用不完全标注数据训练实体识别模型效果，通常将所有符合文本已标注情况的标签序列都考虑到模型训练当中，估计所有可能的标签序列的概率分布信息，并将该信息融入模型训练，使模型可以关注到所有可能的标签序列。但发明人意识到由于命名实体通常有多个类别，且命名实体在文本中稀疏分布，导致候选标签序列的数量随文本中未标记内容的增加而呈指数增长，致使命名实体模型的注意力被分散，不易关注到真实的标签序列，影响识别效果。

技术问题

由于命名实体通常有多个类别，且命名实体在文本中稀疏分布，导致候选标签序列的数量随文本中未标记内容的增加而呈指数增长，致使命名实体模型的注意力被分散，不易关注到真实的标签序列，影响识别效果。

技术方案

本申请的主要目的为提供实体识别模型的训练方法，旨在解决不能及时识别与评估在线语音的发音的技术问题。

第一方面，本申请提出一种实体识别模型的训练方法，包括：

获取不完全标注的指定训练样本，其中，所述指定训练样本为不完全标注数据集中的任一样本；

将所述指定训练样本输入概率预测模型，得到所述指定训练样本中所有未标注文字分别对应的标签概率；

根据所述指定训练样本中所有未标注文字分别对应的标签概率，通过维特比算法计算得到概率最高的标签序列；

根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签；

根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合；

根据所述指定训练样本对应的标签序列集合的获取方式，获取所述不完全标注数据集中所有训练样本分别对应的标签序列集合；

在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型。

第二方面，本申请还提供了一种实体识别模型的训练装置，包括：

第一获取模块，用于获取不完全标注的指定训练样本，其中，所述指定训练样本为不完全标注数据集中的任一样本；

输入模块，用于将所述指定训练样本输入概率预测模型，得到所述指定训练

样本中所有未标注文字分别对应的标签概率；

计算模块，用于根据所述指定训练样本中所有未标注文字分别对应的标签概率，通过维特比算法计算得到概率最高的标签序列；

确定模块，用于根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签；

得到模块，用于根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合；

第二获取模块，用于根据所述指定训练样本对应的标签序列集合的获取方式，获取所述不完全标注数据集中所有训练样本分别对应的标签序列集合；

训练模块，用于在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型。

第三方面，本申请还提供了一种计算机设备，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一种实体识别模型的训练方法；其中，所述一种实体识别模型的训练方法包括：获取不完全标注的指定训练样本，其中，所述指定训练样本为不完全标注数据集中的任一样本；将所述指定训练样本输入概率预测模型，得到所述指定训练样本中所有未标注文字分别对应的标签概率；根据所述指定训练样本中所有未标注文字分别对应的标签概率，通过维特比算法计算得到概率最高的标签序列；根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签；根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合；根据所述指定训练样本对应的标签序列集合的获取方式，获取所述不完全标注数据集中所有训练样本分别对应的标签序列集合；在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型。

第四方面，本申请还提供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一种实体识别模型的训练方法；其中，所述一种实体识别模型的训练方法包括：获取不完全标注的指定训练样本，其中，所述指定训练样本为不完全标注数据集中的任一样本；将所述指定训练样本输入概率预测模型，得到所述指定训练样本中所有未标注文字分别对应的标签概率；根据所述指定训练样本中所有未标注文字分别对应的标签概率，通过维特比算法计算得到概率最高的标签序列；根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签；根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合；根据所述指定训练样本对应的标签序列集合的获取方式，获取所述不完全标注数据集中所有训练样本分别对应的标签序列集合；在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型。

有益效果

本申请通过预测概率模型预测未标注文字的标签概率，并结合维特比算法，得到训练语句对应的标签序列概率，然后选择最有可能的标签序列，并根据最有可能的标签序列减少未标注文字的可能标签数量，以有效减少需要估计概率分布的标签序列的数量，使得实体识别模型更容易识别真实的标签序列，且减少了计算复杂度。

附图说明

图1 本申请一实施例的实体识别模型的训练方法流程示意图；

图2 本申请一实施例的实体识别模型的训练系统流程示意图；

图3 本申请一实施例的计算机设备内部结构示意图。

本发明的最佳实施方式

参照图1，本申请一实施例的实体识别模型的训练方法，包括：

S1：获取不完全标注的指定训练样本，其中，所述指定训练样本为不完全标注数据集中的任一样本；

S2：将所述指定训练样本输入概率预测模型，得到所述指定训练样本中所有未标注文字分别对应的标签概率；

S3：根据所述指定训练样本中所有未标注文字分别对应的标签概率，通过维特比算法计算得到概率最高的标签序列；

S4：根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签；

S5：根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合；

S6：根据所述指定训练样本对应的标签序列集合的获取方式，获取所述不完全标注数据集中所有训练样本分别对应的标签序列集合；

S7：在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型。

本申请实施例中不完全标注的指定训练样本，指用于实体识别的文本语句样本中存在未标注实体标签类型的文字。上述实体标签类型根据任务领域不同而不同。举例地，在描述公司业务的宣传文本中，实体标签包括但不限于公司、地址、人名、时间、组织等。

本申请根据人工标注标签的完整情况构建不完全标注数据集，假设一条不完全标注的训练样本表示为 $x = (x_1, x_2, \dots, x_n)$ ，其中每个 $x_i (i = 1, 2, \dots, n)$ 代表训练样本的文本序列中的一个字。不完全标注的训练样本 x 对应的不完全标注的标签序列表示为 $y_u = (-, y_2, -, \dots, -)$ ，其中 $y_i \in Y$ ， Y 表示实体标签集合， y_i 代表标记人员真实标注的 x_i 的标签，“-”代表未被标记为实体的标签，即该位置的标签可以填补为非实体标签或当前任务领域对应实体标签集中的任意实体标签。通过对未被标记为实体的标签从当前任务领域对应实体标签集进行全排列选择，组合成该训练样本可能存在的所有完全标签序列，组成的集合 $C(y_u)$ 。假设 x 真实的完全标签序列为 $y = (y_1, y_2, \dots, y_n)$ ，则 y 应该是 $C(y_u)$ 中的一条完全标签序列。

本申请通过训练好的概率预测模型预测得到训练文本 x 的 x_i 文字的标签为 $\hat{y}_i$ 的概率， $\hat{y}_i \in Y$ ，再通过维特比算法计算 x 的完全标签序列分别对应的概率，即每个 x_i 是标签 $\hat{y}_i$ 时 x 的完全标签序列为 $\hat{y}$ 的概率， $\hat{y} \in C(y_u)$ 。通过维特比算法，获得概率最高的完全标签序列，并确定遮盖标签进行遮盖删除。通过维特比算法，获得概率最高的完全标签序列的过程如下，假设给定隐式马尔可夫模型HMM状态空间，共有的状态数量，根据初始文字的标签概率，以及初始文字到其他文字的标签序列对应的状态转移概率，产生最有可能的标签序列的概率。通过保存递推中用到的各文字的标签概率，获得维特比路径。然后通过穷举未标注文字已出现过的标签，对应组合训练样本的标签序列，并将训练样本的标签序列和训练样本一起组成训练数据训练实体识别模型。上述实体识别模型和概率预测模型均为BERT (Bidirectional Encoder Representation from Transformers) +CRF (Conditional Random Field, 即条件随机场) 模型架构，区别在于使用的模型参数和输出变量不同。

本申请通过预测概率模型预测未标注文字的标签概率，并结合维特比算法，

得到训练语句对应的标签序列概率，然后选择最有可能的标签序列，并根据最有可能的标签序列减少未标注文字的可能标签数量，以有效减少需要估计概率分布的标签序列的数量，使得实体识别模型更容易识别真实的标签序列，且减少了计算复杂度。

进一步地，所述根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签的步骤 S4，包括：

S41：获取当前实体识别任务中所有标签对应的标签类型集合；

S42：确定概率最高的标签序列组中指定文字分别对应的指定标签，其中，所述指定文字为所述指定训练样本中所有未标注文字中的任意一个，所述指定标签为所述标签类型集合中一种或多种标签；

S43：将所述标签类型集合中所述指定标签之外的标签，作为所述指定文字对应的遮盖标签；

S44：根据所述指定文字对应的遮盖标签的确定方式，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签。

本申请实施例通过 k-best 情况下的维特比算法，获得概率最高的 k 条完全标签序列，记为 $K(x) = \{K_i(x)\}_{i=1,2,\dots,k}$ ，其中第 i 条完全标签序列为 $K_i(x) = [K_i(x_1), K_i(x_2), \dots, K_i(x_n)]$ 。上述 k-best 情况下的维特比算法，表示算法中的每个结点不是只保留最小值，而是保留 K 个最小值，即从升序排列中保留排序靠前的 top K 的最值。然后通过概率最高的 k 条完全标签序列，确定之前未标注文字出现过的标签类型，并将该未标注文字未出现在 k 条完全标签序列中的实体标签作为遮盖标签进行遮盖删除。

进一步地，所述根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合的步骤 S5，包括：

S51：从所述标签类型集合中删除所述遮盖标签，得到所述指定文字对应的预测标签集；

S52：将所述预测标签集标注于所述指定文字上；

S53：根据所述指定文字对应的预测标签集的标注过程，标注所述指定训练样本中所有未标注文字分别对应的预测标签集；

S54：根据所述指定训练样本中所有未标注文字分别对应的预测标签集，按照所述指定训练样本中的文字排布次序，一一对应组合成所述指定训练样本对应的所有标签序列，形成所述标签序列集合。

本申请实施例通过 $K(x)$ 确定遮盖标签并构建预测标签集。对于 x 中的一个未被标记的文字 x_j ，其可能的标签为遮盖标签之外的非实体标签或者并集 $\bigcup_{i=1}^k K_i(x_j)$ 中的一个标签，并集 $\bigcup_{i=1}^k K_i(x_j)$ 表示 $K(x)$ 中文字 x_j 出现的标签并集。通过遮盖或删除未在 k 条标签序列的对应位置出现过的实体标签，得到预测标签集。举例地，文字 x_j 出现过的标签包括“公司”、“人名”和“组织”，文字 x_j 未出现的标签包括“地址”和“时间”，则“地址”和“时间”为遮盖标签，预测标签集由标签“公司”、“人名”和“组织”组成。使得每个未标注文字的可选标签的数量降低，仅可以从预测标签集中选择。对 $y_u = (-, y_2, -, \dots, -)$ 中每个“-”的位置，从各自对应的预测标签集中选择一个标签标注上，就可构成一条完整标签序列，通过删除遮盖标签减少了每个未标注文字的可选标签数量，最后确定可能标签序列的集合为 $S(y_u, K(x))$ ，则 $S(y_u, K(x))$

中的标签序列数量会远少于 $C(y_u)$ 中的标签序列数量。

进一步地，所述根据所述指定训练样本中所有未标注文字分别对应的预测标

签集,按照所述指定训练样本中的文字排布次序,一一对应组合成所述指定训练样本对应的所有标签序列,形成所述标签序列集合的步骤 S53,包括:

S531:判断第一文字是否为标注文字,其中,所述第一文字为所述指定训练样本中的任一文字;

S532:若是,则获取所述第一文字对应的标注标签;

S533:判断排布于所述第一文字之后的第二文字是否为标注文字;

S534:若否,则获取所述第二文字对应的预测标签集;

S535:将所述第二文字对应的预测标签集分别标注于所述第二文字上,并分别连接所述第一文字对应的标注标签,形成所述第一文字到所述第二文字的标签路径;

S536:按照所述第一文字到所述第二文字的标签路径的形成方式,形成所述指定训练样本中所有文字对应的所有标签路径;

S537:将所述指定训练样本中所有文字对应的所有标签路径,作为所述指定训练样本对应的所有标签序列,形成所述标签序列集合。

本申请实施例以训练样本中一个标注标签的字接续一个未标注标签的字为例进行详细说明,但并不限定文字的排布规律,只是标注标签的字的标签获取过程,以及未标注标签的字的标签获取过程类似。根据未标注标签的字对应的预测标签集中存在的标签数量,在标注标签的字接续一个未标注标签的字之间,形成相同数量的标签路径,通过将训练样本中所有标注的字和所有未标注的字按照上述第一文字和第二文字间的标签路径的形成过程,得到任意两个相邻字之间的所有标签路径,然后按照文字的排布次序,依次接续连接两个相邻字之间的所有标签路径,得到所有的标签序列,形成所述标签序列集合。

进一步地,所述在预设损失函数约束下,通过所有训练样本分别对应的标签序列集合,训练所述实体识别模型的步骤 S7,包括:

S71:将所有训练样本分别对应的标签序列集合输入交叉验证模型,得到所有训练样本的标签序列分别对应的标签序列概率;

S72:将所有训练样本的标签序列分别对应的标签序列概率,一一对应设置为各所述标签序列分别对应的分配权重;

S73:将携带所述分配权重的标签序列和各所述标签序列分别对应的训练样本,组成训练数据;

S74:将所述训练数据输入所述实体识别模型,训练至所述预设损失函数收敛。

本申请实施例中,交叉验证模型为 BERT+CRF 模型架构,区别在于使用的模型参数和输出变量与实体识别模型不同。上述预设损失函数为 $L(w, x) = -\log \sum_{y' \in S(y_u, K(x))} q(y' | x) p_w(y' | x)$, w 为所述实体识别模型的模型参数, $q(y' | x)$ 表示分配权重。本申请通过交叉验证模型估计标签序列集合中各标签序列的概率分布 $p(y' | x)$, 其中 $y' \in S(y_u, K(x))$ 。从而获得标签序列集合中各标签序列的概率分布 $q(y' | x)$, $q(y' | x) = \frac{\exp(\log p(y' | x)/T)}{\sum_{y'} \exp(\log p(y' | x)/T)}$, 其中 T 为温度参数,

$T > 0$ 。将估计得到的标签序列集合中各标签序列的概率分布 $q(y' | x)$, 作为每条 x 对应的每条可能标签序列的分配权重。

进一步地,所述将所有训练样本分别对应的标签序列集合输入交叉验证模型,得到所有训练样本的标签序列分别对应的标签序列概率的步骤 S71, 包括:

S711: 将所有训练样本分别对应的标签序列集合中的标签序列, 均分为第一部分数据和第二部分数据;

S712: 将所述第一部分数据输入至所述交叉验证模型, 训练得到第一验证模型, 将所述第二部分数据输入至所述交叉验证模型, 训练得到第二验证模型;

S713: 将所述第二部分数据输入所述第一验证模型, 得到所述第二部分数据中每条标签序列分别对应的标签序列概率, 将所述第一部分数据输入所述第二验证模型, 得到所述第一部分数据中每条标签序列分别对应的标签序列概率。

本申请实施例利用交叉验证模型进行交叉验证时, 先将每条 x 对应的每条可能标签序列的分配权重和每条 x 一一对应匹配, 得到所有训练样本的标签序列组成的训练数据, 然后将训练数据按照数量均分为两部分, 一部分作为训练集, 另一部分作为验证集。即在一部分训练数据上训练序列标注模型, 并通过训练的序列标注模型预测另一部分训练数据中每条 x 的标签序列概率 $p(y' | x)$ 。

参照图 2, 本申请一实施例的实体识别模型的训练装置, 包括:

第一获取模块 1, 用于获取不完全标注的指定训练样本, 其中, 所述指定训练样本为不完全标注数据集中的任一样本;

输入模块 2, 用于将所述指定训练样本输入概率预测模型, 得到所述指定训练样本中所有未标注文字分别对应的标签概率;

计算模块 3, 用于根据所述指定训练样本中所有未标注文字分别对应的标签概率, 通过维特比算法计算得到概率最高的标签序列;

确定模块 4, 用于根据概率最高的标签序列, 确定所述指定训练样本中所有未标注文字分别对应的遮盖标签;

得到模块 5, 用于根据所述遮盖标签, 得到所述指定训练样本对应的标签序列集合;

第二获取模块 6, 用于根据所述指定训练样本对应的标签序列集合的获取方式, 获取所述不完全标注数据集中所有训练样本分别对应的标签序列集合;

训练模块 7, 用于在预设损失函数约束下, 通过所有训练样本分别对应的标签序列集合, 训练所述实体识别模型。

本申请实施例的相关解释, 适用方法对应部分的解释, 不赘述。

进一步地, 确定模块 4, 包括:

获取单元, 用于获取当前实体识别任务中所有标签对应的标签类型集合;

第一确定单元, 用于确定概率最高的标签序列组中指定文字分别对应的指定标签, 其中, 所述指定文字为所述指定训练样本中所有未标注文字中的任意一个, 所述指定标签为所述标签类型集合中一种或多种标签;

作为单元, 用于将所述标签类型集合中所述指定标签之外的标签, 作为所述指定文字对应的遮盖标签;

第二确定单元, 用于根据所述指定文字对应的遮盖标签的确定方式, 确定所述指定训练样本中所有未标注文字分别对应的遮盖标签。

进一步地, 得到模块 5, 包括:

删除单元, 用于从所述标签类型集合中删除所述遮盖标签, 得到所述指定文字对应的预测标签集;

第一标注单元, 用于将所述预测标签集标注于所述指定文字上;

第二标注单元, 用于根据所述指定文字对应的预测标签集的标注过程, 标注所述指定训练样本中所有未标注文字分别对应的预测标签集;

组合单元,用于根据所述指定训练样本中所有未标注文字分别对应的预测标签集,按照所述指定训练样本中的文字排布次序,一一对应组合成所述指定训练样本对应的所有标签序列,形成所述标签序列集合。

进一步地,组合单元,包括:

第一判断子单元,用于判断第一文字是否为标注文字,其中,所述第一文字为所述指定训练样本中的任一文字;

第一获取子单元,用于若为标注文字,则获取所述第一文字对应的标注标签;

第二判断子单元,用于判断排布于所述第一文字之后的第二文字是否为标注文字;

第二获取子单元,用于若为标注文字,则获取所述第二文字对应的预测标签集;

标注子单元,用于将所述第二文字对应的预测标签集分别标注于所述第二文字上,并分别连接所述第一文字对应的标注标签,形成所述第一文字到所述第二文字的标签路径;

形成子单元,用于按照所述第一文字到所述第二文字的标签路径的形成方式,形成所述指定训练样本中所有文字对应的所有标签路径;

作为子单元,用于将所述指定训练样本中所有文字对应的所有标签路径,作为所述指定训练样本对应的所有标签序列,形成所述标签序列集合。

进一步地,训练模块7,包括:

输入单元,用于将所有训练样本分别对应的标签序列集合输入交叉验证模型,得到所有训练样本的标签序列分别对应的标签序列概率;

设置单元,用于将所有训练样本的标签序列分别对应的标签序列概率,一一对应设置为各所述标签序列分别对应的分配权重;

组成单元,用于将携带所述分配权重的标签序列和各所述标签序列分别对应的训练样本,组成训练数据;

训练单元,用于将所述训练数据输入所述实体识别模型,训练至所述预设损失函数收敛。

进一步地,输入单元,包括:

均分子单元,用于将所有训练样本分别对应的标签序列集合中的标签序列,均分为第一部分数据和第二部分数据;

训练子单元,用于将所述第一部分数据输入至所述交叉验证模型,训练得到第一验证模型,将所述第二部分数据输入至所述交叉验证模型,训练得到第二验证模型;

输入子单元,用于将所述第二部分数据输入所述第一验证模型,得到所述第二部分数据中每条标签序列分别对应的标签序列概率,将所述第一部分数据输入所述第二验证模型,得到所述第一部分数据中每条标签序列分别对应的标签序列概率。

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。该内存储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存储

实体识别模型的训练过程需要的所有数据。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实体识别模型的训练方法。

上述处理器执行上述实体识别模型的训练方法，包括：获取不完全标注的指定训练样本，其中，所述指定训练样本为不完全标注数据集中的任一样本；将所述指定训练样本输入概率预测模型，得到所述指定训练样本中所有未标注文字分别对应的标签概率；根据所述指定训练样本中所有未标注文字分别对应的标签概率，通过维特比算法计算得到概率最高的标签序列；根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签；根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合；根据所述指定训练样本对应的标签序列集合的获取方式，获取所述不完全标注数据集中所有训练样本分别对应的标签序列集合；在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型。

上述计算机设备，通过预测概率模型预测未标注文字的标签概率，并结合维特比算法，得到训练语句对应的标签序列概率，然后选择最有可能的标签序列，并根据最有可能的标签序列减少未标注文字的可能标签数量，以有效减少需要估计概率分布的标签序列的数量，使得实体识别模型更容易识别真实的标签序列，且减少了计算复杂度。

在一个实施例中，上述处理器根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签的步骤，包括：获取当前实体识别任务中所有标签对应的标签类型集合；确定概率最高的标签序列组中指定文字分别对应的指定标签，其中，所述指定文字为所述指定训练样本中所有未标注文字中的任意一个，所述指定标签为所述标签类型集合中一种或多种标签；将所述标签类型集合中所述指定标签之外的标签，作为所述指定文字对应的遮盖标签；根据所述指定文字对应的遮盖标签的确定方式，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签。

在一个实施例中，上述处理器根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合的步骤，包括：从所述标签类型集合中删除所述遮盖标签，得到所述指定文字对应的预测标签集；将所述预测标签集标注于所述指定文字上；根据所述指定文字对应的预测标签集的标注过程，标注所述指定训练样本中所有未标注文字分别对应的预测标签集；根据所述指定训练样本中所有未标注文字分别对应的预测标签集，按照所述指定训练样本中的文字排布次序，一一对应组合成所述指定训练样本对应的所有标签序列，形成所述标签序列集合。

在一个实施例中，上述处理器根据所述指定训练样本中所有未标注文字分别对应的预测标签集，按照所述指定训练样本中的文字排布次序，一一对应组合成所述指定训练样本对应的所有标签序列，形成所述标签序列集合的步骤，包括：判断第一文字是否为标注文字，其中，所述第一文字为所述指定训练样本中的任一文字；若是，则获取所述第一文字对应的标注标签；判断排布于所述第一文字之后的第二文字是否为标注文字；若否，则获取所述第二文字对应的预测标签集；将所述第二文字对应的预测标签集分别标注于所述第二文字上，并分别连接所述第一文字对应的标注标签，形成所述第一文字到所述第二文字的标签路径；按照所述第一文字到所述第二文字的标签路径的形成方式，形成所述指定训练样本中所有文字对应的所有标签路径；将所述指定训练样本中所有文字对应的所有标签路径，作为所述指定训练样本对应的所有标签序列，形成所述标签序列集合。

在一个实施例中，上述处理器在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型的步骤，包括：将所有训练样本分别对应的标签序列集合输入交叉验证模型，得到所有训练样本的标签序列分别对应的标签序列概率；将所有训练样本的标签序列分别对应的标签序列概率，一一对应设置为各所述标签序列分别对应的分配权重；将携带所述分配权重的标签序列和各所述标签序列分别对应的训练样本，组成训练数据；将所述训练数据输入所述实体识别模型，训练至所述预设损失函数收敛。

在一个实施例中，上述处理器将所有训练样本分别对应的标签序列集合输入交叉验证模型，得到所有训练样本的标签序列分别对应的标签序列概率的步骤，包括：将所有训练样本分别对应的标签序列集合中的标签序列，均分为第一部分数据和第二部分数据；将所述第一部分数据输入至所述交叉验证模型，训练得到第一验证模型，将所述第二部分数据输入至所述交叉验证模型，训练得到第二验证模型；将所述第二部分数据输入所述第一验证模型，得到所述第二部分数据中每条标签序列分别对应的标签序列概率，将所述第一部分数据输入所述第二验证模型，得到所述第一部分数据中每条标签序列分别对应的标签序列概率。

本领域技术人员可以理解，图3中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定。

本申请一实施例还提供一种计算机可读存储介质，所述存储介质为易失性存储介质或非易失性存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现实体识别模型的训练方法，包括：获取不完全标注的指定训练样本，其中，所述指定训练样本为不完全标注数据集中的任一样本；将所述指定训练样本输入概率预测模型，得到所述指定训练样本中所有未标注文字分别对应的标签概率；根据所述指定训练样本中所有未标注文字分别对应的标签概率，通过维特比算法计算得到概率最高的标签序列；根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签；根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合；根据所述指定训练样本对应的标签序列集合的获取方式，获取所述不完全标注数据集中所有训练样本分别对应的标签序列集合；在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型。

上述计算机可读存储介质，通过预测概率模型预测未标注文字的标签概率，并结合维特比算法，得到训练语句对应的标签序列概率，然后选择最有可能的标签序列，并根据最有可能的标签序列减少未标注文字的可能标签数量，以有效减少需要估计概率分布的标签序列的数量，使得实体识别模型更容易识别真实的标签序列，且减少了计算复杂度。

在一个实施例中，上述处理器根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签的步骤，包括：获取当前实体识别任务中所有标签对应的标签类型集合；确定概率最高的标签序列组中指定文字分别对应的指定标签，其中，所述指定文字为所述指定训练样本中所有未标注文字中的任意一个，所述指定标签为所述标签类型集合中一种或多种标签；将所述标签类型集合中所述指定标签之外的标签，作为所述指定文字对应的遮盖标签；根据所述指定文字对应的遮盖标签的确定方式，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签。

在一个实施例中，上述处理器根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合的步骤，包括：从所述标签类型集合中删除所述遮盖标签，得

到所述指定文字对应的预测标签集；将所述预测标签集标注于所述指定文字上；根据所述指定文字对应的预测标签集的标注过程，标注所述指定训练样本中所有未标注文字分别对应的预测标签集；根据所述指定训练样本中所有未标注文字分别对应的预测标签集，按照所述指定训练样本中的文字排布次序，一一对应组合成所述指定训练样本对应的所有标签序列，形成所述标签序列集合。

在一个实施例中，上述处理器根据所述指定训练样本中所有未标注文字分别对应的预测标签集，按照所述指定训练样本中的文字排布次序，一一对应组合成所述指定训练样本对应的所有标签序列，形成所述标签序列集合的步骤，包括：判断第一文字是否为标注文字，其中，所述第一文字为所述指定训练样本中的任意一文字；若是，则获取所述第一文字对应的标注标签；判断排布于所述第一文字之后的第二文字是否为标注文字；若否，则获取所述第二文字对应的预测标签集；将所述第二文字对应的预测标签集分别标注于所述第二文字上，并分别连接所述第一文字对应的标注标签，形成所述第一文字到所述第二文字的标签路径；按照所述第一文字到所述第二文字的标签路径的形成方式，形成所述指定训练样本中所有文字对应的所有标签路径；将所述指定训练样本中所有文字对应的所有标签路径，作为所述指定训练样本对应的所有标签序列，形成所述标签序列集合。

在一个实施例中，上述处理器在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型的步骤，包括：将所有训练样本分别对应的标签序列集合输入交叉验证模型，得到所有训练样本的标签序列分别对应的标签序列概率；将所有训练样本的标签序列分别对应的标签序列概率，一一对应设置为各所述标签序列分别对应的分配权重；将携带所述分配权重的标签序列和各所述标签序列分别对应的训练样本，组成训练数据；将所述训练数据输入所述实体识别模型，训练至所述预设损失函数收敛。

在一个实施例中，上述处理器将所有训练样本分别对应的标签序列集合输入交叉验证模型，得到所有训练样本的标签序列分别对应的标签序列概率的步骤，包括：将所有训练样本分别对应的标签序列集合中的标签序列，均分为第一部分数据和第二部分数据；将所述第一部分数据输入至所述交叉验证模型，训练得到第一验证模型，将所述第二部分数据输入至所述交叉验证模型，训练得到第二验证模型；将所述第二部分数据输入所述第一验证模型，得到所述第二部分数据中每条标签序列分别对应的标签序列概率，将所述第一部分数据输入所述第二验证模型，得到所述第一部分数据中每条标签序列分别对应的标签序列概率。

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，上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。

权利要求书

- 1、一种实体识别模型的训练方法，其中，包括：
获取不完全标注的指定训练样本，其中，所述指定训练样本为不完全标注数据集中的任一样本；
将所述指定训练样本输入概率预测模型，得到所述指定训练样本中所有未标注文字分别对应的标签概率；
根据所述指定训练样本中所有未标注文字分别对应的标签概率，通过维特比算法计算得到概率最高的标签序列；
根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签；
根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合；
根据所述指定训练样本对应的标签序列集合的获取方式，获取所述不完全标注数据集中所有训练样本分别对应的标签序列集合；
在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型。
- 2、根据权利要求1所述的实体识别模型的训练方法，其中，所述根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签的步骤，包括：
获取当前实体识别任务中所有标签对应的标签类型集合；
确定概率最高的标签序列组中指定文字分别对应的指定标签，其中，所述指定文字为所述指定训练样本中所有未标注文字中的任意一个，所述指定标签为所述标签类型集合中一种或多种标签；
将所述标签类型集合中所述指定标签之外的标签，作为所述指定文字对应的遮盖标签；
根据所述指定文字对应的遮盖标签的确定方式，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签。
- 3、根据权利要求1所述的实体识别模型的训练方法，其中，所述根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合的步骤，包括：
从所述标签类型集合中删除所述遮盖标签，得到所述指定文字对应的预测标签集；
将所述预测标签集标注于所述指定文字上；
根据所述指定文字对应的预测标签集的标注过程，标注所述指定训练样本中所有未标注文字分别对应的预测标签集；
根据所述指定训练样本中所有未标注文字分别对应的预测标签集，按照所述指定训练样本中的文字排布次序，一一对应组合成所述指定训练样本对应的所有标签序列，形成所述标签序列集合。
- 4、根据权利要求3所述的实体识别模型的训练方法，其中，所述根据所述指定训练样本中所有未标注文字分别对应的预测标签集，按照所述指定训练样本中的文字排布次序，一一对应组合成所述指定训练样本对应的所有标签序列，形成所述标签序列集合的步骤，包括：
判断第一文字是否为标注文字，其中，所述第一文字为所述指定训练样本中的任一文字；
若是，则获取所述第一文字对应的标注标签；
判断排布于所述第一文字之后的第二文字是否为标注文字；

若否，则获取所述第二文字对应的预测标签集；

将所述第二文字对应的预测标签集分别标注于所述第二文字上，并分别连接所述第一文字对应的标注标签，形成所述第一文字到所述第二文字的标签路径；

按照所述第一文字到所述第二文字的标签路径的形成方式，形成所述指定训练样本中所有文字对应的所有标签路径；

将所述指定训练样本中所有文字对应的所有标签路径，作为所述指定训练样本对应的所有标签序列，形成所述标签序列集合。

5、根据权利要求1所述的实体识别模型的训练方法，其中，所述在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型的步骤，包括：

将所有训练样本分别对应的标签序列集合输入交叉验证模型，得到所有训练样本的标签序列分别对应的标签序列概率；

将所有训练样本的标签序列分别对应的标签序列概率，一一对应设置为各所述标签序列分别对应的分配权重；

将携带所述分配权重的标签序列和各所述标签序列分别对应的训练样本，组成训练数据；

将所述训练数据输入所述实体识别模型，训练至所述预设损失函数收敛。

6、根据权利要求5所述的实体识别模型的训练方法，其中，所述将所有训练样本分别对应的标签序列集合输入交叉验证模型，得到所有训练样本的标签序列分别对应的标签序列概率的步骤，包括：

将所有训练样本分别对应的标签序列集合中的标签序列，均分为第一部分数据和第二部分数据；

将所述第一部分数据输入至所述交叉验证模型，训练得到第一验证模型，将所述第二部分数据输入至所述交叉验证模型，训练得到第二验证模型；

将所述第二部分数据输入所述第一验证模型，得到所述第二部分数据中每条标签序列分别对应的标签序列概率，将所述第一部分数据输入所述第二验证模型，得到所述第一部分数据中每条标签序列分别对应的标签序列概率。

7、一种实体识别模型的训练装置，其中，包括：

第一获取模块，用于获取不完全标注的指定训练样本，其中，所述指定训练样本为不完全标注数据集中的任一样本；

输入模块，用于将所述指定训练样本输入概率预测模型，得到所述指定训练样本中所有未标注文字分别对应的标签概率；

计算模块，用于根据所述指定训练样本中所有未标注文字分别对应的标签概率，通过维特比算法计算得到概率最高的标签序列；

确定模块，用于根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签；

得到模块，用于根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合；

第二获取模块，用于根据所述指定训练样本对应的标签序列集合的获取方式，获取所述不完全标注数据集中所有训练样本分别对应的标签序列集合；

训练模块，用于在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型。

8、根据权利要求 7 所述的实体识别模型的训练装置，其中，确定模块，包括：

获取单元，用于获取当前实体识别任务中所有标签对应的标签类型集合；

第一确定单元，用于确定概率最高的标签序列组中指定文字分别对应的指定标签，其中，所述指定文字为所述指定训练样本中所有未标注文字中的任意一个，所述指定标签为所述标签类型集合中一种或多种标签；

作为单元，用于将所述标签类型集合中所述指定标签之外的标签，作为所述指定文字对应的遮盖标签；

第二确定单元，用于根据所述指定文字对应的遮盖标签的确定方式，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签。

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一种实体识别模型的训练方法；其中，所述一种实体识别模型的训练方法包括：

获取不完全标注的指定训练样本，其中，所述指定训练样本为不完全标注数据集中的任一样本；

将所述指定训练样本输入概率预测模型，得到所述指定训练样本中所有未标注文字分别对应的标签概率；

根据所述指定训练样本中所有未标注文字分别对应的标签概率，通过维特比算法计算得到概率最高的标签序列；

根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签；

根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合；

根据所述指定训练样本对应的标签序列集合的获取方式，获取所述不完全标注数据集中所有训练样本分别对应的标签序列集合；

在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型。

10、根据权利要求 9 所述的计算机设备，其中，所述根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签的步骤，包括：

获取当前实体识别任务中所有标签对应的标签类型集合；

确定概率最高的标签序列组中指定文字分别对应的指定标签，其中，所述指定文字为所述指定训练样本中所有未标注文字中的任意一个，所述指定标签为所述标签类型集合中一种或多种标签；

将所述标签类型集合中所述指定标签之外的标签，作为所述指定文字对应的遮盖标签；

根据所述指定文字对应的遮盖标签的确定方式，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签。

11、根据权利要求 9 所述的计算机设备，其中，所述根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合的步骤，包括：

从所述标签类型集合中删除所述遮盖标签，得到所述指定文字对应的预测标签集；

将所述预测标签集标注于所述指定文字上；

根据所述指定文字对应的预测标签集的标注过程，标注所述指定训练样本中所有未标注文字分别对应的预测标签集；

根据所述指定训练样本中所有未标注文字分别对应的预测标签集，按照所述指定训练样本中的文字排布次序，一一对应组合成所述指定训练样本对应的所有标签序列，形成所述标签序列集合。

12、根据权利要求 11 所述的计算机设备，其中，所述根据所述指定训练样本中所有未标注文字分别对应的预测标签集，按照所述指定训练样本中的文字排布次序，一一对应组合成所述指定训练样本对应的所有标签序列，形成所述标签序列集合的步骤，包括：

判断第一文字是否为标注文字，其中，所述第一文字为所述指定训练样本中的任一文字；

若是，则获取所述第一文字对应的标注标签；

判断排布于所述第一文字之后的第二文字是否为标注文字；

若否，则获取所述第二文字对应的预测标签集；

将所述第二文字对应的预测标签集分别标注于所述第二文字上，并分别连接所述第一文字对应的标注标签，形成所述第一文字到所述第二文字的标签路径；

按照所述第一文字到所述第二文字的标签路径的形成方式，形成所述指定训练样本中所有文字对应的所有标签路径；

将所述指定训练样本中所有文字对应的所有标签路径，作为所述指定训练样本对应的所有标签序列，形成所述标签序列集合。

13、根据权利要求 9 所述的计算机设备，其中，所述在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型的步骤，包括：

将所有训练样本分别对应的标签序列集合输入交叉验证模型，得到所有训练样本的标签序列分别对应的标签序列概率；

将所有训练样本的标签序列分别对应的标签序列概率，一一对应设置为各所述标签序列分别对应的分配权重；

将携带所述分配权重的标签序列和各所述标签序列分别对应的训练样本，组成训练数据；

将所述训练数据输入所述实体识别模型，训练至所述预设损失函数收敛。

14、根据权利要求 13 所述的计算机设备，其中，所述将所有训练样本分别对应的标签序列集合输入交叉验证模型，得到所有训练样本的标签序列分别对应的标签序列概率的步骤，包括：

将所有训练样本分别对应的标签序列集合中的标签序列，均分为第一部分数据和第二部分数据；

将所述第一部分数据输入至所述交叉验证模型，训练得到第一验证模型，将所述第二部分数据输入至所述交叉验证模型，训练得到第二验证模型；

将所述第二部分数据输入所述第一验证模型，得到所述第二部分数据中每条标签序列分别对应的标签序列概率，将所述第一部分数据输入所述第二验证模型，得到所述第一部分数据中每条标签序列分别对应的标签序列概率。

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一种实体识别模型的训练方法；其中，所述一种实体识别模型的训练方法包括：

获取不完全标注的指定训练样本，其中，所述指定训练样本为不完全标注数据集中的任一样本；

将所述指定训练样本输入概率预测模型，得到所述指定训练样本中所有未标注文字分别对应的标签概率；

根据所述指定训练样本中所有未标注文字分别对应的标签概率，通过维特比算法计算得到概率最高的标签序列；

根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签；

根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合；

根据所述指定训练样本对应的标签序列集合的获取方式，获取所述不完全标注数据集中所有训练样本分别对应的标签序列集合；

在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型。

16、根据权利要求 15 所述的计算机可读存储介质，其中，所述根据概率最高的标签序列，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签的步骤，包括：

获取当前实体识别任务中所有标签对应的标签类型集合；

确定概率最高的标签序列组中指定文字分别对应的指定标签，其中，所述指定文字为所述指定训练样本中所有未标注文字中的任意一个，所述指定标签为所述标签类型集合中一种或多种标签；

将所述标签类型集合中所述指定标签之外的标签，作为所述指定文字对应的遮盖标签；

根据所述指定文字对应的遮盖标签的确定方式，确定所述指定训练样本中所有未标注文字分别对应的遮盖标签。

17、根据权利要求 15 所述的计算机可读存储介质，其中，所述根据所述遮盖标签，得到所述指定训练样本对应的标签序列集合的步骤，包括：

从所述标签类型集合中删除所述遮盖标签，得到所述指定文字对应的预测标签集；

将所述预测标签集标注于所述指定文字上；

根据所述指定文字对应的预测标签集的标注过程，标注所述指定训练样本中所有未标注文字分别对应的预测标签集；

根据所述指定训练样本中所有未标注文字分别对应的预测标签集，按照所述指定训练样本中的文字排布次序，一一对应组合成所述指定训练样本对应的所有标签序列，形成所述标签序列集合。

18、根据权利要求 17 所述的计算机可读存储介质，其中，所述根据所述指定训练样本中所有未标注文字分别对应的预测标签集，按照所述指定训练样本中的文字排布次序，一一对应组合成所述指定训练样本对应的所有标签序列，形成所述标签序列集合的步骤，包括：

判断第一文字是否为标注文字，其中，所述第一文字为所述指定训练样本中的任一文字；

若是，则获取所述第一文字对应的标注标签；

判断排布于所述第一文字之后的第二文字是否为标注文字；

若否，则获取所述第二文字对应的预测标签集；

将所述第二文字对应的预测标签集分别标注于所述第二文字上，并分别连接所述第一文字对应的标注标签，形成所述第一文字到所述第二文字的标签路径；

按照所述第一文字到所述第二文字的标签路径的形成方式，形成所述指定训练样本中所有文字对应的所有标签路径；

将所述指定训练样本中所有文字对应的所有标签路径，作为所述指定训练样本对应的所有标签序列，形成所述标签序列集合。

19、根据权利要求 15 所述的计算机可读存储介质，其中，所述在预设损失函数约束下，通过所有训练样本分别对应的标签序列集合，训练所述实体识别模型的步骤，包括：

将所有训练样本分别对应的标签序列集合输入交叉验证模型，得到所有训练样本的标签序列分别对应的标签序列概率；

将所有训练样本的标签序列分别对应的标签序列概率，一一对应设置为各所述标签序列分别对应的分配权重；

将携带所述分配权重的标签序列和各所述标签序列分别对应的训练样本，组成训练数据；

将所述训练数据输入所述实体识别模型，训练至所述预设损失函数收敛。

20、根据权利要求 19 所述的计算机可读存储介质，其中，所述将所有训练样本分别对应的标签序列集合输入交叉验证模型，得到所有训练样本的标签序列分别对应的标签序列概率的步骤，包括：

将所有训练样本分别对应的标签序列集合中的标签序列，均分为第一部分数据和第二部分数据；

将所述第一部分数据输入至所述交叉验证模型，训练得到第一验证模型，将所述第二部分数据输入至所述交叉验证模型，训练得到第二验证模型；

将所述第二部分数据输入所述第一验证模型，得到所述第二部分数据中每条标签序列分别对应的标签序列概率，将所述第一部分数据输入所述第二验证模型，得到所述第一部分数据中每条标签序列分别对应的标签序列概率。

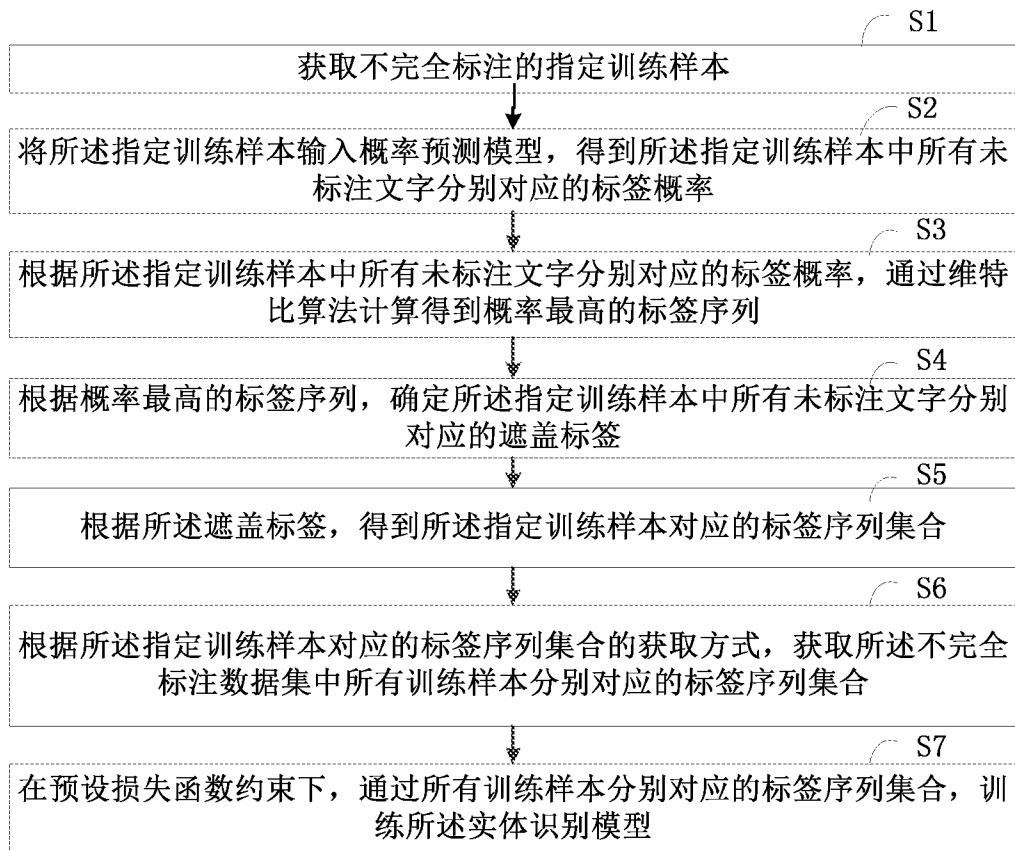


图 1

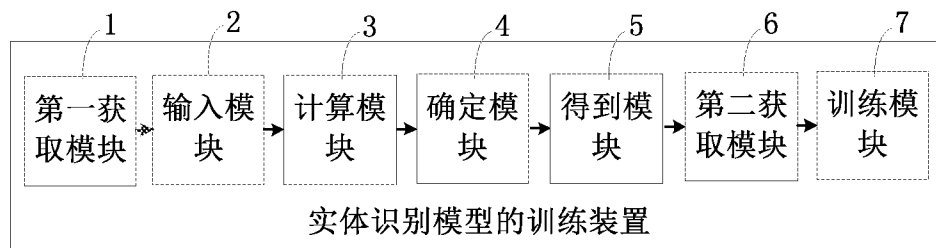


图 2

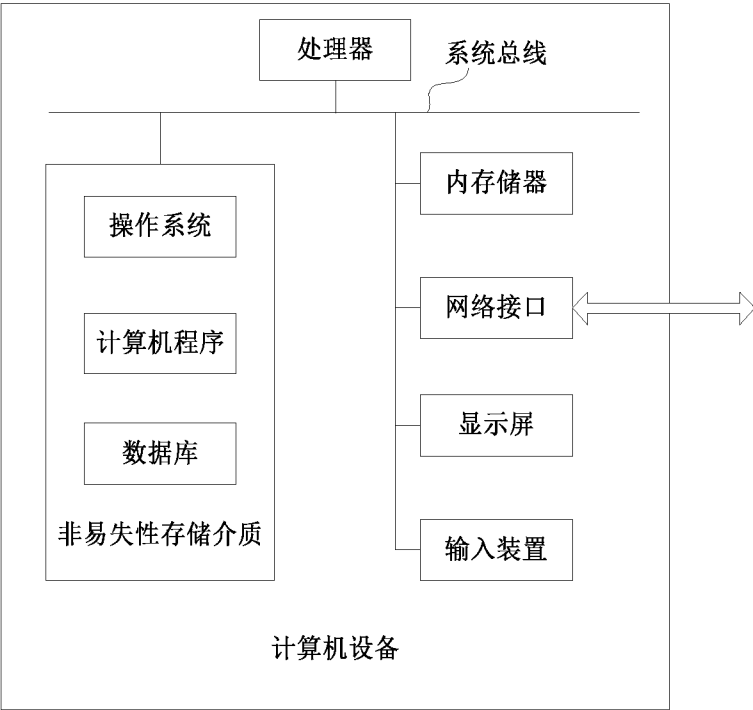


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2021/097543

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/62(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)

G06K, G06N, 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)

CNTXT, CNABS, DWPI, SIPOABS, CNKI: 弱标注, 不完全标注, 局部标注, 少量标注, 较少标注, 命名实体识别, 实体识别, labeled, named entity recognition

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 112733911 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 30 April 2021 (2021-04-30) description paragraphs [0001]-[0153]	1-20
X	CN 111611802 A (SOOCHOW UNIVERSITY) 01 September 2020 (2020-09-01) description, paragraphs [0050]-[0088]	1, 7, 9, 15
A	CN 110070183 A (INSTITUTE OF INFORMATION ENGINEERING, CHINESE ACADEMY OF SCIENCES) 30 July 2019 (2019-07-30) entire document	1-20
A	CN 108363701 A (DAERGUAN INFORMATION TECHNOLOGY (SHANGHAI) CO., LTD.) 03 August 2018 (2018-08-03) entire document	1-20
A	US 2015286629 A1 (MICROSOFT CORPORATION) 08 October 2015 (2015-10-08) 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

29 September 2021

Date of mailing of the international search report

12 October 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/CN2021/097543

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	112733911	A	30 April 2021	None	
CN	111611802	A	01 September 2020	None	
CN	110070183	A	30 July 2019	None	
CN	108363701	A	03 August 2018	None	
US	2015286629	A1	08 October 2015	US 9971763 B2	15 May 2018

国际检索报告

国际申请号

PCT/CN2021/097543

A. 主题的分类

G06K 9/62 (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)

G06K, G06N, 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))

CNTXT, CNABS, DWPI, SIPOABS, CNKI: 弱标注, 不完全标注, 局部标注, 少量标注, 较少标注, 命名实体识别, 实体识别, labeled, named entity recognition

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 112733911 A (平安科技深圳有限公司) 2021年 4月 30日 (2021 - 04 - 30) 说明书第[0001]-[0153]段	1-20
X	CN 111611802 A (苏州大学) 2020年 9月 1日 (2020 - 09 - 01) 说明书第[0050]-[0088]段	1、7、9、15
A	CN 110070183 A (中国科学院信息工程研究所) 2019年 7月 30日 (2019 - 07 - 30) 全文	1-20
A	CN 108363701 A (达而观信息科技上海有限公司) 2018年 8月 3日 (2018 - 08 - 03) 全文	1-20
A	US 2015286629 A1 (MICROSOFT CORPORATION) 2015年 10月 8日 (2015 - 10 - 08) 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

2021年 9月 29日

国际检索报告邮寄日期

2021年 10月 12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

刘瑛

电话号码 010-62412234

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2021/097543

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	112733911	A	2021年 4月 30日	无	
CN	111611802	A	2020年 9月 1日	无	
CN	110070183	A	2019年 7月 30日	无	
CN	108363701	A	2018年 8月 3日	无	
US	2015286629	A1	2015年 10月 8日	US 9971763 B2	2018年 5月 15日