

(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
2021 年 7 月 15 日 (15.07.2021)



(10) 国际公布号
WO 2021/139239 A1

(51) 国际专利分类号:
G06F 40/30 (2020.01) *G06F 16/36* (2019.01)

(21) 国际申请号: PCT/CN2020/118331

(22) 国际申请日: 2020 年 9 月 28 日 (28.09.2020)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202010738252.X 2020年7月28日 (28.07.2020) CN

安社区益田路 5033 号平安金融中心 23 楼, Guangdong 518000 (CN)。

(72) 发明人: 柴玲(CHAI, Ling); 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)。

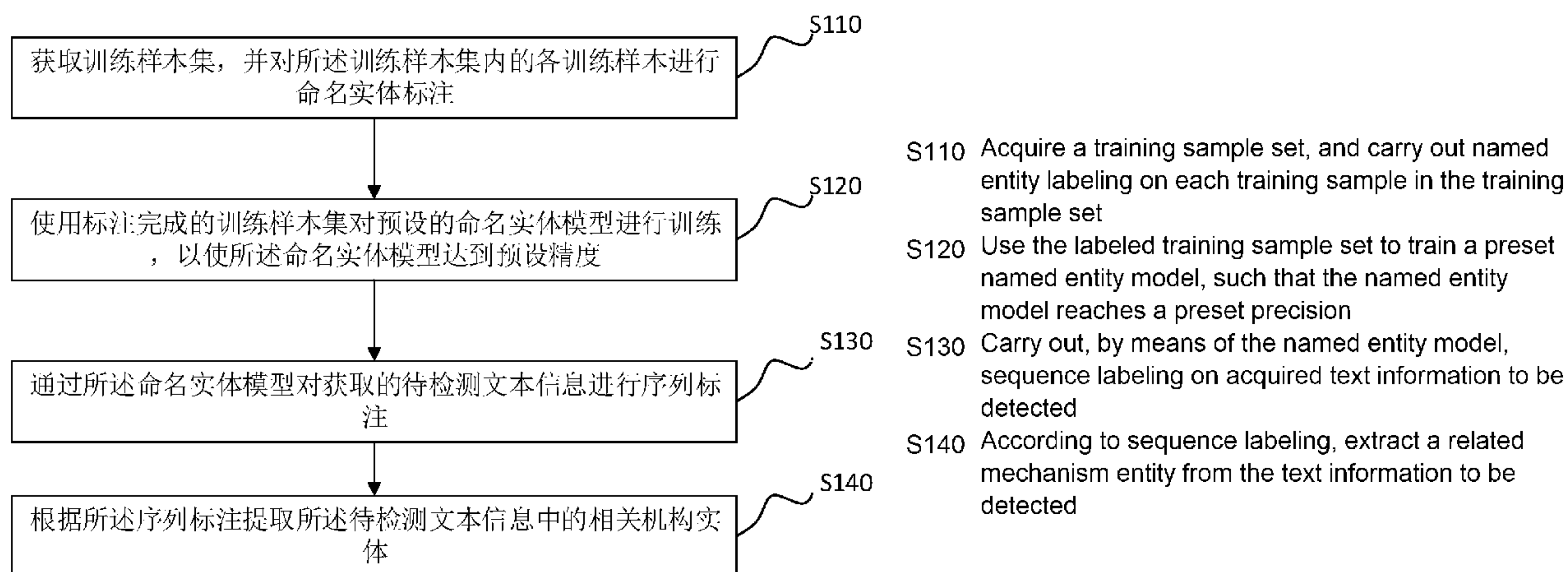
(74) 代理人: 广州三环专利商标代理有限公司 (SCIHEAD IP LAW FIRM); 中国广东省广州市越秀区先烈中路 80 号汇华商贸大厦 1508 室, Guangdong 510070 (CN)。

(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

(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,

(54) **Title:** MECHANISM ENTITY EXTRACTION METHOD, SYSTEM AND DEVICE BASED ON MULTIPLE TRAINING TARGETS

(54) 发明名称: 基于多训练目标的机构实体抽取方法、系统及装置



(57) **Abstract:** A mechanism entity extraction method based on multiple training targets. The method comprises: acquiring a training sample set, and carrying out named entity labeling on each training sample in the training sample set; using the labeled training sample set to train a preset named entity model, such that the named entity model reaches a preset precision; carrying out, by means of the named entity model, sequence labeling on acquired text information to be detected; and according to sequence labeling, extracting a related mechanism entity from the text information to be detected. The present invention further relates to blockchain technology. The training sample set is stored in a blockchain. The method can effectively solve the problems of the low efficiency and poor quality of an existing mechanism entity extraction method.

(57) **摘要:** 一种基于多训练目标的机构实体抽取方法, 包括: 获取训练样本集, 并对训练样本集内的各训练样本进行命名实体标注; 使用标注完成的训练样本集对预设的命名实体模型进行训练, 以使命名实体模型达到预设精度; 通过所述命名实体模型对获取的待检测文本信息进行序列标注; 根据序列标注提取待检测文本信息中的相关机构实体。还涉及区块链技术, 训练样本集存储于区块链中。该方法能够有效地解决现有的机构实体抽取方法效率低质量差的问题。

[见续页]

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))。
- 在修改权利要求的期限届满之前进行，在收到该修改后将重新公布(细则48.2(h))。
- 根据申请人的请求，在条约第21条(2)(a)所规定的期限届满之前进行。

基于多训练目标的机构实体抽取方法、系统及装置

本申请要求于 2020 年 7 月 28 日提交中国专利局、申请号为 202010738252.X，发明名称为“基于多训练目标的机构实体抽取方法、系统及装置”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。

5 技术领域

本申请涉及信息提取技术领域，尤其涉及一种基于多训练目标的机构实体抽取方法、系统、装置及存储介质。

背景技术

10 当前很多学术学者库如AMINER、ORCID等，提供学者信息，方便用户去追踪某一学者或其所在团队的研究方向与进展。比如，一些专家团队项目深耕医学科研领域，致力于搭建一个医学领域的专家库，构建完整的专家知识图谱。

15 然而，在专家知识图谱的构建过程中，建立专家与机构的关系网络是一个有价值且有难度的工作，原因在于专家是一个有自我选择能力的行为人，他是会在各机构之间进行流动；比如，专家A可能在机构A读博士、在机构B工作，期间又在机构C进修过。然而，目前的常见的学者库（如AMINER、ORCID等）一般仅提供学者的当前所在机构。实际上，发明人意识到，一个学者的完整的科研画像与他曾经待过的机构是息息相关的。

20 显然，通过人力去搜集数十万专家（以中国医疗相关学者为例）的教育经历、工作经历、进修经历中涉及的机构是不现实的。因此，发明人意识到可以考虑从互联网中获取到学者主页上大段的文本信息。所以该问题即变成如何使用人工智能的算法从糅杂的文本信息中提取与该学者相关的结构化的知识。

25 如某互联网中的某学者的相关简介字段为“1990年6月毕业后,到广州医学院第一附属医院从事肿瘤血液内科工作,2008年6月取得南方医科大学临床医学博士学位。2008年12月至2009年5月在中山大学肿瘤防治中心进修。”，则需要从中抽取出三个机构实体为“广州医学院第一附属医院”、“南方医科大学”以及“中山大学”，并且识别出三者分别属于“工作经历”、“教育经历”、“短期进修经历”。

基于以上问题，亟需一种高效且高质的机构实体抽取方法。

发明内容

30 本申请提供一种基于多训练目标的机构实体抽取方法、系统、电子装置以及计算机存储介质，其主要目的在于解决现有的机构实体抽取方法效率低质量差的问题。

为实现上述目的，本申请提供一种基于多训练目标的机构实体抽取方法，该方法包括如下步骤：

获取训练样本集，并对所述训练样本集内的各训练样本进行命名实体标注；

35 使用标注完成的训练样本集对预设的命名实体模型进行训练，以使所述命名实体模型达到预设精度；其中，所述命名实体模型包括第一主干路和第二主干路，所述第一主干路用于提取输入文本信息的第一向量特征集，所述第二主干路用于提取所述输入文本信息的第二向量特征集；并且，所述第二主干路还用于根据所述第一向量特征集和所述第二向量特征集对所述输入文本信息进行序列标注；

通过所述命名实体模型对获取的待检测文本信息进行序列标注；

根据所述序列标注提取所述待检测文本信息中的相关机构实体。

40 另外，本申请还提供一种基于多训练目标的机构实体抽取系统，所述系统包括：

样本标注单元，用于获取训练样本集并对所述训练样本集内的各训练样本进行命名实体标注；

模型训练单元，用于使用标注完成的训练样本集对预设的命名实体模型进行训练，以使所述命名实体模型达到预设精度；其中，所述命名实体模型包括第一主干路和第二主干

路，所述第一主干路用于提取输入文本信息的第一向量特征集，所述第二主干路用于提取所述输入文本信息的第二向量特征集；并且，所述第二主干路还用于根据所述第一向量特征集和所述第二向量特征集对所述输入文本信息进行序列标注；

模型应用单元，用于通过所述命名实体模型对获取的待检测文本信息进行序列标注；

5 机构实体提取单元，用于根据所述序列标注提取所述待检测文本信息中的相关机构实体。

另外，为实现上述目的，本申请还提供一种电子装置，该电子装置包括：存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的基于多训练目标的机构实体抽取程序，所述基于多训练目标的机构实体抽取程序被所述处理器执行时实现如下步骤：

10 获取训练样本集，并对所述训练样本集内的各训练样本进行命名实体标注；

使用标注完成的训练样本集对预设的命名实体模型进行训练，以使所述命名实体模型达到预设精度；其中，所述命名实体模型包括第一主干路和第二主干路，所述第一主干路用于提取输入文本信息的第一向量特征集，所述第二主干路用于提取所述输入文本信息的第二向量特征集；并且，所述第二主干路还用于根据所述第一向量特征集和所述第二向量特征集对所述输入文本信息进行序列标注；

15 通过所述命名实体模型对获取的待检测文本信息进行序列标注；

根据所述序列标注提取所述待检测文本信息中的相关机构实体。

另外，为实现上述目的，本申请还提供一种计算机可读存储介质，所述计算机可读存储介质中存储有基于多训练目标的机构实体抽取程序，所述基于多训练目标的机构实体抽取程序被处理器执行时，实现以下步骤：

20 获取训练样本集，并对所述训练样本集内的各训练样本进行命名实体标注；

使用标注完成的训练样本集对预设的命名实体模型进行训练，以使所述命名实体模型达到预设精度；其中，所述命名实体模型包括第一主干路和第二主干路，所述第一主干路用于提取输入文本信息的第一向量特征集，所述第二主干路用于提取所述输入文本信息的第二向量特征集；并且，所述第二主干路还用于根据所述第一向量特征集和所述第二向量特征集对所述输入文本信息进行序列标注；

25 通过所述命名实体模型对获取的待检测文本信息进行序列标注；

根据所述序列标注提取所述待检测文本信息中的相关机构实体。

本申请能够有效地避免误差传播，另外，本申请设计的多个训练目标的命名实体模型，强化了边界特征和语义特征的抽取，能够显著提高最终的预测精度，尤其是对边界的抓取，比传统的NER模型稳定得多。

附图说明

图1为根据本申请实施例的基于多训练目标的机构实体抽取方法的较佳实施例流程图；

图2为根据本申请实施例的电子装置的较佳实施例结构示意图；

35 图3为根据本申请实施例的基于多训练目标的机构实体抽取程序的内部逻辑示意图。

本申请目的的实现、功能特点及优点将结合实施例，参照附图做进一步说明。

具体实施方式

在下面的描述中，出于说明的目的，为了提供对一个或多个实施例的全面理解，阐述了许多具体细节。然而，很明显，也可以在没有这些具体细节的情况下实现这些实施例。

40 本申请的技术方案可应用于人工智能、区块链和/或大数据技术领域，涉及的数据如训练样本集等可存储于数据库中，或者可以存储于区块链中，比如通过区块链分布式存储，本申请不做限定。

在具体介绍本申请提供的基于多训练目标的机构实体抽取方法的实施例之前，需要说明的是，传统的实体抽取选用的方法主要有两种，一种是分阶段训练模型：首先训练命名

实体提取模型识别出所有机构实体，“南方医科大学”被标注为[B-ORG,I-ORG,I-ORG,I-ORG,E-ORG]，然后文本分类模型识别该机构实体所属经历类型为“工作经历(JOB)”、“教育经历(EDU)”还是“短期进修经历(STU)”。但这种解决方案很明显的缺陷是第一个模型的误差会传递到第二个模型并且放大误差。

5 另外一种常见的解决方案为训练端到端的命名实体提取模型，如 LSTM+CRF 等。对每个实体进行统一的标注（unified tag），“南方医科大学”直接被标注为[B-EDU,I-EDU,I-EDU,I-EDU,I-EDU,E-EDU]，“中山大学”直接标注为[B-STU,I-STU,I-STU,E-STU]，训练端到端的命名实体提取模型。这样可以避免分别训练的传统方案中的误差传播。

10 但是 LSTM+CRF 简单的命名实体抽取模型仍无法很好地解决简介机构细抽取这一场景中特有的两个问题：一：同一个实体在不同语境下标签不同：“上海第六人民医院”既是医生 A 的“教育经历”机构，又是医生 B 的“工作经历”，还是医生 C 的“教育经历”以及“工作经历”。所以对上下文场景的信息捕捉难度高于一般的命名实体提取问题。二是边界问题，为保证输入的结构化知识的统一，对机构的提取均保留到独立单位的粒度（大学、医院等），
15 比如“中山大学肿瘤防治中心”，我们希望最后的结果能识别到“中山大学”的级别同时忽略“肿瘤防治中心”，与此同时“北京肿瘤防治中心”又是一个独立的实体。显然，传统的端到端的命名实体提取模型做不到这一点，因此，亟需一种更加高效且高质的机构实体抽取方法。

以下将结合附图对本申请的具体实施例进行详细描述。

20 实施例 1

为了说明本申请提供的基于多训练目标的机构实体抽取方法，图 1 示出了根据本申请提供的基于多训练目标的机构实体抽取方法的流程。

如图 1 所示，本申请提供的基于多训练目标的机构实体抽取方法，包括：

S110：获取训练样本集，并对该训练样本集内的各训练样本进行命名实体标注。

25 需要说明的是，此处的样本即为一段包含机构实体的一段文字信息，例如，可以是入职简历中的一段话，也可以是网络中的学者主页上的一段文字信息。

具体地，在对该训练样本集内的各训练样本进行命名实体标注的过程中，本申请使用的命名实体标注方法为 BIO 标注方式，其中，B 用于标注机构实体的开头，I 用于标注机构实体本身，O 用于标注样本中与机构实体不相关的信息。

30 此外，为实现后期模型的多目标训练，本申请需要对训练样本集中的每一个样本进行多种类型的标注，至少包括四种类型，例如：Boundary-tag, End-tag, Type-tage, 以及 unified-tag 四类标签，不同类型的标签的标注方式不同，当然对应的标注功能也不同，Boundary-tag 类型主要用于标注样本中的机构实体边界；End-tag 类型主要用于标注样本中的机构实体的结束位置；Type-tage 类型主要用于标注机构实体的实体类型，比如，毕业院校、工作场所，
35 实习场所等等。unified-tag 类型为最终的目标标签，待四种类型的标注都完成后，将该样本保存至训练样本集。

另外，需要强调的是，为进一步保证上述该训练样本集内数据的私密和安全性，该训练样本集可以存储于区块链的节点中。

40 S120：使用标注完成的训练样本集对预设的命名实体模型进行训练，以使该命名实体模型达到预设精度；其中，该命名实体模型包括第一主干路和第二主干路，该第一主干路用于提取输入文本信息的第一向量特征集，该第二主干路用于提取该输入文本信息的第二向量特征集；并且，该第二主干路还用于根据该第一向量特征集和该第二向量特征集对该输入文本信息进行序列标注。

需要说明的是，命名实体模型为自行设计的一个新型的序列标注模型，该模型结合了

多个目标的训练环节；具体地，该命名实体模型包括两条主干路，该第一主干路内设置有第一神经网络模型 LSTM1，该第一主干路通过该第一神经网络模型 LSTM1 提取该输入文本信息（对应训练样本或后期的待检测文本信息）的第一向量特征集；该第二主干路内设置有第二神经网络模型 LSTM2，该第二主干路通过该第二神经网络模型 LSTM1 提取该输入文本信息的第二向量特征集。

该第一主干路分支为第一支路和第二支路，该第一支路内设置有第一预测分类器（predictor），该第一预测分类器用于根据该 Boundary-tag 标注类型标注出该第一向量特征集的实体边界；该第二支路内设置有第二预测分类器，该第二预测分类器用于根据该 End-tag 标注类型标注出该第一向量特征集的结束位置。

具体地，输入文本信息经 LSTM1 进行特征提取后会输出一组相应的第一向量特征集，记作 h_1 ，然后将该第一向量特征集 h_1 分别同时传给第一支路和第二支路，其中，第一支路对应第一预测分类器，用于根据文本信息的第一向量特征集 h_1 标注出其实体边界 $y^{\text{boundary_tag}}$ ，对应 Boundary-tag 标注；第二支路对应第二预测分类器，用于根据文本信息的第一向量特征集 h_1 标注出其结束位置 $y^{\text{end_tag}}$ ，对应 End-tag 标注。

此处，对于第二主干路而言，文本信息经 LSTM2 进行特征提取后会输出一组相应的第二向量特征集，记作 h_2 ，并且，在该第二主干路提取完该第二向量特征集后，该第二主干路分支为第三支路和最终输出支路；其中，该第三支路内设置有第三预测分类器，该第三预测分类器用于根据该 Type-tage 标注类型标注出该第一向量特征集的实体类型；该最终输出支路内设置有总预测分类器，该总预测分类器用于根据该第一向量特征集、该第二向量特征集以及该 unified-tag 标注类型标注出该输入文本信息的最终标注。

具体地，将该第二向量特征集 h_2 分别同时传给第三支路和最终输出支路，其中，第三支路对应第三预测分类器（predictor），用于根据输入文本信息的第二向量特征集 h_2 标注出其实体类型 $y^{\text{type_tage}}$ ，对应 Type-tage；最终输出支路对应总预测分类器（SC-BG），用于根据输入文本信息的第一向量特征集 h_1 以及第二向量特征集 h_2 对输入文本信息进行最终标注 $y^{\text{unified_tag}}$ ，对应 unified-tag。

需要说明的是，LSTM（包括 LSTM1 和 LSTM2）为一种现有的常用的神经网络模型，其具体结构为现有技术，在此不再赘述。通过使用该类神经网络模型可以获取输入文本信息的一组向量特征（ h_1 或 h_2 ），当然 LSTM 需要搭配预测分类器使用，当使用训练样本集内训练样本对 LSTM 和与其相连的预测分类器（第一预测分类器、第一预测分类器、第一预测分类器以及总预测分类器）训练完毕后，LSTM 输出的特征向量即会与各预测分类器之间产生所需关联，该关联可用模型训练参数 W^1 表示， W^1 达到预设精度时，LSTM 输出的特征向量为所需的特征向量。

需要进一步说明的是，当使用训练样本集内的所有训练样本对上述模型训练完毕后，该命名实体模型的精度即可到达所需的预设精度，此时，命名实体模型提取的向量特征 h_1 与 Boundary-tag 和 End-tage 产生所需的联系，命名实体模型提取的向量特征 h_2 与 Type-tage 产生所需的联系。当使用向量特征 h_1 和 h_2 识别待检测文本信息时，即可直接应用到 boundary-tag, End-tag 以及 Type-tage 的标记特性，从而通过 $y^{\text{boundary_tag}}$ ， $y^{\text{type_tage}}$ ， $y^{\text{unified_tag}}$ 辅助提高 unified_tag 的准确率。

更为具体地，该第一预测分类器、该第二预测分类器以及该第三预测分类器内均设置有激活函数，该第一向量特征集或该第二向量特征集通过该激活函数后实现对该第一向量特征集或该第二向量特征集的标注；其中，该激活函数的计算过程如下：

$$z_t^b = \text{Softmax}(W^1 h_t^0)$$

$$y_t^1 = \arg \max_y z_t^b$$

其中, W^1 为该命名实体模型中需要训练的参数, 与预测分类器的标注类型相关联; h_t^0 代指第一向量特征集或第二向量特征集, y_t^1 为输出的标注结果; Softmax 函数为归一化函数, 是用于将 $W^1 h_t^0$ 的值映射到 (0,1) 区间内:

$$f(i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

其中, $[x_1, x_2, \dots, x_i, x_n]$ 为一个数组, 代指 $W^1 h_t^1$, $f(i)$ 为第 i 个元素的 softmax 值。

5 为便于理解上述命名实体模型的数据处理流程, 下面以“毕业于上海交通大学医学院”为具体的输入文本信息的示例对数据在模型中的流动与计算过程进行详细介绍。

具体地, 命名实体模型的前端第一主干路中的 LSTM 层记作 LSTM1, 输出第一向量特征集, 记作 h_1 , 在 t 时刻的输出向量记为 h_t^1 , 经过 Softmax 激活函数之后用来预测 boundary-tag 标注, 例如, “毕业于上海交通大学医学院”对应的输出应该为“毕(O)业(O)于(O)上(B)海(I)交(I)通(I)大(I)学(I)医(O)学(O)院(O)”。输出记作 $y^{\text{boundary_tag}}$ 。

10 计算过程如下:

$$z_t^b = \text{Softmax}(W^b h_t^1)$$

$$y_t^{\text{boundary_tag}} = \arg \max_y z_t^b$$

其中, W^b 为模型中第一预测分类器中需要训练的参数。其中 Softmax 函数为归一化函数, 是将 h_t^1 的值映射到 (0,1) 区间内:

$$f(i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

其中, $[x_1, x_2, \dots, x_i, x_n]$ 为一个数组, 代指 $W^b h_t^1$, $f(i)$ 为第 i 个元素的 softmax 值。

15 与此同时, 通过第二预测分类器来预测 end-tag 标注, 即预测对应文本信息为 0 (非实体结束位置), 还是 1 (实体结束位置); 例如, “毕业于上海交通大学医学院”的对应输出为“毕(O)业(O)于(O)上(O)海(O)交(O)通(O)大(O)学(1)医(O)学(O)院(O)”。输出记作为 $y^{\text{end_tag}}$ 。

计算过程如下:

$$z_t^e = \text{Softmax}(W^e h_t^1)$$

$$y_t^{\text{end_tag}} = \arg \max_y z_t^e$$

20 其中, W^e 为模型中第二预测分类器中需要训练的参数。

通过使用训练样本集不断训练优化模型中的第一主干路 (第一预测分类器和第二预测

分类器), 第一条主干路中的 LSTM1 的输出 h1 即可学习到的 boundary-tag 和 end-tag 两种分类特性。

然而, 在传统的 CRF 中, 对于为本信息中的每一个名词, 都是一视同仁的, 由于很多机构都是以“学院”结尾, 所以在 CRF 的预测中“上海交通大学医学院”常常会标记为一个整体实体, 但是, 我们希望最后的结果能识别到“上海交通大学”的这一粒度同时忽略“医学院”这个更低级别的实体。所以需要加强对实体的边界的识别, 而本申请提供的命名实体模型的第一主干路相当于增加了实体的边界约束, 可实现相应的预测功能。

此外, 命名实体模型的前端的第二主干路中的 LSTM 层记作 LSTM2, 在输入文本信息输入模型后, 输出第二向量特征集, 记作 h2, 在 t 时刻的输出向量记为 h_t^2 , 通过 softmax 函数来预测签 type_tag 标注, 即预测对应分类类型为 JOB (工作单位)、EDU (教育经历的单位) 等; 例如“毕业于上海交通大学医学院”的对应输出应该为“毕(O)业(O)于(O)上(EDU)海(EDU)交(EDU)通(EDU)大(EDU)学(EDU)医(O)学(O)院(O)”。记作为 y^{type_tag} 。

计算过程如下:

$$z_t^t = Softmax(W^t h_t^2)$$

$$y_t^{type_tag} = \arg \max_y z_t^t$$

其中, W^t 模型中第三预测分类器中需要训练的参数。

另外, 对于主要预测部分 (对应总预测分类器 (SC-BG)), 引入预测组件 BG (boundary guide) 和 SC (sentiment consistency), 进一步整合第一向量特征集 h1 和第二向量特征集 h2 的数据以及内部隐藏特性, 得到最终的预测结果, 对应 unified-tag 分类标签, 例如: “毕业于上海交通大学医学院”的对应输出为“毕(O)业(O)于(O)上(B-EDU)海(I-EDU)交(I-EDU)通(I-EDU)大(I-EDU)学(I-EDU)医(O)学(O)院(O)”, 此处即为最终的目标标签。最终的输出记作为 $y^{unified_tag}$ 。

具体地, 该总预测分类器内设置有第一预测组件 SC 和第二预测组件 BG; 其中, 第一预测组件 SC 用于对该第二向量特征集进行优化, 以增强该第二向量特征集中当前向量特征与前一时刻特征之间的联系;

该第二预测组件 BG 用于根据优化后的第二向量特征集、该第一向量特征集以及该 unified-tag 标注类型标注出该输入文本信息的最终标注。

其中, 对于 SC 组件, 输入为 h2, 输出为一组向量特征集, 记为 h3, 在 t 时刻的输出向量记为 h_t^3 , 其中的计算过程为:

$$g_t = sigmoid(W^g h_t^2 + b^g)$$

$$h_t^3 = g_t \odot h_t^2 + (1 - g_t) \odot h_{t-1}^3$$

其中, sigmoid 函数如下:

$$f(x) = \frac{1}{1 + \exp(-x)}$$

需要说明的是, 对于 $\odot$ 运算符, 其为预先设定的线性运算符, 例如, $A \odot B = 3A + 2B$, 此处只要满足线性关系即可。

对于 BG 组件, 输入 h1 为和 h3, 输出为最终的标签 (unified-tag), 记作 $y^{unified_tag}$, 预测过程如下, BG 中含有一个 boundary-tag 到 unified-tag 的转移矩阵 W^u ,

$$W_{i,j}^{tr} = \begin{cases} \frac{1}{|B_i|} & \text{if } j \in B_i \\ 0 & \text{otherwise} \end{cases}$$

其中, B_i 为 unified-tag 的标签集合 $\{B-EUD, I-EDU, B-STU, I-STU, O, \dots\}$, $|B_i|$ 为该集合的大小。

然后通过转移矩阵, 原本的 z^b 变化为:

$$z_t^{u'} = W^{trT} z_t^b$$

其中, $z^{u'}$ 可以看作是由边界信息预测的最终标签, z^b 为第一预测分类器中的中间参数 (可参照对于第一预测分类器的具体实施例), 用 z^b 自身计算得到 $z^{u'}$ 这一标签的权重 a^t

$$c_t = (z_t^b)^T z_t^b$$

$$a_t = \epsilon c_t$$

其中, ϵ 为预测的超参数, 最终标签的计算过程为:

$$z_t^u = a_t z_t^{u'} + (1 - a_t) \text{Softmax}(W^3 h_3^t)$$

$$y_t^{\text{unified_tag}} = \arg \max_y z_t^u$$

需要说明的是, 随着样本训练集对命名实体模型的训练, W^1 (包括 $W^b W^e W^t W^{tr}$) 也会随之进行变化, 并越来越接近最优值, 当命名实体模型训练完毕后, W^1 一般会在最优值附近, 此时即可使用该命名实体模型对待抽取文本信息进行机构实体信息提取。

S130: 获取待检测文本信息, 并通过该命名实体模型对该待检测文本信息进行序列标注。

具体地, 可以用互联网或数据库中获取与行为人相关的待检测文本信息, 如个人简历信息、个人主页信息等。

需要说明的是, 待检测文本信息经命名实体模型标注后, 会在四个输出端输出相应的标注序列, 包括 $y^{\text{boundary_tag}}$, $y^{\text{end_tag}}$, $y^{\text{type_tag}}$ 和 $y^{\text{unified_tag}}$, 由于 $y^{\text{unified_tag}}$ 中已经包含了 $y^{\text{boundary_tag}}$, $y^{\text{end_tag}}$, $y^{\text{type_tag}}$ 的相关序列特性信息, 因此, 在实际应用中, 只需要获取待检测文本信息的 $y^{\text{unified_tag}}$ 序列标注即可。

S140: 根据该序列标注提取该待检测文本信息中的相关机构实体。

具体地, 根据该 $y^{\text{unified_tag}}$ 序列标注, 提取待检测文本信息中行为人的相关机构实体。

需要说明的是, 由于 $y^{\text{unified_tag}}$ 序列标注中包含了 $y^{\text{end_tag}}$ 序列标注的相关特性, 因此, 可以根据 $y^{\text{unified_tag}}$ 序列标注精准地确定所需机构实体的结束位置, 避免机构实体定位不准确的问题。此外, 由于 $y^{\text{unified_tag}}$ 序列标注中 $y^{\text{type_tag}}$ 序列标注的相关特性, 因此, 可以根据 $y^{\text{unified_tag}}$ 序列标注精准地确定机构实体的类别为“工作经历”、“教育经历”还是“短期进修经历”。

当然, 还可以通过修改训练目标进一步拓展到更细的机构抽取, 比如抽取二级机构 (“上海交通大学医学院”中的“医学院”), unified-tag 的标签为“上(B-EDU)海(I-EDU)交(I-EDU)通(I-EDU)大(I-EDU)学(I-EDU)医(I-EDU)学(I-EDU)院(I-EDU)”, boundary-tag 标注为“上(B)海(I)交(I)通(I)大(I)学(I)医(I)学(I)院(I)”, 与上述过程同理, 结束标签 1 应该打在“院”的位置。模型的框架无需改动, 此时, 即可实现二级机构的抽取。

通过上述技术方案的表述可知，本申请提出的基于多训练目标的机构实体抽取方法，通过设计一个多目标训练的命名实体模型，与传统的分别训练命名实体提取模型和文本分类模型相比，能够有效地避免误差传播，另外，针对常规的命名实体提取模型如 LSTM+CRF 不能很好地判断同一实体的不同类型以及边界精度识别不准的问题，本申请设计的多个训练目标的命名实体模型，强化了边界特征和语义特征的抽取，能够显著提高最终的预测精度，尤其是对边界的抓取，比传统的 NER 模型稳定得多。

应理解，上述实施例中各步骤的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不对本申请实施例的实施过程构成任何限定。

实施例 2

与上述方法相对应，本申请还提供一种基于多训练目标的机构实体抽取系统，该系统包括：

样本标注单元，用于获取训练样本集并对该训练样本集内的各训练样本进行命名实体标注；

模型训练单元，用于使用标注完成的训练样本集对预设的命名实体模型进行训练，以使该命名实体模型达到预设精度；其中，该命名实体模型包括第一主干路和第二主干路，该第一主干路用于提取输入文本信息的第一向量特征集，该第二主干路用于提取该输入文本信息的第二向量特征集；并且，该第二主干路还用于根据该第一向量特征集和该第二向量特征集对该输入文本信息进行序列标注；

模型应用单元，用于获取待检测文本信息，并通过该命名实体模型对该待检测文本信息进行序列标注；

机构实体提取单元，用于根据该序列标注提取该待检测文本信息中的相关机构实体。

实施例 3

本申请还提供一种电子装置 70。参照图 2 所示，该图为本申请提供的电子装置 70 的较佳实施例结构示意图。

在本实施例中，电子装置 70 可以是服务器、智能手机、平板电脑、便携计算机、桌上型计算机等具有运算功能的终端设备。

该电子装置 70 包括：处理器 71 以及存储器 72。

存储器 72 包括至少一种类型的可读存储介质。至少一种类型的可读存储介质可为如闪存、硬盘、多媒体卡、卡型存储器等的非易失性存储介质。在一些实施例中，可读存储介质可以是该电子装置 70 的内部存储单元，例如该电子装置 70 的硬盘。在另一些实施例中，可读存储介质也可以是电子装置 1 的外部存储器，例如电子装置 70 上配备的插接式硬盘，智能存储卡(Smart Media Card, SMC)，安全数字(Secure Digital, SD)卡，闪存卡(Flash Card)等。

在本实施例中，存储器 72 的可读存储介质通常用于存储安装于电子装置 70 的基于多训练目标的机构实体抽取程序 73。存储器 72 还可以用于暂时地存储已经输出或者将要输出的数据。

处理器 72 在一些实施例中可以是一中央处理器(Central Processing Unit, CPU)，微处理器或其他数据处理芯片，用于运行存储器 72 中存储的程序代码或处理数据，例如基于多训练目标的机构实体抽取程序 73 等。

在一些实施例中，电子装置 70 为智能手机、平板电脑、便携计算机等的终端设备。在其他实施例中，电子装置 70 可以为服务器。

图 2 仅示出了具有组件 71-73 的电子装置 70，但是应理解的是，并不要求实施所有示出的组件，可以替代的实施更多或者更少的组件。

可选地，该电子装置 70 还可以包括用户接口，用户接口可以包括输入单元比如键盘

(Keyboard)、语音输入装置比如麦克风(microphone)等具有语音识别功能的设备、语音输出装置比如音响、耳机等,可选地用户接口还可以包括标准的有线接口、无线接口。

可选地,该电子装置 70 还可以包括显示器,显示器也可以称为显示屏或显示单元。在一些实施例中可以是 LED 显示器、液晶显示器、触控式液晶显示器以及有机发光二极管(Organic Light-Emitting Diode, OLED)触摸器等。显示器用于显示在电子装置 70 中处理的信息以及用于显示可视化的用户界面。

可选地,该电子装置 70 还可以包括触摸传感器。触摸传感器所提供的供用户进行触摸操作的区域称为触控区域。此外,这里的触摸传感器可以为电阻式触摸传感器、电容式触摸传感器等。而且,触摸传感器不仅包括接触式的触摸传感器,也可包括接近式的触摸传感器等。此外,触摸传感器可以为单个传感器,也可以为例如阵列布置的多个传感器。

此外,该电子装置 70 的显示器的面积可以与触摸传感器的面积相同,也可以不同。可选地,将显示器与触摸传感器层叠设置,以形成触摸显示屏。该装置基于触摸显示屏侦测用户触发的触控操作。

可选地,该电子装置 70 还可以包括射频(Radio Frequency, RF)电路,传感器、音频电路等等,在此不再赘述。

在图 2 所示的装置实施例中,作为一种计算机存储介质的存储器 72 中可以包括操作系统、以及基于多训练目标的机构实体抽取程序 73;处理器 71 执行存储器 72 中存储基于多训练目标的机构实体抽取程序 73 时实现如下步骤:

获取训练样本集,并对该训练样本集内的各训练样本进行命名实体标注;

使用标注完成的训练样本集对预设的命名实体模型进行训练,以使该命名实体模型达到预设精度;其中,该命名实体模型包括第一主干路和第二主干路,该第一主干路用于提取输入文本信息的第一向量特征集,该第二主干路用于提取该输入文本信息的第二向量特征集;并且,该第二主干路还用于根据该第一向量特征集和该第二向量特征集对该输入文本信息进行序列标注;

获取待检测文本信息,并通过该命名实体模型对该待检测文本信息进行序列标注;

根据该序列标注提取该待检测文本信息中的相关机构实体。

在该实施例中,图 3 为根据本申请实施例的基于多训练目标的机构实体抽取程序的内部逻辑示意图,如图 3 所示,基于多训练目标的机构实体抽取程序 73 还可以被分割为一个或者多个模块,一个或者多个模块被存储于存储器 72 中,并由处理器 71 执行,以完成本申请。本申请所称的模块是指能够完成特定功能的一系列计算机程序指令段。参照图 3 所示,为图 2 中基于多训练目标的机构实体抽取程序 73 较佳实施例的程序模块图。基于多训练目标的机构实体抽取程序 73 可以被分割为:样本标注模块 74、模型训练模块 75、模型应用模块 76 以及机构实体提取模块 77。模块 74-76 所实现的功能或操作步骤均与上文类似,此处不再详述,示例性地,例如,其中:

样本标注模块 74,用于获取训练样本集并对该训练样本集内的各训练样本进行命名实体标注;

模型训练模块 75,用于使用标注完成的训练样本集对预设的命名实体模型进行训练,以使该命名实体模型达到预设精度;其中,该命名实体模型包括第一主干路和第二主干路,该第一主干路用于提取输入文本信息的第一向量特征集,该第二主干路用于提取该输入文本信息的第二向量特征集;并且,该第二主干路还用于根据该第一向量特征集和该第二向量特征集对该输入文本信息进行序列标注;

模型应用模块 76,用于获取待检测文本信息,并通过该命名实体模型对该待检测文本信息进行序列标注;

机构实体提取模块 77,用于根据该序列标注提取该待检测文本信息中的相关机构实体。

实施例 4

本申请还提供一种计算机可读存储介质，计算机可读存储介质中存储有基于多训练目标的机构实体抽取程序 73，基于多训练目标的机构实体抽取程序 73 被处理器执行时实现如下操作：

5 获取训练样本集，并对该训练样本集内的各训练样本进行命名实体标注；

10 使用标注完成的训练样本集对预设的命名实体模型进行训练，以使该命名实体模型达到预设精度；其中，该命名实体模型包括第一主干路和第二主干路，该第一主干路用于提取输入文本信息的第一向量特征集，该第二主干路用于提取该输入文本信息的第二向量特征集；并且，该第二主干路还用于根据该第一向量特征集和该第二向量特征集对该输入文本信息进行序列标注；

获取待检测文本信息，并通过该命名实体模型对该待检测文本信息进行序列标注；

根据该序列标注提取该待检测文本信息中的相关机构实体。

本申请提供的计算机可读存储介质的具体实施方式与上述基于多训练目标的机构实体抽取方法、电子装置的具体实施方式大致相同，在此不再赘述。

15 可选的，该计算机可读存储介质可以是非易失性的，也可以是易失性的。

20 需要说明的是，本申请所指区块链是分布式数据存储、点对点传输、共识机制、加密算法等计算机技术的新型应用模式。区块链(Blockchain)，本质上是一个去中心化的数据库，是一串使用密码学方法相关联产生的数据块，每一个数据块中包含了一批次网络交易的信息，用于验证其信息的有效性(防伪)和生成下一个区块。区块链可以包括区块链底层平台、平台产品服务层以及应用服务层等。

25 需要进一步说明的是，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。

30 上述本申请实施例序号仅仅为了描述，不代表实施例的优劣。通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到上述实施例方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在如上的一个存储介质(如 ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，或者网络设备等)执行本申请各个实施例的方法。

35 以上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

1、一种基于多训练目标的机构实体抽取方法，应用于电子装置，其中，所述方法包括：
获取训练样本集，并对所述训练样本集内的各训练样本进行命名实体标注；

5 使用标注完成的训练样本集对预设的命名实体模型进行训练，以使所述命名实体模型达到预设精度；其中，所述命名实体模型包括第一主干路和第二主干路，所述第一主干路用于提取输入文本信息的第一向量特征集，所述第二主干路用于提取所述输入文本信息的第二向量特征集；并且，所述第二主干路还用于根据所述第一向量特征集和所述第二向量特征集对所述输入文本信息进行序列标注；

通过所述命名实体模型对获取的待检测文本信息进行序列标注；

10 根据所述序列标注提取所述待检测文本信息中的相关机构实体。

2、根据权利要求1所述的基于多训练目标的机构实体抽取方法，其中，

所述训练样本集存储于区块链中；并且，在对所述训练样本集内的各训练样本进行命名实体标注的过程中，使用 BIO 的标注方式进行标注；其中，

15 B 用于标注机构实体的开头，I 用于标注机构实体本身，O 用于标注所述训练样本中与机构实体不相关的信息。

3、根据权利要求2所述的基于多训练目标的机构实体抽取方法，其中，所述在对所述训练样本集内的各训练样本进行命名实体标注的过程中，使用的标注类型包括：Boundary-tag、End-tag、Type-tage 以及 unified-tag；其中，

20 Boundary-tag 类型用于标注出所述训练样本中的机构实体的边界，End-tag 类型用于标注出所述训练样本中的机构实体的结束位置，Type-tage 类型用于标注出训练样本中的机构实体的实体类型，unified-tag 类型用于作为最终的目标标签。

4、根据权利要求3所述的基于多训练目标的机构实体抽取方法，其中，在所述第一主干路提取完所述第一向量特征集后，所述第一主干路分支为第一支路和第二支路；在所述第二主干路提取完所述第二向量特征集后，所述第二主干路分支为第三支路和最终输出支路；其中，

25 在所述第一支路内设置有第一预测分类器，所述第一预测分类器用于根据所述 Boundary-tag 标注类型标注出所述第一向量特征集的实体边界；在所述第二支路内设置有第二预测分类器，所述第二预测分类器用于根据所述 End-tag 标注类型标注出所述第一向量特征集的结束位置；

30 在所述第三支路内设置有第三预测分类器，所述第三预测分类器用于根据所述 Type-tage 标注类型标注出所述第一向量特征集的实体类型；在所述最终输出支路内设置有总预测分类器，所述总预测分类器用于根据所述第一向量特征集、所述第二向量特征集以及所述 unified-tag 标注类型标注出所述输入文本信息的最终标注。

5、根据权利要求4所述的基于多训练目标的机构实体抽取方法，其中，

35 在所述第一主干路内设置有第一神经网络模型 LSTM1，所述第一主干路通过所述第一神经网络模型 LSTM1 提取所述输入文本信息的第一向量特征集；

在所述第二主干路内设置有第二神经网络模型 LSTM2，所述第二主干路通过所述第二神经网络模型 LSTM1 提取所述输入文本信息的第二向量特征集。

6、根据权利要求5所述的基于多训练目标的机构实体抽取方法，其中，

40 所述第一预测分类器、所述第二预测分类器以及所述第三预测分类器内均设置有激活函数，所述第一向量特征集或所述第二向量特征集通过所述激活函数后实现对所述第一向量特征集或所述第二向量特征集的标注；其中，所述激活函数的计算过程如下：

$$z_t^b = \text{Softmax}(W^1 h_t^0)$$

$$y_t^1 = \arg \max_y z_t^b$$

其中, W^1 为所述命名实体模型中需要训练的参数, 与预测分类器的标注类型相关联; h_t^0 代指第一向量特征集或第二向量特征集, y_t^1 为输出的标注结果; Softmax 函数为归一化函数, 是用于将 $W^1 h_t^0$ 的值映射到 (0,1) 区间内:

$$f(i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

其中, $[x_1, x_2, \dots, x_i, x_n]$ 为一个数组, 代指 $W^1 h_t^1$, $f(i)$ 为第 i 个元素的 softmax 值。

- 5 7、根据权利要求 6 所述的基于多训练目标的机构实体抽取方法, 其中,

在所述总预测分类器内设置有第一预测组件 SC 和第二预测组件 BG; 其中, 第一预测组件 SC 用于对所述第二向量特征集进行优化, 以增强所述第二向量特征集中当前向量特征与前一时刻向量特征之间的联系;

所述第二预测组件 BG 用于根据优化后的第二向量特征集、所述第一向量特征集以及
- 10 所述 unified-tag 标注类型标注出所述输入文本信息的最终标注。
- 8、一种基于多训练目标的机构实体抽取系统, 其中, 所述系统包括:

样本标注单元, 用于获取训练样本集并对所述训练样本集内的各训练样本进行命名实体标注;

模型训练单元, 用于使用标注完成的训练样本集对预设的命名实体模型进行训练, 以
- 15 使所述命名实体模型达到预设精度; 其中, 所述命名实体模型包括第一主干路和第二主干路, 所述第一主干路用于提取输入文本信息的第一向量特征集, 所述第二主干路用于提取所述输入文本信息的第二向量特征集; 并且, 所述第二主干路还用于根据所述第一向量特征集和所述第二向量特征集对所述输入文本信息进行序列标注;

模型应用单元, 用于通过所述命名实体模型对获取的待检测文本信息进行序列标注;
- 20 机构实体提取单元, 用于根据所述序列标注提取所述待检测文本信息中的相关机构实体。
- 9、一种电子装置, 其中, 所述电子装置包括: 存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的基于多训练目标的机构实体抽取程序, 所述基于多训练目标的机构实体抽取程序被所述处理器执行时实现如下步骤:

获取训练样本集, 并对所述训练样本集内的各训练样本进行命名实体标注;

使用标注完成的训练样本集对预设的命名实体模型进行训练, 以使所述命名实体模型达到预设精度; 其中, 所述命名实体模型包括第一主干路和第二主干路, 所述第一主干路用于提取输入文本信息的第一向量特征集, 所述第二主干路用于提取所述输入文本信息的第二向量特征集; 并且, 所述第二主干路还用于根据所述第一向量特征集和所述第二向量特征集对所述输入文本信息进行序列标注;
- 25 通过所述命名实体模型对获取的待检测文本信息进行序列标注;
- 根据所述序列标注提取所述待检测文本信息中的相关机构实体。
- 10、根据权利要求 9 所述的电子装置, 其中,

所述训练样本集存储于区块链中; 并且, 在对所述训练样本集内的各训练样本进行命名
- 30 实体标注的过程中, 使用 BIO 的标注方式进行标注; 其中,
- B 用于标注机构实体的开头, I 用于标注机构实体本身, O 用于标注所述训练样本中与机构实体不相关的信息;

其中，所述在对所述训练样本集内的各训练样本进行命名实体标注的过程中，使用的标注类型包括：Boundary-tag、End-tag、Type-tage 以及 unified-tag；其中，

Boundary-tag 类型用于标注出所述训练样本中的机构实体的边界，End-tag 类型用于标注出所述训练样本中的机构实体的结束位置，Type-tage 类型用于标注出训练样本中的机构实体的实体类型，unified-tag 类型用于作为最终的目标标签。

11、根据权利要求 10 所述的电子装置，其中，在所述第一主干路提取完所述第一向量特征集后，所述第一主干路分支为第一支路和第二支路；在所述第二主干路提取完所述第二向量特征集后，所述第二主干路分支为第三支路和最终输出支路；其中，

在所述第一支路内设置有第一预测分类器，所述第一预测分类器用于根据所述 Boundary-tag 标注类型标注出所述第一向量特征集的实体边界；在所述第二支路内设置有第二预测分类器，所述第二预测分类器用于根据所述 End-tag 标注类型标注出所述第一向量特征集的结束位置；

在所述第三支路内设置有第三预测分类器，所述第三预测分类器用于根据所述 Type-tage 标注类型标注出所述第一向量特征集的实体类型；在所述最终输出支路内设置有总预测分类器，所述总预测分类器用于根据所述第一向量特征集、所述第二向量特征集以及所述 unified-tag 标注类型标注出所述输入文本信息的最终标注。

12、根据权利要求 11 所述的电子装置，其中，

在所述第一主干路内设置有第一神经网络模型 LSTM1，所述第一主干路通过所述第一神经网络模型 LSTM1 提取所述输入文本信息的第一向量特征集；

在所述第二主干路内设置有第二神经网络模型 LSTM2，所述第二主干路通过所述第二神经网络模型 LSTM2 提取所述输入文本信息的第二向量特征集。

13、根据权利要求 12 所述的电子装置，其中，

所述第一预测分类器、所述第二预测分类器以及所述第三预测分类器内均设置有激活函数，所述第一向量特征集或所述第二向量特征集通过所述激活函数后实现对所述第一向量特征集或所述第二向量特征集的标注；其中，所述激活函数的计算过程如下：

$$z_t^b = \text{Softmax}(W^1 h_t^0)$$

$$y_t^1 = \arg \max_y z_t^b$$

其中， W^1 为所述命名实体模型中需要训练的参数，与预测分类器的标注类型相关联； h_t^0 代指第一向量特征集或第二向量特征集， y_t^1 为输出的标注结果；Softmax 函数为归一化函数，是用于将 $W^1 h_t^0$ 的值映射到 (0,1) 区间内：

$$f(i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

其中， $[x_1, x_2, \dots, x_i, x_n]$ 为一个数组，代指 $W^1 h_t^1$ ， $f(i)$ 为第 i 个元素的 softmax 值。

14、根据权利要求 13 所述的电子装置，其中，

在所述总预测分类器内设置有第一预测组件 SC 和第二预测组件 BG；其中，第一预测组件 SC 用于对所述第二向量特征集进行优化，以增强所述第二向量特征集中当前向量特征与前一时刻向量特征之间的联系；

所述第二预测组件 BG 用于根据优化后的第二向量特征集、所述第一向量特征集以及

所述 unified-tag 标注类型标注出所述输入文本信息的最终标注。

15、一种计算机可读存储介质，其中，所述计算机可读存储介质中存储有基于多训练目标的机构实体抽取程序，所述基于多训练目标的机构实体抽取程序被处理器执行时，实现如以下步骤：

- 5 获取训练样本集，并对所述训练样本集内的各训练样本进行命名实体标注；
使用标注完成的训练样本集对预设的命名实体模型进行训练，以使所述命名实体模型达到预设精度；其中，所述命名实体模型包括第一主干路和第二主干路，所述第一主干路用于提取输入文本信息的第一向量特征集，所述第二主干路用于提取所述输入文本信息的第二向量特征集；并且，所述第二主干路还用于根据所述第一向量特征集和所述第二向量特征集对所述输入文本信息进行序列标注；
- 10 通过所述命名实体模型对获取的待检测文本信息进行序列标注；
根据所述序列标注提取所述待检测文本信息中的相关机构实体。
- 16、根据权利要求 15 所述的计算机可读存储介质，其中，
所述训练样本集存储于区块链中；并且，在对所述训练样本集内的各训练样本进行命名实体标注的过程中，使用 BIO 的标注方式进行标注；其中，
15 B 用于标注机构实体的开头，I 用于标注机构实体本身，O 用于标注所述训练样本中与机构实体不相关的信息；
其中，所述在对所述训练样本集内的各训练样本进行命名实体标注的过程中，使用的标注类型包括：Boundary-tag、End-tag、Type-tage 以及 unified-tag；其中，
- 20 Boundary-tag 类型用于标注出所述训练样本中的机构实体的边界，End-tag 类型用于标注出所述训练样本中的机构实体的结束位置，Type-tage 类型用于标注出训练样本中的机构实体的实体类型，unified-tag 类型用于作为最终的目标标签。
- 17、根据权利要求 16 所述的计算机可读存储介质，其中，在所述第一主干路提取完所述第一向量特征集后，所述第一主干路分支为第一支路和第二支路；在所述第二主干路提取完所述第二向量特征集后，所述第二主干路分支为第三支路和最终输出支路；其中，
- 25 在所述第一支路内设置有第一预测分类器，所述第一预测分类器用于根据所述 Boundary-tag 标注类型标注出所述第一向量特征集的实体边界；在所述第二支路内设置有第二预测分类器，所述第二预测分类器用于根据所述 End-tag 标注类型标注出所述第一向量特征集的结束位置；
- 30 在所述第三支路内设置有第三预测分类器，所述第三预测分类器用于根据所述 Type-tage 标注类型标注出所述第一向量特征集的实体类型；在所述最终输出支路内设置有总预测分类器，所述总预测分类器用于根据所述第一向量特征集、所述第二向量特征集以及所述 unified-tag 标注类型标注出所述输入文本信息的最终标注。
- 18、根据权利要求 17 所述的计算机可读存储介质，其中，
- 35 在所述第一主干路内设置有第一神经网络模型 LSTM1，所述第一主干路通过所述第一神经网络模型 LSTM1 提取所述输入文本信息的第一向量特征集；
在所述第二主干路内设置有第二神经网络模型 LSTM2，所述第二主干路通过所述第二神经网络模型 LSTM2 提取所述输入文本信息的第二向量特征集。
- 19、根据权利要求 18 所述的计算机可读存储介质，其中，
- 40 所述第一预测分类器、所述第二预测分类器以及所述第三预测分类器内均设置有激活函数，所述第一向量特征集或所述第二向量特征集通过所述激活函数后实现对所述第一向量特征集或所述第二向量特征集的标注；其中，所述激活函数的计算过程如下：

$$z_t^b = \text{Softmax}(W^1 h_t^0)$$

$$y_t^1 = arg \max_y z_t^b$$

其中， W^1 为所述命名实体模型中需要训练的参数，与预测分类器的标注类型相关联； h_t^0 代指第一向量特征集或第二向量特征集， y_t^1 为输出的标注结果；Softmax 函数为归一化函数，是用于将 $W^1 h_t^0$ 的值映射到(0,1)区间内：

$$f(i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

其中， $[x_1, x_2, \dots, x_i, x_n]$ 为一个数组，代指 $W^1 h_t^1$ ， $f(i)$ 为第 i 个元素的 softmax 值。

5 20、根据权利要求 19 所述的计算机可读存储介质，其中，

在所述总预测分类器内设置有第一预测组件 SC 和第二预测组件 BG；其中，第一预测组件 SC 用于对所述第二向量特征集进行优化，以增强所述第二向量特征集中当前向量特征与前一时刻向量特征之间的联系；

10 所述第二预测组件 BG 用于根据优化后的第二向量特征集、所述第一向量特征集以及所述 unified-tag 标注类型标注出所述输入文本信息的最终标注。

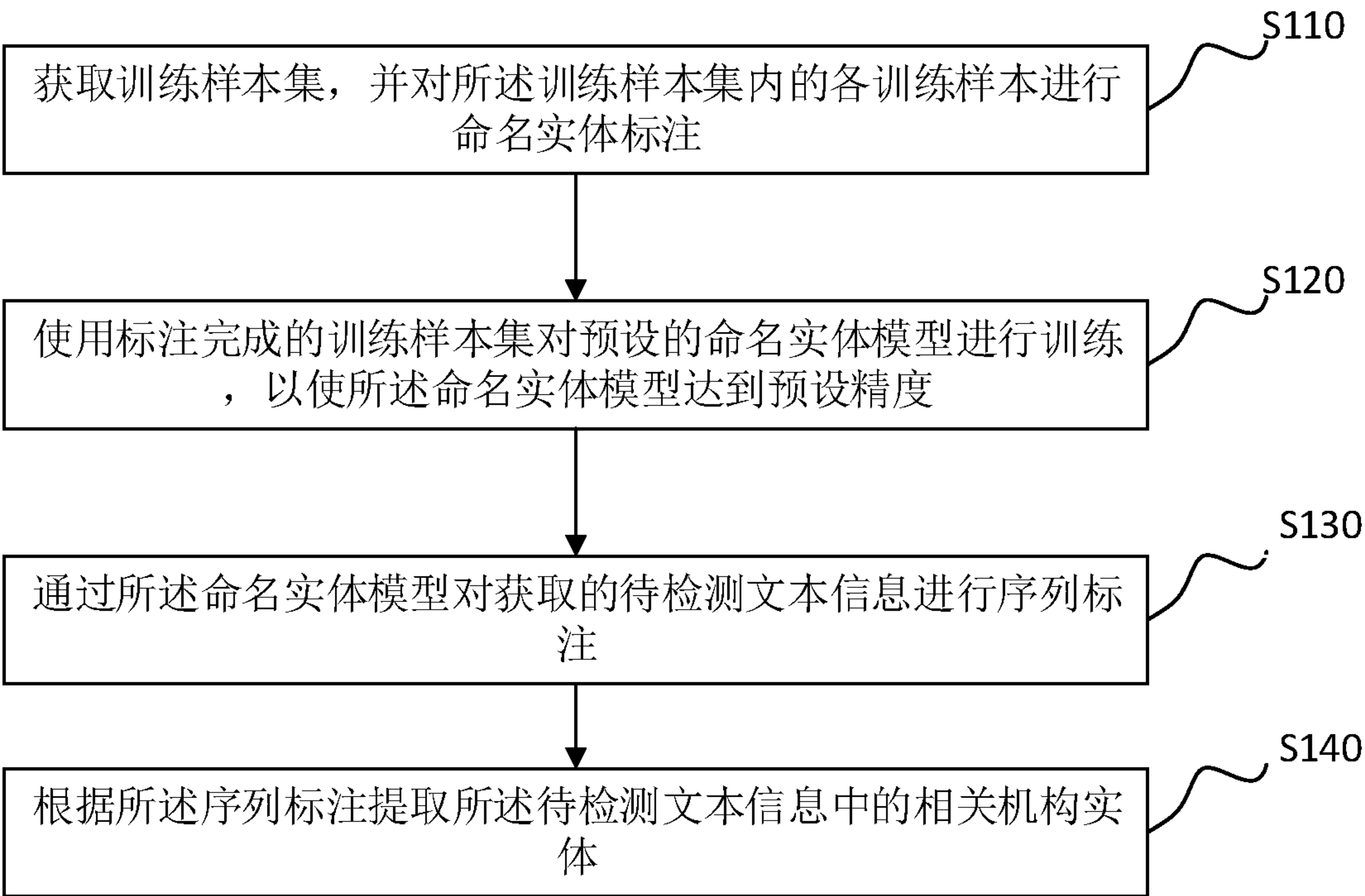


图 1

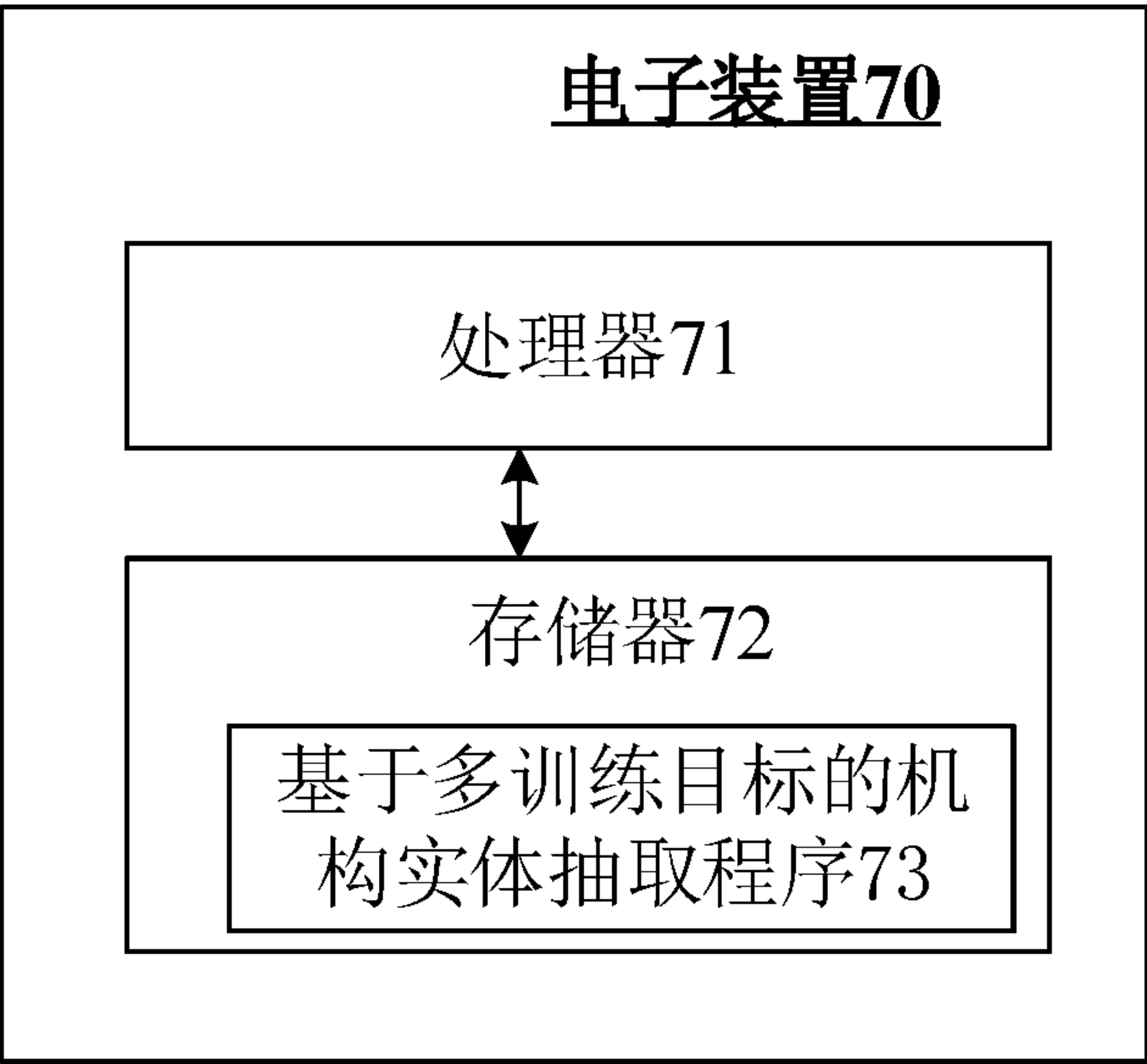


图 2

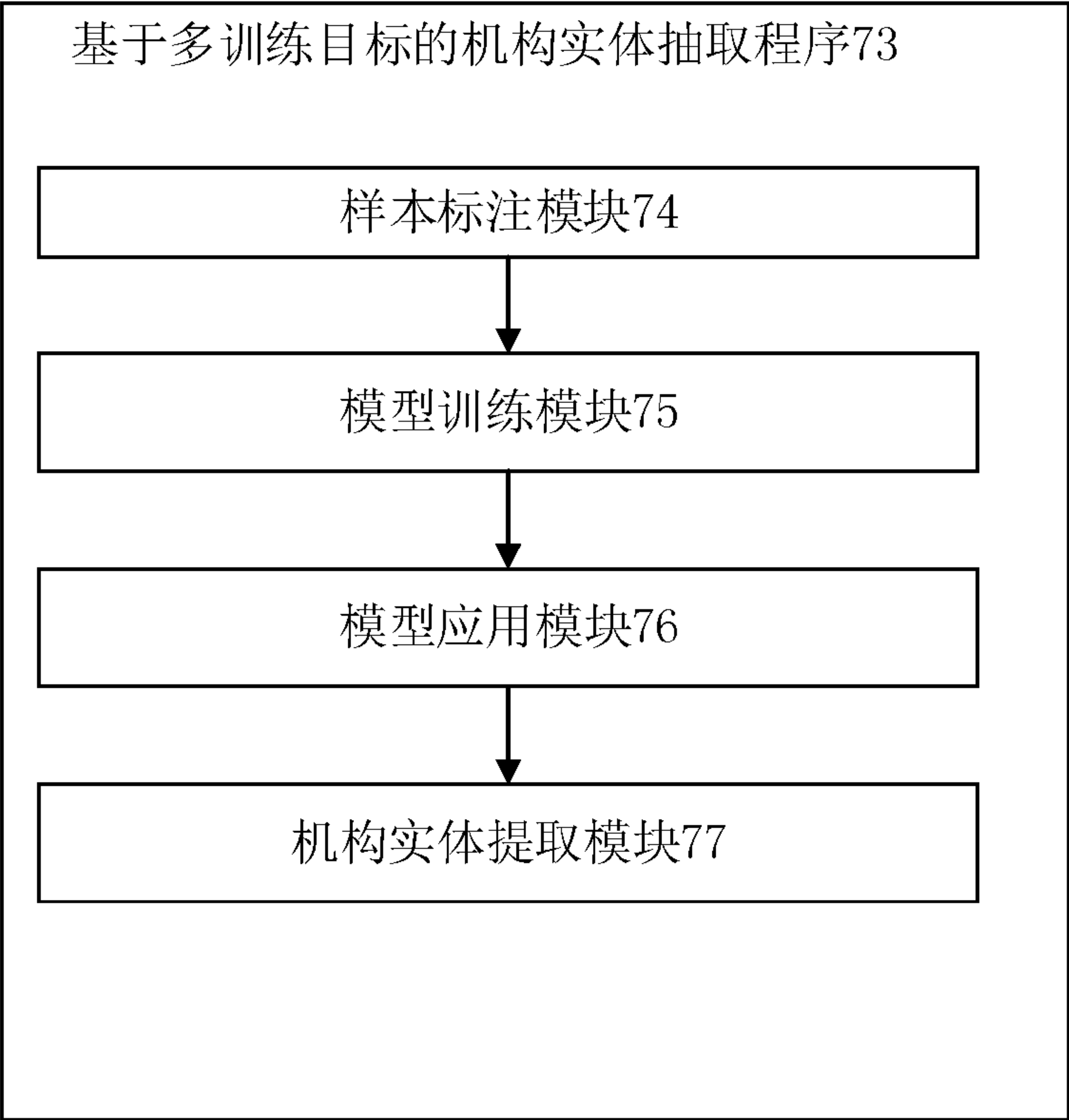


图 3

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/118331

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/30(2020.01)i; G06F 16/36(2019.01)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)

G06F; G06K; G06N

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)

WPI, EPODOC, CNPAT, CNKI: 标签, 训练, 标注, 实体, 序列, 命名, 模型, 向量, 特征, LSTM, 预测, 分类, entit+, label, train +, vector, feature, named, model

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 111428501 A (PEKING UNIVERSITY FOUNDER GROUP CORP. et al.) 17 July 2020 (2020-07-17) description, paragraphs [0044]-[0117]	1-3, 8-10, 15, 16
A	CN 110287480 A (GUANGZHOU DUOYI NETWORK CO., LTD. et al.) 27 September 2019 (2019-09-27) entire document	1-20
A	CN 110705294 A (SUNING CLOUD COMPUTING CO., LTD.) 17 January 2020 (2020-01-17) entire document	1-20
A	CN 109871538 A (SOUTH CHINA UNIVERSITY OF TECHNOLOGY) 11 June 2019 (2019-06-11) entire document	1-20
A	CN 109635280 A (YUANBAO TECHNOLOGY (WUHAN) CO., LTD.) 16 April 2019 (2019-04-16) 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

06 April 2021

Date of mailing of the international search report

27 April 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

International application No.

PCT/CN2020/118331

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 110162772 A (BEIJING SCIENCE AND TECHNOLOGY CO., THREE FAST ONLINE) 23 August 2019 (2019-08-23) entire document	1-20

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/CN2020/118331

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	111428501	A	17 July 2020	None	
CN	110287480	A	27 September 2019	None	
CN	110705294	A	17 January 2020	None	
CN	109871538	A	11 June 2019	None	
CN	109635280	A	16 April 2019	None	
CN	110162772	A	23 August 2019	None	

国际检索报告		国际申请号 PCT/CN2020/118331
A. 主题的分类 G06F 40/30(2020.01)i; G06F 16/36(2019.01)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) G06F; G06K; G06N 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)) WPI, EPDOC, CNPAT, CNKI: 标签, 训练, 标注, 实体, 序列, 命名, 模型, 向量, 特征, LSTM, 预测, 分类, entit+, label, train+, vector, feature, named, model		
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
X	CN 111428501 A (北大方正集团有限公司 等) 2020年 7月 17日 (2020 - 07 - 17) 说明书第[0044]-[0117]段	1-3、8-10、15、16
A	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) 全文	1-20
A	CN 110705294 A (苏宁云计算有限公司) 2020年 1月 17日 (2020 - 01 - 17) 全文	1-20
A	CN 109871538 A (华南理工大学) 2019年 6月 11日 (2019 - 06 - 11) 全文	1-20
A	CN 109635280 A (园宝科技武汉有限公司) 2019年 4月 16日 (2019 - 04 - 16) 全文	1-20
A	CN 110162772 A (北京三快在线科技有限公司) 2019年 8月 23日 (2019 - 08 - 23) 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 4月 6日		国际检索报告邮寄日期 2021年 4月 27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		授权官员 丁文勃 电话号码 86-(10)-53961330

国际检索报告 关于同族专利的信息				国际申请号 PCT/CN2020/118331	
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	111428501	A	2020年 7月 17日	无	
CN	110287480	A	2019年 9月 27日	无	
CN	110705294	A	2020年 1月 17日	无	
CN	109871538	A	2019年 6月 11日	无	
CN	109635280	A	2019年 4月 16日	无	
CN	110162772	A	2019年 8月 23日	无	