

(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



(10) 国际公布号

WO 2023/065635 A1

(43) 国际公布日

2023 年 4 月 27 日 (27.04.2023)

(51) 国际专利分类号:

G06F 40/295 (2020.01) G06N 3/08 (2006.01)

G06N 3/04 (2006.01)

(21) 国际申请号: PCT/CN2022/089993

(22) 国际申请日: 2022 年 4 月 28 日 (28.04.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:

202111233302.X 2021 年 10 月 22 日 (22.10.2021) 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: 司世景(SI, Shijing); 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)。王健宗(WANG, Jianzong); 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: 深圳市赛恩倍吉知识产权代理有限公司(SHENZHEN SCIENBIZIP INTELLECTUAL PROPERTY AGENCY CO., LTD.); 中国广东省深圳市龙华新区龙观东路 83 号荣群大厦 9 楼, Guangdong 518109 (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: NAMED ENTITY RECOGNITION METHOD AND APPARATUS, STORAGE MEDIUM AND TERMINAL DEVICE

(54) 发明名称: 命名实体识别方法、装置、存储介质及终端设备

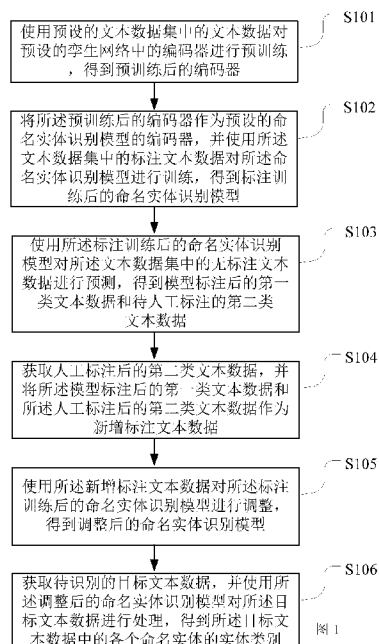


图 1

S101 Pre-train an encoder in a preset Siamese network by using text data in a preset text data set to obtain a pre-trained encoder
S102 Use the pre-trained encoder as an encoder of a preset named entity recognition (NER) model, and train the NER model by using labeled text data in the text data set to obtain a labeled and trained NER model
S103 Predict unlabeled text data in the text data set by using the labeled and trained NER model to obtain a first type of text data after model labeling and a second type of text data to be manually labeled
S104 Obtain a second type of text data after manual labeling, and use the first type of text data after the model labeling and the second type of text data after manual labeling as newly added and labeled text data
S105 Adjust the labeled and trained NER model by using the newly added and labeled text data to obtain an adjusted NER model
S106 Obtain target text data to be recognized, and process the target text data by using the adjusted NER model to obtain an entity category of each named entity in the target text data

(57) Abstract: The present application belongs to the technical field of artificial intelligence, and relates in particular to a named entity recognition (NER) method and apparatus, a computer-readable storage medium and a terminal device. The method comprises: pre-training an encoder in a Siamese network, using same as an encoder of a NER model, and training the model by using labeled text data to obtain a labeled and trained model; predicting unlabeled text data by using the labeled and trained model to obtain first and second types of text data; obtaining a second type of text data after manual labeling, and using the first type of text data after model

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, JM, 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))。

labeling and the second type of text data after manual labeling as newly added and labeled text data; adjusting the labeled and trained model by using the newly added and labeled text data to obtain an adjusted model; and obtaining target text data to be recognized, and processing the target text data by using the adjusted model to obtain an entity category of each named entity in the target text data.

(57) 摘要: 本申请属于人工智能技术领域, 尤其涉及一种命名实体识别方法、装置、计算机可读存储介质及终端设备。所述方法包括: 对孪生网络中的编码器进行预训练, 将其作为命名实体识别模型的编码器, 使用标注文本数据对模型进行训练, 得到标注训练后的模型; 使用标注训练后的模型对无标注文本数据进行预测, 得到第一类和第二类文本数据; 获取人工标注后的第二类文本数据, 并将模型标注后的第一类文本数据和人工标注后的第二类文本数据作为新增标注文本数据; 使用新增标注文本数据对标注训练后的模型进行调整, 得到调整后的模型; 获取待识别的目标文本数据, 使用调整后的模型对目标文本数据进行处理, 得到目标文本数据中的各个命名实体的实体类别。

命名实体识别方法、装置、存储介质及终端设备

本申请要求于 2021 年 10 月 22 日提交中国专利局，申请号为 202111233302.X 申请名称为“命名实体识别方法、装置、存储介质及终端设备”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

本申请属于人工智能技术领域，尤其涉及一种命名实体识别方法、装置、计算机可读存储介质及终端设备。

背景技术

命名实体识别（Named Entity Recognition，NER）也称为实体识别、实体分块或实体提取，是信息提取的一个子任务，旨在对非结构化文本中的命名实体进行定位和分类，将其分类为预定义类别，如人名、组织、位置、医疗代码、时间表达式、数量、货币价值、百分比等。现有技术中的方法通常采用 IOB 标记方案将其转化为序列标记问题，其中双向长短期记忆网络结合条件随机场是典型的模型，这些方法取得的巨大成功得益于大量正确人工标注的数据。但是，发明人发现在实际的命名实体识别的场景中，即使只人工标注几千或者几万个训练数据，所消耗的时间和金钱成本也是巨大的，而为了获得更高的识别准确率，则需要更多的标注数据，也会造成更大的成本消耗。

发明内容

本申请提出一种命名实体识别方法、装置、存储介质及终端设备，以解决现有的命名实体识别方法成本消耗较大的问题。

本申请的第一方面提供一种命名实体识别方法，所述方法包括：

使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练，得到预训练后的编码器；

将所述预训练后的编码器作为预设的命名实体识别模型的编码器，并使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型；

使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测，得到模型标注后的第一类文本数据和待人工标注的第二类文本数据；

获取人工标注后的第二类文本数据，并将所述模型标注后的第一类文本数据和所述人工标注后的第二类文本数据作为新增标注文本数据；

使用所述新增标注文本数据对所述标注训练后的命名实体识别模型进行调整，得到调整后的命名实体识别模型；

获取待识别的目标文本数据，并使用所述调整后的命名实体识别模型对所述目标文本数据进行处理，得到所述目标文本数据中的各个命名实体的实体类别。

本申请的第二方面提供一种计算机可读存储介质，所述计算机可读存储介质存储有至少一个计算机可读指令，所述至少一个计算机可读指令被处理器执行时实现以下步骤：

使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练，得到预训练后的编码器；

将所述预训练后的编码器作为预设的命名实体识别模型的编码器，并使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型；

使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测，得到模型标注后的第一类文本数据和待人工标注的第二类文本数据；

获取人工标注后的第二类文本数据，并将所述模型标注后的第一类文本数据和所述人工标注后的第二类文本数据作为新增标注文本数据；

使用所述新增标注文本数据对所述标注训练后的命名实体识别模型进行调整，得到调整后的命名实体识别模型；

获取待识别的目标文本数据，并使用所述调整后的命名实体识别模型对所述目标文本数据进行处理，得到所述目标文本数据中的各个命名实体的实体类别。

本申请的第三方面提供一种终端设备，所述终端设备包括存储器及处理器，所述存储器用于存储至少一个计算机可读指令，所述处理器用于执行所述至少一个计算机可读指令以实现以下步骤：

使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练，得到预训练后的编码器；

将所述预训练后的编码器作为预设的命名实体识别模型的编码器，并使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型；

使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测，得到模型标注后的第一类文本数据和待人工标注的第二类文本数据；

获取人工标注后的第二类文本数据，并将所述模型标注后的第一类文本数据和所述人工标注后的第二类文本数据作为新增标注文本数据；

使用所述新增标注文本数据对所述标注训练后的命名实体识别模型进行调整，得到调整后的命名实体识别模型；

获取待识别的目标文本数据，并使用所述调整后的命名实体识别模型对所述目标文本数据进行处理，得到所述目标文本数据中的各个命名实体的实体类别。

本申请的第四方面提供一种命名实体识别装置，所述装置包括：

编码器预训练模块，用于使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练，得到预训练后的编码器；

模型训练模块，用于将所述预训练后的编码器作为预设的命名实体识别模型的编码器，并使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型；

模型预测模块，用于使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测，得到模型标注后的第一类文本数据和待人工标注的第二类文本数据；

标注文本数据新增模块，用于获取人工标注后的第二类文本数据，并将所述模型标注后的第一类文本数据和所述人工标注后的第二类文本数据作为新增标注文本数据；

模型调整模块，用于使用所述新增标注文本数据对所述标注训练后的命名实体识别模型进行调整，得到调整后的命名实体识别模型；

命名实体识别模块，用于获取待识别的目标文本数据，并使用所述调整后的命名实体识别模型对所述目标文本数据进行处理，得到所述目标文本数据中的各个命名实体的实体类别。

本申请所述的命名实体识别方法、装置、存储介质及终端设备，有效减少了对于时间和金钱成本的消耗。

附图说明

图1为本申请实施例一种命名实体识别方法的一个实施例流程图；

图 2 为使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练的示意图；

图 3 为使用文本数据集中的标注文本数据对命名实体识别模型进行训练的示意图；

图 4 为命名实体识别模型的示意图；

图 5 为本申请实施例中一种命名实体识别装置的一个实施例结构图；

图 6 为本申请实施例中一种终端设备的示意框图。

具体实施方式

为使得本申请的发明目的、特征、优点能够更加的明显和易懂，下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，下面所描述的实施例仅仅是本申请一部分实施例，而非全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例，都属于本申请保护的范围。

本申请实施例可以基于人工智能技术对相关的数据进行获取和处理。其中，人工智能（Artificial Intelligence, AI）是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。

人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、机器人技术、生物识别技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

本申请实施例的执行主体可以为基于人工智能的终端设备，用于执行本申请实施例中的命名实体识别方法。

请参阅图 1，本申请实施例中一种命名实体识别方法的一个实施例可以包括：

步骤 S101、使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练，得到预训练后的编码器。

如图 2 所示，步骤 S101 具体可以包括如下过程：

步骤 S1011、对所述文本数据集中的文本数据进行数据增强，得到预设数量的增强文本数据对。

其中，所述文本数据集中可以包括标注文本数据和无标注文本数据。任意一个所述增强文本数据对中均包括由同一文本数据进行数据增强后得到的两个不同的增强文本数据。

在本申请实施例中，通过对同一样本进行数据增强，能够得到一系列自相似的文本数据对作为正例。具体采用何种数据增强方法，可以根据实际情况进行设置，例如，在一种实现方式中，可以使用 dropout 方法进行数据增强，即通过随机采样 dropout 掩码（dropout mask）来对文本数据进行 dropout 操作，同一个文本数据进行连续两次的 dropout 操作，可以分别得到两个不同的增强文本数据，从而组成一个增强文本数据对。

需要注意的是，以上仅为举例，而非对数据增强方法的限定，在实际应用中，可以根据具体情况采用现有技术中其它的数据增强方法，本申请实施例对此不作具体限定。

步骤 S1012、使用所述孪生网络对所述增强文本数据对进行处理，分别得到第一特征向量和第二特征向量。

所述孪生网络可以为 SimSiam 网络，SimSiam 网络中包括两条处理分支（分别记为分支 1 和分支 2），分别对增强文本数据对中的两个增强文本数据（分别记为文本数据 1 和文本数据 2）进行处理，其中，分支 1 中的编码器（encoder）对文本数据 1 进行编码，得到第一特征向量；分支 2 中的编码器对文本数据 2 进行编码，编码结果再经过预测器（predictor）的非线性变化，得到第二特征向量。需要注意的是，分支 1 中的编码器和分支 2 中的编码器共用相同的参数，可以将其认为是同一个编码器。

步骤 S1013、根据所述第一特征向量和所述第二特征向量计算第一损失函数。

具体地，可以根据下式计算所述第一损失函数：

$$\mathcal{D}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}$$

其中， p_1 为所述第一特征向量， z_2 为所述第二特征向量， $\|p_1\|_2$ 为所述第一特征向量的模， $\|z_2\|_2$ 为所述第二特征向量的模， $\mathcal{D}(p_1, z_2)$ 为所述第一损失函数。

步骤 S1014、以最小化所述第一损失函数为目标对所述孪生网络中的编码器进行预训练，得到预训练后的编码器。

在计算得到所述第一损失函数之后，则可以根据所述第一损失函数对所述孪生网络的模型参数进行调整。在本申请实施例中，假设所述孪生网络的模型参数为 W_1 ，将第一损失函数反向传播修改所述孪生网络的模型参数 W_1 ，得到修改后的模型参数 W_2 。修改参数之后再继续执行下一次的训练过程，在该次训练过程中，重新计算得到所述第一损失函数，将所述第一损失函数反向传播修改所述孪生网络的模型参数 W_2 ，得到修改后的模型参数 W_3 ，……，以此类推，不断重复以上过程，每次训练过程均可对所述孪生网络的模型参数进行修改，直至满足预设的训练条件，其中，训练条件可以是训练次数达到预设的次数阈值，次数阈值可以根据实际情况进行设置，例如，可以将其设置为数千、数万、数十万甚至更大的数值；训练条件也可以是所述孪生网络收敛；由于可能出现训练次数还未达到次数阈值，但所述孪生网络已经收敛，可能导致重复不必要的工作；或者所述孪生网络始终无法收敛，可能导致无限循环，无法结束训练的过程，基于上述两种情况，训练条件还可以是训练次数达到次数阈值或所述孪生网络收敛。当满足训练条件，即可得到预训练后的孪生网络，此时的编码器即为预训练后的编码器。

通过本申请实施例中的这种对比学习方式，编码器的参数已得到优化，可用于后续的命名实体识别过程。

步骤 S102、将所述预训练后的编码器作为预设的命名实体识别模型的编码器，并使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型。

如图 3 所示，步骤 S102 具体可以包括如下过程：

步骤 S1021、使用所述命名实体识别模型的编码器对所述文本数据集中的标注文本数据进行编码，得到编码后的特征向量。

图 4 所示即为所述命名实体识别模型的示意图，所述命名实体识别模型可以包括编码器和多层感知器（Multi-Layer Perception）。 i 为所述文本数据集中的标注文本数据的序号， x_i 为所述文本数据集中的第 i 个标注文本数据， $1 \leq i \leq n$ ， n 为所述文本数据集中的标注文本数据的数目，所述命名实体识别模型的编码器对 x_i 进行编码，得到与 x_i 对应的编码后的特征向量，将其记为 h_i 。

步骤 S1022、使用所述命名实体识别模型的多层感知器对所述编码后的特征向量进行处理，得到实体类别的概率分布。

具体地，可以根据下式计算实体类别的概率分布：

$$p_i = \text{Softmax}(U \tanh(Vh_i))$$

其中， U 和 V 均为预设的模型参数， Softmax 为预设的激励函数， p_i 为与 x_i 对应的实体类别的概率分布。

步骤 S1023、根据所述概率分布计算第二损失函数。

具体地，可以根据下式计算所述第二损失函数：

$$\text{loss} = \sum_{1 \leq i \leq n} -\log p_i[z_i]$$

其中， z_i 为与 x_i 对应的实体类别标签，即预先人工标注的实体类别， loss 为所述第二损失

函数。

步骤 S1024、以最小化所述第二损失函数为目标对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型。

在计算得到所述第二损失函数之后，则可以根据所述第二损失函数对所述命名实体识别模型的模型参数进行调整。在本申请实施例中，假设所述命名实体识别模型的模型参数为V1，将第二损失函数反向传播修改所述命名实体识别模型的模型参数V1，得到修改后的模型参数V2。修改参数之后再继续执行下一次的训练过程，在该次训练过程中，重新计算得到所述第二损失函数，将所述第二损失函数反向传播修改所述命名实体识别模型的模型参数V2，得到修改后的模型参数V3，……，以此类推，不断重复以上过程，每次训练过程均可对所述命名实体识别模型的模型参数进行修改，直至满足预设的训练条件，其中，训练条件可以是训练次数达到预设的次数阈值，次数阈值可以根据实际情况进行设置，例如，可以将其设置为数千、数万、数十万甚至更大的数值；训练条件也可以是所述命名实体识别模型收敛；由于可能出现训练次数还未达到次数阈值，但所述命名实体识别模型已经收敛，可能导致重复不必要的工作；或者所述命名实体识别模型始终无法收敛，可能导致无限循环，无法结束训练的过程，基于上述两种情况，训练条件还可以是训练次数达到次数阈值或所述命名实体识别模型收敛。当满足训练条件，即可得到标注训练后的命名实体识别模型。

步骤 S103、使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测，得到模型标注后的第一类文本数据和待人工标注的第二类文本数据。

具体地，可以使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测，并计算预测结果的置信度。然后将预测结果的置信度大于或等于预设的置信度阈值的无标注文本数据作为第一类文本数据，即较易通过模型进行标注的文本数据；将预测结果的置信度小于所述置信度阈值的无标注文本数据作为第二类文本数据，即难以通过模型进行标注的文本数据。例如，如果是对“Mary”进行标注，如果模型给出的预测标注为实体类别“PER”和“LOC”的概率较为接近，则无法确定其最终的标注结果，因此可将其作为第二类文本数据，需要筛选出来进行人工标注。所述置信度阈值的具体取值可以根据实际情况进行设置，本申请实施例对此不作具体限定。

步骤 S104、获取人工标注后的第二类文本数据，并将所述模型标注后的第一类文本数据和所述人工标注后的第二类文本数据作为新增标注文本数据。

在本申请实施例中，可以安排专家对筛选出来的第二类文本数据进行人工标注，通过这样的主动学习方式，对较易通过模型进行标注的文本数据直接使用模型进行标注，仅将经过筛选的难以通过模型进行标注的文本数据进行人工标注，极大提高模型的整体效率。

步骤 S105、使用所述新增标注文本数据对所述标注训练后的命名实体识别模型进行调整，得到调整后的命名实体识别模型。

经过上述过程，一部分的无标注文本数据转变成了标记文本数据，标记文本数据的数量在该过程中不断增多，可以继续使用新增标注文本数据对命名实体识别模型进行调整，得到调整后的命名实体识别模型。

需要注意的是，模型的调整是一个持续迭代的过程，即不断地重复步骤 S103 至步骤 S105，持续地将无标注文本数据转变为标记文本数据，并使用新增标注文本数据进行新一轮的模型调整，直至最终得到的命名实体识别模型达到预定的识别精度为止。在后续的命名实体识别任务中，即可使用该模型进行命名实体识别。

步骤 S106、获取待识别的目标文本数据，并使用所述调整后的命名实体识别模型对所述目标文本数据进行处理，得到所述目标文本数据中的各个命名实体的实体类别。

所述目标文本数据可以是预先存储于所述终端设备中，也可以是由其他设备通过预设的通信方式发送至所述终端设备中，还可以是由用户通过预设的人机交互界面输入到所述终端设备中。当需要进行命名实体识别时，所述终端设备即可使用所述调整后的命名实体识别模

型对所述目标文本数据进行处理，从而得到所述目标文本数据中的各个命名实体的实体类别。例如，若所述目标文本数据为“小明早上 8 点去学校上课。”，则经过所述调整后的命名实体识别模型处理之后，最终得到的识别结果为：（命名实体：小明，实体类别：人名）；（命名实体：早上 8 点，实体类别：时间）；（命名实体：学校，实体类别：地点）。

综上所述，本申请实施例使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练，得到预训练后的编码器；将所述预训练后的编码器作为预设的命名实体识别模型的编码器，并使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型；使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测，得到模型标注后的第一类文本数据和待人工标注的第二类文本数据；获取人工标注后的第二类文本数据，并将所述模型标注后的第一类文本数据和所述人工标注后的第二类文本数据作为新增标注文本数据；使用所述新增标注文本数据对所述标注训练后的命名实体识别模型进行调整，得到调整后的命名实体识别模型；获取待识别的目标文本数据，并使用所述调整后的命名实体识别模型对所述目标文本数据进行处理，得到所述目标文本数据中的各个命名实体的实体类别。通过本申请实施例，将主动学习与对比学习进行融合，首先利用对比学习的方法预训练编码器并用于命名实体识别模型的训练，在随后的主动学习过程中，模型通过收到的反馈不断进行自我的迭代训练与优化，整个过程仅需少量的人工标注，有效减少了对于时间和金钱成本的消耗。

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。

对应于上文实施例所述的一种命名实体识别方法，图 5 示出了本申请实施例提供的一种命名实体识别装置的一个实施例结构图。

本实施例中，一种命名实体识别装置可以包括：

编码器预训练模块 501，用于使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练，得到预训练后的编码器；

模型训练模块 502，用于将所述预训练后的编码器作为预设的命名实体识别模型的编码器，并使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型；

模型预测模块 503，用于使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测，得到模型标注后的第一类文本数据和待人工标注的第二类文本数据；

标注文本数据新增模块 504，用于获取人工标注后的第二类文本数据，并将所述模型标注后的第一类文本数据和所述人工标注后的第二类文本数据作为新增标注文本数据；

模型调整模块 505，用于使用所述新增标注文本数据对所述标注训练后的命名实体识别模型进行调整，得到调整后的命名实体识别模型；

命名实体识别模块 506，用于获取待识别的目标文本数据，并使用所述调整后的命名实体识别模型对所述目标文本数据进行处理，得到所述目标文本数据中的各个命名实体的实体类别。

在本申请实施例的一种具体实现方式中，所述编码器预训练模块可以包括：

数据增强单元，用于对所述文本数据集中的文本数据进行数据增强，得到预设数量的增强文本数据对；其中，任意一个所述增强文本数据对中均包括由同一文本数据进行数据增强后的得到的两个不同的增强文本数据；

孪生网络处理单元，用于使用所述孪生网络对所述增强文本数据对进行处理，分别得到第一特征向量和第二特征向量；

第一损失函数计算单元，用于根据所述第一特征向量和所述第二特征向量计算第一损失函数；

预训练单元,用于以最小化所述第一损失函数为目标对所述孪生网络中的编码器进行预训练,得到预训练后的编码器。

在本申请实施例的一种具体实现方式中,所述第一损失函数计算单元可以具体用于根据下式计算所述第一损失函数:

$$\mathcal{D}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}$$

其中, p_1 为所述第一特征向量, z_2 为所述第二特征向量, $\|p_1\|_2$ 为所述第一特征向量的模, $\|z_2\|_2$ 为所述第二特征向量的模, $\mathcal{D}(p_1, z_2)$ 为所述第一损失函数。

在本申请实施例的一种具体实现方式中,所述模型训练模块可以包括:

编码单元,用于使用所述命名实体识别模型的编码器对所述文本数据集中的标注文本数据进行编码,得到编码后的特征向量;

多层感知单元,用于使用所述命名实体识别模型的多层感知器对所述编码后的特征向量进行处理,得到实体类别的概率分布;

第二损失函数计算单元,用于根据所述概率分布计算第二损失函数;

模型训练单元,用于以最小化所述第二损失函数为目标对所述命名实体识别模型进行训练,得到标注训练后的命名实体识别模型。

在本申请实施例的一种具体实现方式中,所述多层感知单元可以具体用于根据下式计算实体类别的概率分布:

$$p_i = \text{Softmax}(U \tanh(Vh_i))$$

其中, i 为所述文本数据集中的标注文本数据的序号, h_i 为与所述文本数据集中的第 i 个标注文本数据对应的编码后的特征向量, U 和 V 均为预设的模型参数, Softmax 为预设的激励函数, p_i 为与所述文本数据集中的第 i 个标注文本数据对应的实体类别的概率分布。

在本申请实施例的一种具体实现方式中,所述第二损失函数计算单元可以具体用于根据下式计算所述第二损失函数:

$$\text{loss} = \sum_{1 \leq i \leq n} -\log p_i[z_i]$$

其中, n 为所述文本数据集中的标注文本数据的数目, z_i 为与所述文本数据集中的第 i 个标注文本数据对应的实体类别标签, loss 为所述第二损失函数。

在本申请实施例的一种具体实现方式中,所述模型预测模块可以包括:

预测单元,用于使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测,并计算预测结果的置信度;

第一类文本数据确定单元,用于将预测结果的置信度大于或等于预设的置信度阈值的无标注文本数据作为所述第一类文本数据;

第二类文本数据确定单元,用于将预测结果的置信度小于所述置信度阈值的无标注文本数据作为所述第二类文本数据。

所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的装置,模块和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中未详述或记载的部分,可以参见其它实施例的相关描述。

图6示出了本申请实施例提供的一种终端设备的示意框图,为了便于说明,仅示出了与本申请实施例相关的部分。

在本实施例中,所述终端设备6可以是桌上型计算机、笔记本、掌上电脑等计算设备。该终端设备6可包括:处理器60、存储器61以及存储在所述存储器61中并可在所述处理器60上运行的计算机可读指令62,例如执行上述的命名实体识别方法的计算机可读指令。所述处理器60执行所述计算机可读指令62时实现上述各个命名实体识别方法实施例中的步骤,

例如图 1 所示的步骤 S101 至 S106。或者，所述处理器 60 执行所述计算机可读指令 62 时实现上述各装置实施例中各模块/单元的功能，例如图 5 所示模块 501 至 506 的功能。

示例性的，所述计算机可读指令 62 可以被分割成一个或多个模块/单元，所述一个或者多个模块/单元被存储在所述存储器 61 中，并由所述处理器 60 执行，以完成本申请。所述一个或多个模块/单元可以是能够完成特定功能的一系列计算机可读指令段，该指令段用于描述所述计算机可读指令 62 在所述终端设备 6 中的执行过程。

所述处理器 60 可以是中央处理单元（Central Processing Unit，CPU），还可以是其它通用处理器、数字信号处理器（Digital Signal Processor，DSP）、专用集成电路（Application Specific Integrated Circuit，ASIC）、现场可编程门阵列（Field-Programmable Gate Array，FPGA）或者其它可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

所述存储器 61 可以是所述终端设备 6 的内部存储单元，例如终端设备 6 的硬盘或内存。所述存储器 61 也可以是所述终端设备 6 的外部存储设备，例如所述终端设备 6 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。进一步地，所述存储器 61 还可以既包括所述终端设备 6 的内部存储单元也包括外部存储设备。所述存储器 61 用于存储所述计算机可读指令以及所述终端设备 6 所需的其它指令和文本数据。所述存储器 61 还可以用于暂时地存储已经输出或者将要输出的文本数据。

进一步地，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。

进一步地，所述计算机可读存储介质可主要包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需的应用程序等；存储数据区可存储根据区块链节点的使用所创建的数据等。

在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。

所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个计算机可读存储介质中。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或者该技术方案的全部或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一个存储介质中，包括若干计算机可读指令用以使得一台计算机设备（可以是个人计算机，服务器，或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U 盘、移动硬盘、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、磁碟或者光盘等各种可以存储计算机可读指令的介质。

以上所述实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

权 利 要 求 书

1. 一种命名实体识别方法，其中，所述方法包括：

使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练，得到预训练后的编码器；

将所述预训练后的编码器作为预设的命名实体识别模型的编码器，并使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型；

使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测，得到模型标注后的第一类文本数据和待人工标注的第二类文本数据；

获取人工标注后的第二类文本数据，并将所述模型标注后的第一类文本数据和所述人工标注后的第二类文本数据作为新增标注文本数据；

使用所述新增标注文本数据对所述标注训练后的命名实体识别模型进行调整，得到调整后的命名实体识别模型；

获取待识别的目标文本数据，并使用所述调整后的命名实体识别模型对所述目标文本数据进行处理，得到所述目标文本数据中的各个命名实体的实体类别。

2. 根据权利要求1所述的命名实体识别方法，其中，所述使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练，得到预训练后的编码器，包括：

对所述文本数据集中的文本数据进行数据增强，得到预设数量的增强文本数据对；其中，任意一个所述增强文本数据对中均包括由同一文本数据进行数据增强后得到的两个不同的增强文本数据；

使用所述孪生网络对所述增强文本数据对进行处理，分别得到第一特征向量和第二特征向量；

根据所述第一特征向量和所述第二特征向量计算第一损失函数；

以最小化所述第一损失函数为目标对所述孪生网络中的编码器进行预训练，得到预训练后的编码器。

3. 根据权利要求2所述的命名实体识别方法，其中，所述根据所述第一特征向量和所述第二特征向量计算第一损失函数，包括：

根据下式计算所述第一损失函数：

$$\mathcal{D}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}$$

其中， p_1 为所述第一特征向量， z_2 为所述第二特征向量， $\|p_1\|_2$ 为所述第一特征向量的模， $\|z_2\|_2$ 为所述第二特征向量的模， $\mathcal{D}(p_1, z_2)$ 为所述第一损失函数。

4. 根据权利要求1所述的命名实体识别方法，其中，所述使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型，包括：

使用所述命名实体识别模型的编码器对所述文本数据集中的标注文本数据进行编码，得到编码后的特征向量；

使用所述命名实体识别模型的多层感知器对所述编码后的特征向量进行处理，得到实体类别的概率分布；

根据所述概率分布计算第二损失函数；

以最小化所述第二损失函数为目标对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型。

5. 根据权利要求4所述的命名实体识别方法，其中，所述使用所述命名实体识别模型的多层感知器对所述编码后的特征向量进行处理，得到实体类别的概率分布，包括：

根据下式计算实体类别的概率分布：

$$p_i = \text{Softmax}(U \tanh(Vh_i))$$

其中, i 为所述文本数据集中的标注文本数据的序号, h_i 为与所述文本数据集中的第 i 个标注文本数据对应的编码后的特征向量, U 和 V 均为预设的模型参数, Softmax 为预设的激励函数, p_i 为与所述文本数据集中的第 i 个标注文本数据对应的实体类别的概率分布。

6. 根据权利要求 5 所述的命名实体识别方法, 其中, 所述根据所述概率分布计算第二损失函数, 包括:

根据下式计算所述第二损失函数:

$$\text{loss} = \sum_{1 \leq i \leq n} -\log p_i[z_i]$$

其中, n 为所述文本数据集中的标注文本数据的数目, z_i 为与所述文本数据集中的第 i 个标注文本数据对应的实体类别标签, loss 为所述第二损失函数。

7. 根据权利要求 1 至 6 中任一项所述的命名实体识别方法, 其中, 所述使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测, 得到模型标注后的第一类文本数据和待人工标注的第二类文本数据, 包括:

使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测, 并计算预测结果的置信度;

将预测结果的置信度大于或等于预设的置信度阈值的无标注文本数据作为所述第一类文本数据;

将预测结果的置信度小于所述置信度阈值的无标注文本数据作为所述第二类文本数据。

8. 一种计算机可读存储介质, 其中, 所述计算机可读存储介质存储有至少一个计算机可读指令, 所述至少一个计算机可读指令被处理器执行时实现以下步骤:

使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练, 得到预训练后的编码器;

将所述预训练后的编码器作为预设的命名实体识别模型的编码器, 并使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练, 得到标注训练后的命名实体识别模型;

使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测, 得到模型标注后的第一类文本数据和待人工标注的第二类文本数据;

获取人工标注后的第二类文本数据, 并将所述模型标注后的第一类文本数据和所述人工标注后的第二类文本数据作为新增标注文本数据;

使用所述新增标注文本数据对所述标注训练后的命名实体识别模型进行调整, 得到调整后的命名实体识别模型;

获取待识别的目标文本数据, 并使用所述调整后的命名实体识别模型对所述目标文本数据进行处理, 得到所述目标文本数据中的各个命名实体的实体类别。

9. 根据权利要求 8 所述的存储介质, 其中, 所述至少一个计算机可读指令被所述处理器执行以实现所述使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练, 得到预训练后的编码器时, 具体包括:

对所述文本数据集中的文本数据进行数据增强, 得到预设数量的增强文本数据对; 其中, 任意一个所述增强文本数据对中均包括由同一文本数据进行数据增强后得到的两个不同的增强文本数据;

使用所述孪生网络对所述增强文本数据对进行处理, 分别得到第一特征向量和第二特征向量;

根据所述第一特征向量和所述第二特征向量计算第一损失函数;

以最小化所述第一损失函数为目标对所述孪生网络中的编码器进行预训练, 得到预训练后的编码器。

10. 根据权利要求 9 所述的存储介质, 其中, 所述至少一个计算机可读指令被所述处理

器执行以实现所述根据所述第一特征向量和所述第二特征向量计算第一损失函数时，具体包括：

根据下式计算所述第一损失函数：

$$\mathcal{D}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}$$

其中， p_1 为所述第一特征向量， z_2 为所述第二特征向量， $\|p_1\|_2$ 为所述第一特征向量的模， $\|z_2\|_2$ 为所述第二特征向量的模， $\mathcal{D}(p_1, z_2)$ 为所述第一损失函数。

11. 根据权利要求 8 所述的存储介质，其中，所述至少一个计算机可读指令被所述处理器执行以实现所述使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型时，具体包括：

使用所述命名实体识别模型的编码器对所述文本数据集中的标注文本数据进行编码，得到编码后的特征向量；

使用所述命名实体识别模型的多层感知器对所述编码后的特征向量进行处理，得到实体类别的概率分布；

根据所述概率分布计算第二损失函数；

以最小化所述第二损失函数为目标对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型。

12. 根据权利要求 11 所述的存储介质，其中，所述至少一个计算机可读指令被所述处理器执行以实现所述使用所述命名实体识别模型的多层感知器对所述编码后的特征向量进行处理，得到实体类别的概率分布时，具体包括：

根据下式计算实体类别的概率分布：

$$p_i = \text{Softmax}(U \tanh(Vh_i))$$

其中， i 为所述文本数据集中的标注文本数据的序号， h_i 为与所述文本数据集中的第 i 个标注文本数据对应的编码后的特征向量， U 和 V 均为预设的模型参数， Softmax 为预设的激励函数， p_i 为与所述文本数据集中的第 i 个标注文本数据对应的实体类别的概率分布。

13. 根据权利要求 12 所述的存储介质，其中，所述至少一个计算机可读指令被所述处理器执行以实现所述根据所述概率分布计算第二损失函数时，具体包括：

根据下式计算所述第二损失函数：

$$\text{loss} = \sum_{1 \leq i \leq n} -\log p_i[z_i]$$

其中， n 为所述文本数据集中的标注文本数据的数目， z_i 为与所述文本数据集中的第 i 个标注文本数据对应的实体类别标签， loss 为所述第二损失函数。

14. 一种终端设备，其中，所述终端设备包括存储器及处理器，所述存储器用于存储至少一个计算机可读指令，所述处理器用于执行所述至少一个计算机可读指令以实现以下步骤：

使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练，得到预训练后的编码器；

将所述预训练后的编码器作为预设的命名实体识别模型的编码器，并使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型；

使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测，得到模型标注后的第一类文本数据和待人工标注的第二类文本数据；

获取人工标注后的第二类文本数据，并将所述模型标注后的第一类文本数据和所述人工标注后的第二类文本数据作为新增标注文本数据；

使用所述新增标注文本数据对所述标注训练后的命名实体识别模型进行调整，得到调整

后的命名实体识别模型；

获取待识别的目标文本数据，并使用所述调整后的命名实体识别模型对所述目标文本数据进行处理，得到所述目标文本数据中的各个命名实体的实体类别。

15. 根据权利要求 14 所述的终端设备，其中，所述处理器执行所述至少一个计算机可读指令以实现所述使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练，得到预训练后的编码器时，具体包括：

对所述文本数据集中的文本数据进行数据增强，得到预设数量的增强文本数据对；其中，任意一个所述增强文本数据对中均包括由同一文本数据进行数据增强后得到的两个不同的增强文本数据；

使用所述孪生网络对所述增强文本数据对进行处理，分别得到第一特征向量和第二特征向量；

根据所述第一特征向量和所述第二特征向量计算第一损失函数；

以最小化所述第一损失函数为目标对所述孪生网络中的编码器进行预训练，得到预训练后的编码器。

16. 根据权利要求 15 所述的终端设备，其中，所述处理器执行所述至少一个计算机可读指令以实现所述根据所述第一特征向量和所述第二特征向量计算第一损失函数时，具体包括：

根据下式计算所述第一损失函数：

$$\mathcal{D}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}$$

其中， p_1 为所述第一特征向量， z_2 为所述第二特征向量， $\|p_1\|_2$ 为所述第一特征向量的模， $\|z_2\|_2$ 为所述第二特征向量的模， $\mathcal{D}(p_1, z_2)$ 为所述第一损失函数。

17. 根据权利要求 14 所述的终端设备，其中，所述处理器执行所述至少一个计算机可读指令以实现所述使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型时，具体包括：

使用所述命名实体识别模型的编码器对所述文本数据集中的标注文本数据进行编码，得到编码后的特征向量；

使用所述命名实体识别模型的多层感知器对所述编码后的特征向量进行处理，得到实体类别的概率分布；

根据所述概率分布计算第二损失函数；

以最小化所述第二损失函数为目标对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型。

18. 根据权利要求 17 所述的终端设备，其中，所述处理器执行所述至少一个计算机可读指令以实现所述使用所述命名实体识别模型的多层感知器对所述编码后的特征向量进行处理，得到实体类别的概率分布时，具体包括：

根据下式计算实体类别的概率分布：

$$p_i = \text{Softmax}(U \tanh(Vh_i))$$

其中， i 为所述文本数据集中的标注文本数据的序号， h_i 为与所述文本数据集中的第 i 个标注文本数据对应的编码后的特征向量， U 和 V 均为预设的模型参数， Softmax 为预设的激励函数， p_i 为与所述文本数据集中的第 i 个标注文本数据对应的实体类别的概率分布。

19. 根据权利要求 18 所述的终端设备，其中，所述处理器执行所述至少一个计算机可读指令以实现所述根据所述概率分布计算第二损失函数时，具体包括：

根据下式计算所述第二损失函数：

$$\text{loss} = \sum_{1 \leq i \leq n} -\log p_i[z_i]$$

其中， n 为所述文本数据集中的标注文本数据的数目， z_i 为与所述文本数据集中的第 i 个

标注文本数据对应的实体类别标签，loss 为所述第二损失函数。

20.一种命名实体识别装置，其中，所述装置包括：

编码器预训练模块，用于使用预设的文本数据集中的文本数据对预设的孪生网络中的编码器进行预训练，得到预训练后的编码器；

模型训练模块，用于将所述预训练后的编码器作为预设的命名实体识别模型的编码器，并使用所述文本数据集中的标注文本数据对所述命名实体识别模型进行训练，得到标注训练后的命名实体识别模型；

模型预测模块，用于使用所述标注训练后的命名实体识别模型对所述文本数据集中的无标注文本数据进行预测，得到模型标注后的第一类文本数据和待人工标注的第二类文本数据；

标注文本数据新增模块，用于获取人工标注后的第二类文本数据，并将所述模型标注后的第一类文本数据和所述人工标注后的第二类文本数据作为新增标注文本数据；

模型调整模块，用于使用所述新增标注文本数据对所述标注训练后的命名实体识别模型进行调整，得到调整后的命名实体识别模型；

命名实体识别模块，用于获取待识别的目标文本数据，并使用所述调整后的命名实体识别模型对所述目标文本数据进行处理，得到所述目标文本数据中的各个命名实体的实体类别。

1/3

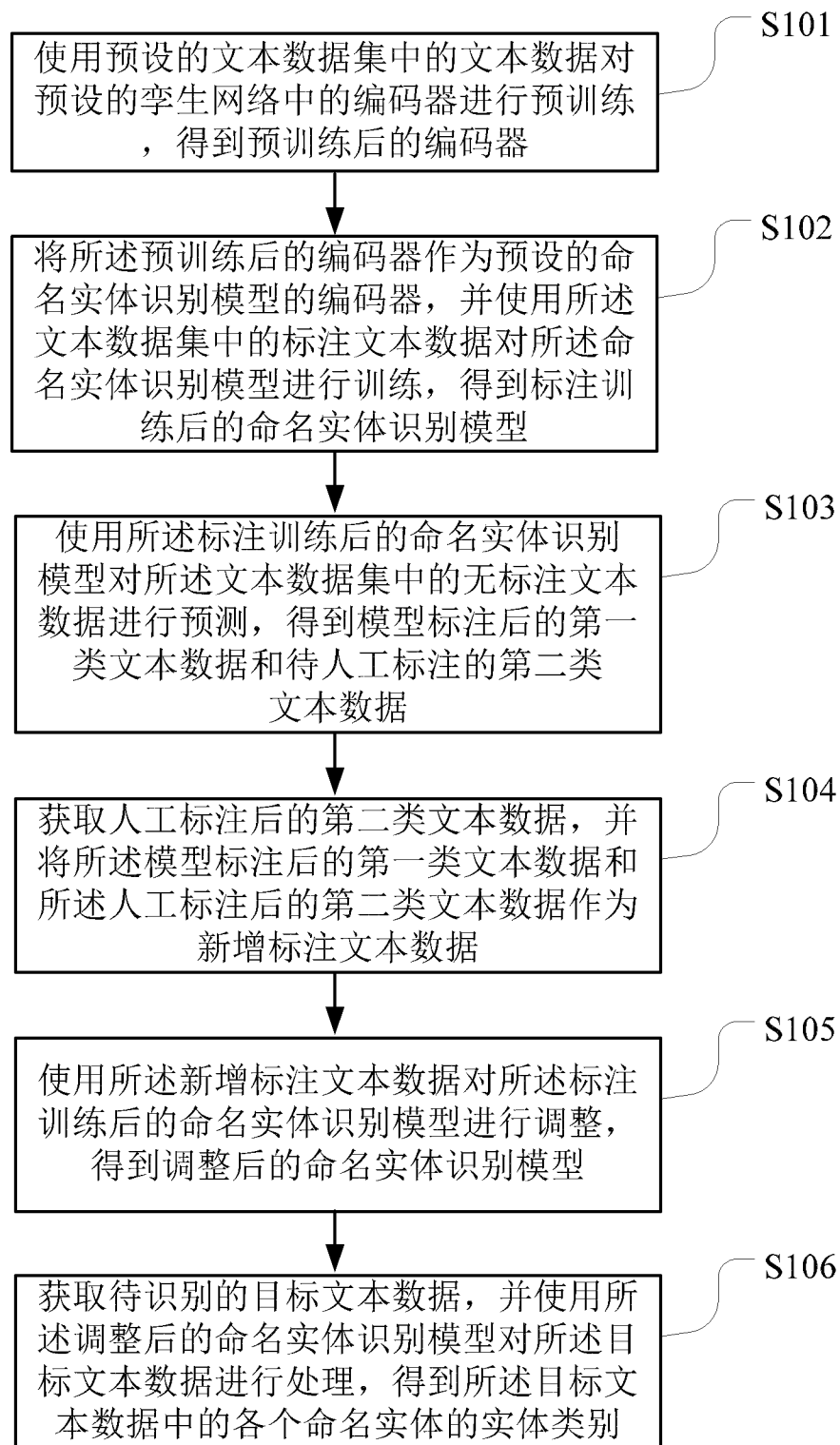


图 1

2/3

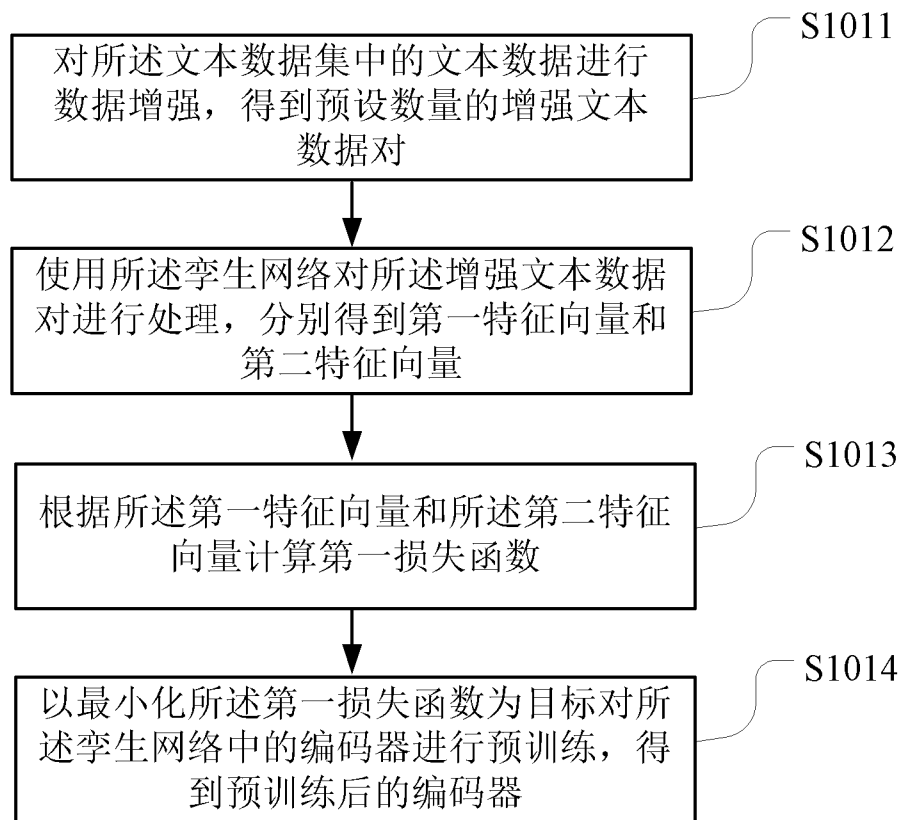


图 2

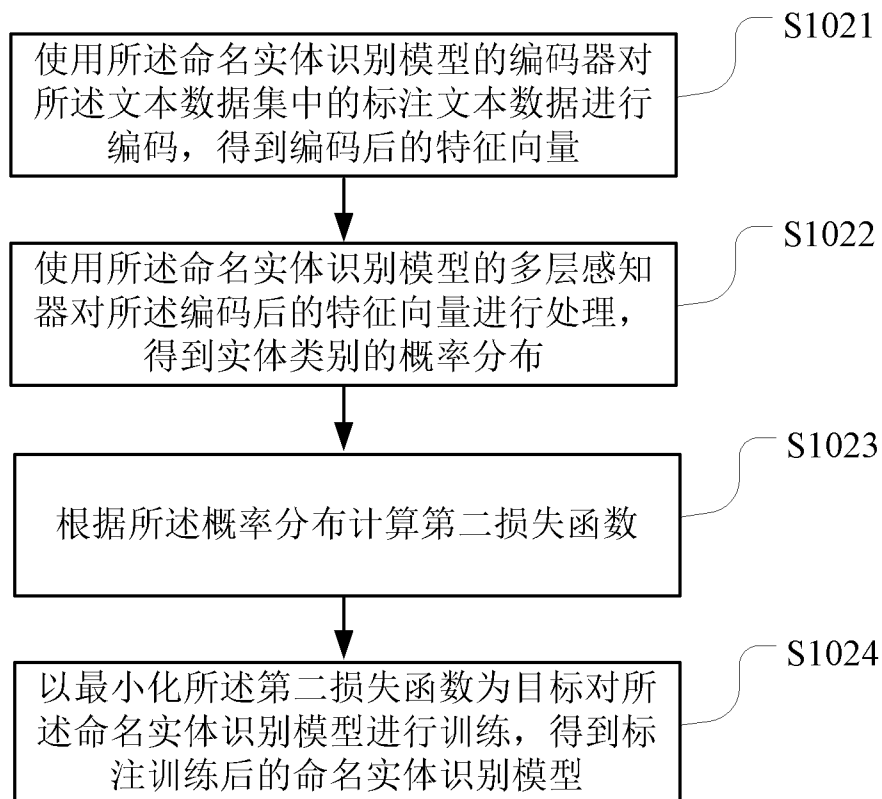


图 3

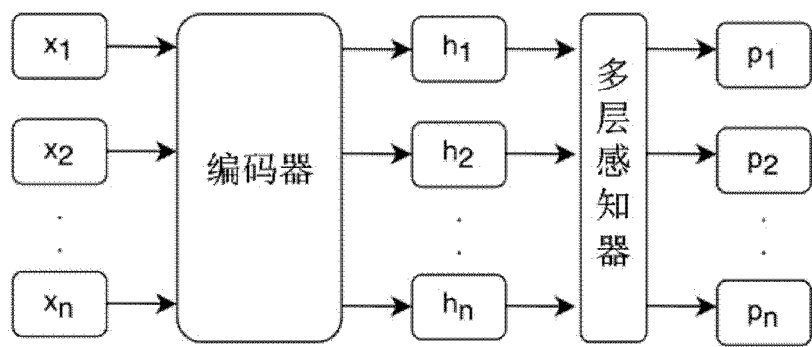


图 4

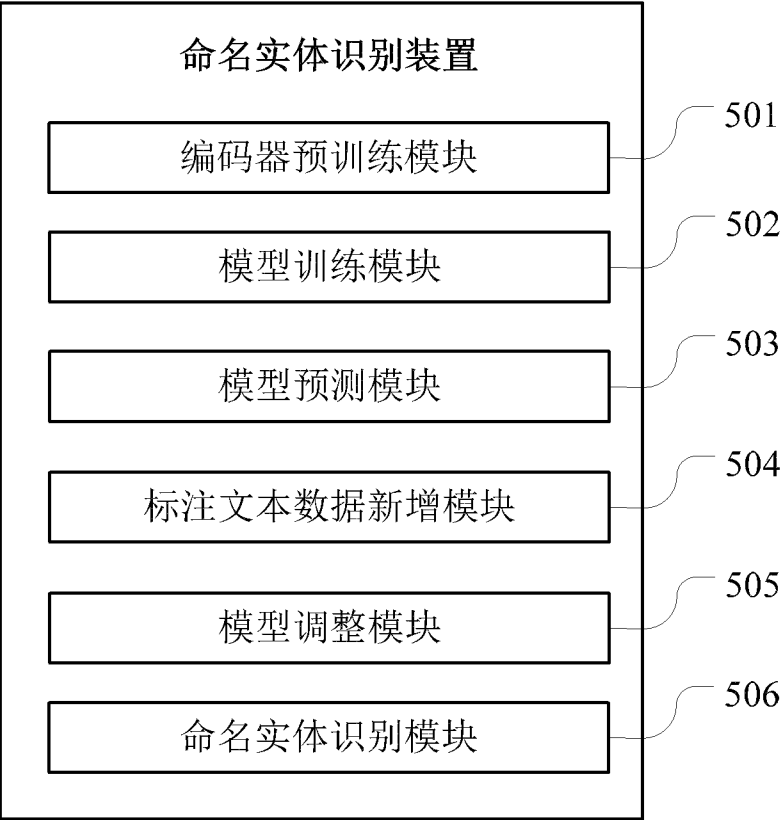


图 5

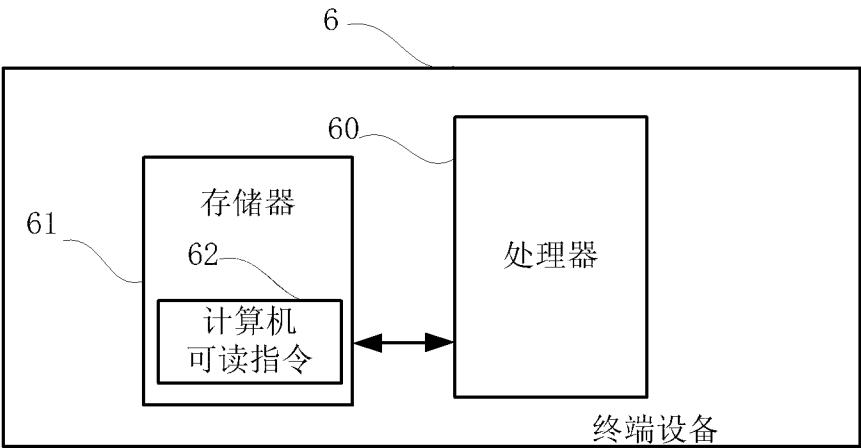


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/089993

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/295(2020.01)i; G06N 3/04(2006.01)i; 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)

G06F; 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)

CNPAT, CNKI, BAIDU, EPODOC, WPI, IEEE: 命名, 实体, 识别, 模型, 预设, 孪生网络, 编码器, 预训练, 标注, 人工标注, 无标注, 新增, 扩充, 文本, 增强, 损失函数, 预测, 置信度, naming, entity, identification, model, preset, siamese network, encoder, pre-train, label, human, add, augment, text, loss function, forecast, confidence

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 113901823 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 07 January 2022 (2022-01-07) claims 1-10	1-20
Y	CN 108959252 A (NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY OF PLA) 07 December 2018 (2018-12-07) description, paragraphs [0007]-[0063], and figures 1-4	1-20
Y	CN 112818691 A (BEIJING KINGSOFT DIGITAL ENTERTAINMENT TECHNOLOGY CO., LTD.) 18 May 2021 (2021-05-18) description, paragraphs [0041]-[0227], and figures 1-7	1-20
Y	weixin_39874196. "A twin network image similarity_twin network is used for carrying out meta-learning by using a double-head neural network" https://blog.csdn.net/weixin_39874196/article/details/110351766 , 20 November 2020 (2020-11-20), text, pp. 1-4	1-20
A	CN 109543009 A (LONGMA ZHIXIN (ZHUHAI HENGQIN) TECHNOLOGY CO., LTD.) 29 March 2019 (2019-03-29) 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

15 July 2022

Date of mailing of the international search report

27 July 2022

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/CN2022/089993

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2021149993 A1 (INTUIT INC.) 20 May 2021 (2021-05-20) entire document	1-20

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/089993

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	113901823	A	07 January 2022	None			
CN	108959252	A	07 December 2018	None			
CN	112818691	A	18 May 2021	None			
CN	109543009	A	29 March 2019	None			
US	2021149993	A1	20 May 2021	WO	2021096571	A1	20 May 2021
				CA	3129721	A1	20 May 2021
				AU	2020382717	A1	02 September 2021

国际检索报告

国际申请号

PCT/CN2022/089993

A. 主题的分类 G06F 40/295(2020.01)i; G06N 3/04(2006.01)i; 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) G06F; 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)) CNPAT, CNKI, BAIDU, EPDOC, WPI, IEEE: 命名, 实体, 识别, 模型, 预设, 孪生网络, 编码器, 预训练, 标注, 人工标注, 无标注, 新增, 扩充, 文本, 增强, 损失函数, 预测, 置信度, naming, entity, identification, model, preset, siamese network, encoder, pre-train, label, human, add, augment, text, loss function, forecast, confidence																							
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 113901823 A (平安科技深圳有限公司) 2022年1月7日 (2022 - 01 - 07) 权利要求1-10</td> <td>1-20</td> </tr> <tr> <td>Y</td> <td>CN 108959252 A (中国人民解放军国防科技大学) 2018年12月7日 (2018 - 12 - 07) 说明书第[0007]-[0063]段, 附图1-4</td> <td>1-20</td> </tr> <tr> <td>Y</td> <td>CN 112818691 A (北京金山数字娱乐科技有限公司) 2021年5月18日 (2021 - 05 - 18) 说明书第[0041]-[0227]段, 附图1-7</td> <td>1-20</td> </tr> <tr> <td>Y</td> <td>weixin_39874196. "孪生网络图像相似度_孪生网络: 使用双头神经网络进行元学习" https://blog.csdn.net/weixin_39874196/article/details/110351766, 2020年11月20日 (2020 - 11 - 20), 正文第1-4页</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>CN 109543009 A (龙马智芯珠海横琴科技有限公司) 2019年3月29日 (2019 - 03 - 29) 全文</td> <td>1-20</td> </tr> <tr> <td>A</td> <td>US 2021149993 A1 (INTUIT INC.) 2021年5月20日 (2021 - 05 - 20) 全文</td> <td>1-20</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 113901823 A (平安科技深圳有限公司) 2022年1月7日 (2022 - 01 - 07) 权利要求1-10	1-20	Y	CN 108959252 A (中国人民解放军国防科技大学) 2018年12月7日 (2018 - 12 - 07) 说明书第[0007]-[0063]段, 附图1-4	1-20	Y	CN 112818691 A (北京金山数字娱乐科技有限公司) 2021年5月18日 (2021 - 05 - 18) 说明书第[0041]-[0227]段, 附图1-7	1-20	Y	weixin_39874196. "孪生网络图像相似度_孪生网络: 使用双头神经网络进行元学习" https://blog.csdn.net/weixin_39874196/article/details/110351766 , 2020年11月20日 (2020 - 11 - 20), 正文第1-4页	1-20	A	CN 109543009 A (龙马智芯珠海横琴科技有限公司) 2019年3月29日 (2019 - 03 - 29) 全文	1-20	A	US 2021149993 A1 (INTUIT INC.) 2021年5月20日 (2021 - 05 - 20) 全文	1-20
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																					
PX	CN 113901823 A (平安科技深圳有限公司) 2022年1月7日 (2022 - 01 - 07) 权利要求1-10	1-20																					
Y	CN 108959252 A (中国人民解放军国防科技大学) 2018年12月7日 (2018 - 12 - 07) 说明书第[0007]-[0063]段, 附图1-4	1-20																					
Y	CN 112818691 A (北京金山数字娱乐科技有限公司) 2021年5月18日 (2021 - 05 - 18) 说明书第[0041]-[0227]段, 附图1-7	1-20																					
Y	weixin_39874196. "孪生网络图像相似度_孪生网络: 使用双头神经网络进行元学习" https://blog.csdn.net/weixin_39874196/article/details/110351766 , 2020年11月20日 (2020 - 11 - 20), 正文第1-4页	1-20																					
A	CN 109543009 A (龙马智芯珠海横琴科技有限公司) 2019年3月29日 (2019 - 03 - 29) 全文	1-20																					
A	US 2021149993 A1 (INTUIT INC.) 2021年5月20日 (2021 - 05 - 20) 全文	1-20																					
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																							
<table border="0"> <tr> <td> * 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 </td> <td> "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 </td> </tr> </table>			* 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																			
* 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																					
2022年7月15日		2022年7月27日																					
ISA/CN的名称和邮寄地址		受权官员																					
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		徐淑娴 电话号码 86-(10)-53961535																					

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/089993

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	113901823	A	2022年1月7日	无			
CN	108959252	A	2018年12月7日	无			
CN	112818691	A	2021年5月18日	无			
CN	109543009	A	2019年3月29日	无			
US	2021149993	A1	2021年5月20日	WO	2021096571	A1	2021年5月20日
				CA	3129721	A1	2021年5月20日
				AU	2020382717	A1	2021年9月2日