

(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 12 月 29 日 (29.12.2022)



(10) 国际公布号
WO 2022/267453 A1

(51) 国际专利分类号:
G06F 16/33 (2019.01)

(21) 国际申请号: PCT/CN2022/071360

(22) 国际申请日: 2022 年 1 月 11 日 (11.01.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110704690.9 2021年6月24日 (24.06.2021) 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: 陈凯(CHEN, Kai); 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)。 徐冰(XU, Bing); 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)。 汪伟(WANG, Wei); 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: 深圳中一联合知识产权代理有限公司(SHENZHEN ZHONGYI UNION INTELLECTUAL PROPERTY AGENCY CO.,LTD.);
中国广东省深圳市福田区莲花街道紫荆社区深南大道 6008 号深圳特区报业大厦 33 层, Guangdong 518034 (CN)。

(54) Title: METHOD FOR TRAINING KEY INFORMATION EXTRACTION MODEL, AND EXTRACTION METHOD, DEVICE AND MEDIUM

(54) 发明名称: 关键信息提取模型的训练方法、提取方法、设备及介质

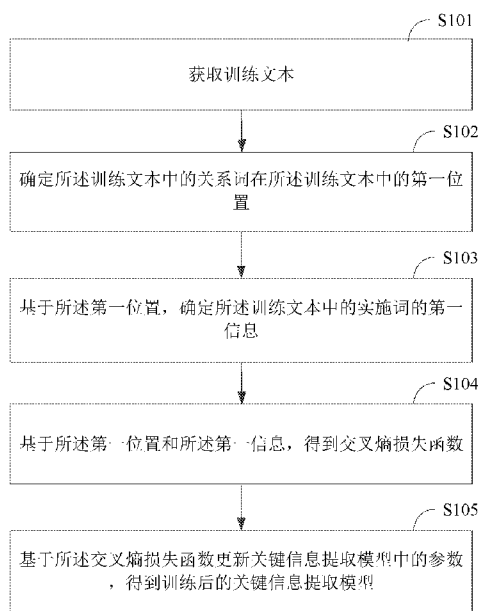


图 2

S101 Acquire training text
S102 Determine, in the training text, a first location of a relative word in the training text
S103 Determine first information of an implementation word in the training text on the basis of the first location
S104 Obtain a cross-entropy loss function on the basis of the first location and the first information
S105 Update parameters in a key information extraction model on the basis of the cross-entropy loss function, so as to obtain a trained key information extraction model

(57) Abstract: A method for training a key information extraction model, and an extraction method, a device and a medium. The method comprises: acquiring training text (S101); determining, in the training text, a first location of a relative word in the training text (S102); determining first information of an implementation word in the training text on the basis of the first location (S103); obtaining a cross-entropy loss function on the basis of the first location and the first information (S104); and updating parameters in a key information extraction model on the basis of the cross-entropy loss function, so as to obtain a trained key information extraction model

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

(S105). By means of the method, a cross-entropy loss function is determined by using a plurality of parameters, such that an obtained trained key information extraction model can be more accurate, and key information extracted by the key information extraction model is more accurate.

(57) 摘要: 一种关键信息提取模型的训练方法、提取方法、设备及介质, 该方法包括: 获取训练文本(S101), 确定训练文本中的关系词在训练文本中的第一位置(S102); 基于第一位置, 确定训练文本中的实施词的第一信息(S103), 基于第一位置和第一信息, 得到交叉熵损失函数(S104), 基于交叉熵损失函数更新关键信息提取模型中的参数, 得到训练后的关键信息提取模型(S105)。该方法利用多个参数确定交叉熵损失函数可以使得到的训练后的关键信息提取模型更准确, 使关键信息提取模型提取的关键信息更准确。

关键信息提取模型的训练方法、提取方法、设备及介质

[0001] 本申请要求于2021年06月24日在中国专利局提交的、申请号为202110704690.9、发明名称为“关键信息提取模型的训练方法、提取方法、设备及介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

[0002] 本申请涉及机器学习的数据处理技术领域，具体涉及一种关键信息提取模型的训练方法、关键信息的提取方法、终端设备及计算机可读存储介质。

背景技术

[0003] 文本的关键信息可以反映文本的主要内容，准确查找文本的关键信息可以准确反映文本的内容。

[0004] 目前，提取文本的关键信息多采用深度学习模型，通过深度学习模型提取文本的关键信息。但是发明人意识到目前使用深度学习模型提取的关键信息存在不够准确的缺点，因此，如何提高深度学习模型的准确度是目前需要解决的问题。

发明概述

技术问题

[0005] 本申请实施例的目的之一在于：提供一种关键信息提取模型的训练方法、关键信息的提取方法、终端设备及计算机可读存储介质。

问题的解决方案

技术解决方案

[0006] 本申请实施例采用的技术方案是：

[0007] 第一方面，提供了一种关键信息提取模型的训练方法，包括：

[0008] 获取训练文本；

[0009] 确定所述训练文本中的关系词在所述训练文本中的第一位置；

[0010] 基于所述第一位置，确定所述训练文本中的实施词的第一信息，其中，所述实施词为与所述关系词存在第一预设关系的词和/或与所述关系词存在第二预设关

系的词，所述第一信息包括所述实施词在所述训练文本中的位置及所述实施词与所述关系词的关系；

[0011] 基于所述第一位置和所述第一信息，得到交叉熵损失函数；

[0012] 基于所述交叉熵损失函数更新所述关键信息提取模型中的参数，得到训练后的关键信息提取模型。

[0013] 第二方面，提供了一种关键信息的提取方法，应用于上述第一方面所述关键信息提取模型的训练方法得到的训练后的关键信息提取模型，包括：

[0014] 获取待处理文本；

[0015] 基于所述训练后的关键信息提取模型，得到所述待处理文本的关键信息。

[0016] 第三方面，提供一种关键信息的提取装置，包括：

[0017] 文本获取模块，用于获取待处理文本；

[0018] 结果输出模块，用于基于所述训练后的关键信息提取模型，得到所述待处理文本的关键信息，其中，所述训练后的关键信息提取模型为上述第一方面所述关键信息提取模型的训练方法得到的模型。

[0019] 第四方面，提供一种终端设备，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计算机程序，其中，所述处理器执行所述计算机程序时实现上述第一方面中任一项所述的关键信息提取模型的训练方法。

[0020] 第五方面，提供一种终端设备，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计算机程序，其中，所述处理器执行所述计算机程序时实现上述第二方面中任一项所述的关键信息的提取方法。

[0021] 第六方面，提供了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计算机程序，其中，所述计算机程序被处理器执行时实现上述第一方面中任一项所述的关键信息提取模型的训练方法。

[0022] 第七方面，提供了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计算机程序，其中，所述计算机程序被处理器执行时实现上述第二方面中任一项所述的关键信息的提取方法。

[0023] 第八方面，提供了一种计算机程序产品，当计算机程序产品在终端设备上运行时，使得终端设备执行上述第一方面中任一项所述的关键信息提取模型的训练

方法，和/或实现上述第二方面中任一项所述的关键信息的提取方法。

发明的有益效果

有益效果

[0024] 本申请实施例提供的关键信息提取模型的训练方法的有益效果在于：获取训练文本；确定所述训练文本中的关系词在所述训练文本中的第一位置；基于所述第一位置，确定所述训练文本中的实施词的第一信息，其中，所述实施词为与所述关系词存在第一预设关系的词和/或与所述关系词存在第二预设关系的词，所述第一信息包括所述实施词在所述训练文本中的位置及所述实施词与所述关系词的关系；基于所述第一位置和所述第一信息，得到交叉熵损失函数；基于所述交叉熵损失函数更新所述关键信息提取模型中的参数，得到训练后的关键信息提取模型。

[0025] 本申请实施例提供的关键信息的提取方法的有益效果在于：获取待处理文本；基于所述训练后的关键信息提取模型，得到所述待处理文本的关键信息。

对附图的简要说明

附图说明

[0026] 为了更清楚地说明本申请实施例中的技术方案，下面将对实施例或示范性技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它的附图。

[0027] 图1是本申请一实施例提供的关键信息提取模型的训练方法的应用场景示意图；

[0028] 图2是本申请一实施例提供的关键信息提取模型的训练方法的流程示意图；

[0029] 图3是本申请一实施例提供的实施词的第一信息的确定方法的流程示意图；

[0030] 图4是本申请一实施例提供的关系结果张量的示意图；

[0031] 图5是本申请一实施例提供的交叉熵损失函数的确定方法的流程示意图；

[0032] 图6是本申请一实施例提供的关键信息的提取方法的流程示意图；

[0033] 图7是本申请一实施例提供的终端设备的结构示意图。

发明实施例

本发明的实施方式

[0034] 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本申请，并不用于限定本申请。

[0035] 术语“第一”、“第二”仅用于便于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明技术特征的数量。“多个”的含义是两个或两个以上，除非另有明确具体的限定。

[0036] 为了说明本申请所提供的技术方案，以下结合具体附图及实施例进行详细说明。

[0037] 图1为本申请实施例提供的关键信息提取模型的训练方法的应用场景示意图，上述关键信息提取模型的训练方法可以用于对模型进行训练。其中，存储设备10用于存储文本，电子设备20用于从存储设备10中获取文本，并基于获取的文本对待训练的关键信息提取模型进行训练，得到训练后的关键信息提取模型。

[0038] 以下结合图1对本申请实施例的关键信息提取模型的训练方法进行详细说明。

[0039] 图2示出了本申请提供的关键信息提取模型的训练方法的示意性流程图，参照图2，对该方法的详述如下：

[0040] S101，获取训练文本。

[0041] 在本实施例中，待训练的关键信息提取模型在获取训练样本时可以从存储设备中获得，还可以从数据库中获得，训练文本可以是一句话或一段话。训练文本还可以是文章题目等。训练文本还可以是由政府或监管机构发布的政策文本。

[0042] 在本实施例中，待训练的关键信息提取模型可以是BERT模型（Bidirectional Encoder Representation from Transformers）。

[0043] S102，确定所述训练文本中的关系词在所述训练文本中的第一位置。

[0044] 在本实施例中，关系词可以是训练文本中词性为动词的词。

[0045] 具体的，在获取到训练文本后，先查找训练文本中是否存在动词，如果训练文本中存在动词，可以将动词的位置找出，本申请中记为第一位置。

[0046] 具体的，对训练文本的处理，实际是对训练文本的文本向量进行处理。文本向

量由训练文本中各个字符的向量组成。因此，在获取到训练文本后，可以先确定训练文本中各个字符的向量。然后在文本向量中确定关系词的第一位置。

[0047] 可选的，关系词在训练文本中的第一位置可以从待训练的关键信息提取模型中的Dense层（全连层）得到。

[0048] S103，基于所述第一位置，确定所述训练文本中的实施词的第一信息，其中，所述实施词为与所述关系词存在第一预设关系的词和/或与所述关系词存在第二预设关系的词，所述第一信息包括所述实施词在所述训练文本中的第二位置及所述实施词与所述关系词的关系。

[0049] 在本实施例，在关系词为动词时，实施词为训练文本中与关系词存在主谓关系的词，实施词还可以为训练文本中与关系词存在动宾关系的词。

[0050] 具体的，在确定了关系词的第一位置，且关系词为动词时，可以确定训练文本中是否存在动词的主语，如果训练文本中存在动词的主语，则可以提取该主语及主语的位置，本申请中记为第二位置，并输出该主语与动词为主谓关系。例如，如果训练文本为“我吃苹果”，吃是关系词，我是实施词，则可以得到“我-主谓关系-吃”。

[0051] 具体的，在确定了关系词的第一位置，且关系词为动词时，可以确定训练文本中是否存在动词的宾语，如果训练文本中存在动词的宾语，则可以提取该宾语和该宾语的位置，本申请中记为第二位置，并输出该宾语与动词为动宾关系。例如，如果训练文本为“从严落实疫情防控责任”，落实是关系词，疫情防控责任是实施词，则可以得到“落实-动宾关系-疫情防控责任”。即使在训练文本中不存在主语时，也可以提取出与动词相关的宾语词。

[0052] 本申请实施例中，可以从训练文本中提取两种关系，一种是动宾关系，一种是主谓关系。

[0053] 可选的，可以采用待训练的关键信息提取模型中的LSTM层（Long Short-Term Memory-长短期记忆网络）和Dense层得到第一信息。

[0054] S104，基于所述第一位置和所述第一信息，得到交叉熵损失函数。

[0055] 在本实施例，可以通过第一位置和第一信息得到两个交叉熵函数，基于上述得到的两个交叉熵函数得到交叉熵损失函数。

- [0056] S105, 基于所述交叉熵损失函数更新关键信息提取模型中的参数, 得到训练后的关键信息提取模型。
- [0057] 在本实施例中, 在得到交叉熵损失函数后, 可以利用交叉熵损失函数更新关键信息提取模型中的参数, 继续对关键信息提取模型进行训练, 直到交叉熵损失函数满足预设要求, 则可以停止对关键信息提取模型的训练, 得到训练后的关键信息提取模型。
- [0058] 本申请实施例中, 首先获取训练文本, 并确定训练文本中关系词在训练文本中的第一位置, 基于第一位置, 确定训练文本中的实施词的第一信息, 其中, 实施词为与关系词存在第一预设关系的词和/或与关系词存在第二预设关系的词, 第一信息包括实施词在训练文本中的位置及实施词与关系词的关系, 基于第一位置和第一信息, 得到交叉熵损失函数, 基于交叉熵损失函数更新关键信息提取模型中的参数, 得到训练后的关键信息提取模型, 本申请利用关系词的第一位置、实施词的第二位置和实施词与关系词的关系确定交叉熵损失函数, 利用交叉熵损失函数更新待训练的关键信息提取模型中的参数, 得到训练后的关键信息提取模型, 本申请利用多个参数确定交叉熵损失函数可以使得到的训练后的关键信息提取模型更准确, 使关键信息提取模型提取的关键信息更准确。
- [0059] 在一种可能的实现方式中, 步骤S102的实现过程可以包括:
- [0060] S1021, 确定所述关系词在所述训练文本中的头位置和尾位置, 所述第一位置包括所述头位置和所述尾位置, 其中, 所述关系词为词性为动词的词。
- [0061] 可选的, 得到训练文本的文本向量后, 可以对文本向量进行BIO标注 (B-begin, I-inside, O-outside), 将关系词的首位置的向量用B标注, 关系词的尾位置的向量用I标注, 训练文本中的非关系词的向量用O标注。
- [0062] 具体的, 确定关系词的第一位置可以是确定关系词的头位置的向量和尾位置的向量。另外, 如果关系词至少包括三个字符, 还可以确定关系词中每个字符的向量。第一位置可以包括关系词中每个字符的向量。
- [0063] 本申请实施例中, 确定关系词的头位置的向量和尾位置的向量, 可以减少后续第一信息确定时对数据的处理, 加快模型的训练速度。
- [0064] 如图3所示, 在一种可能的实现方式中, 步骤S103的实现过程可以包括:

- [0065] S1031, 计算所述头位置的向量和所述尾位置的向量的平均向量。
- [0066] 在本实施例中, 如果第一位置中包括关系词中每个字符的向量, 可以计算每个字符的向量的平均值, 将该平均值作为平均向量。平均向量用于表征关系词的向量。
- [0067] 在本实施例中, 在得到头位置的向量和尾位置的向量后, 可以将头位置的向量和尾位置的向量回传, 以确定第一信息。
- [0068] S1032, 基于所述平均向量, 确定所述实施词的第一信息。
- [0069] 在本实施例中, 在得到平均向量后, 可以将平均向量加入到文本向量中每个字符的向量中, 然后确定第一信息。
- [0070] 在一种可能的实现方式中, 在步骤S1032之后, 上述方法还可以包括:
- [0071] 基于所述实施词的第一信息, 得到所述实施词的关系结果张量。
- [0072] 在本实施例中, 在得到实施词的第一信息后, 可以根据第一信息确定实施词的关系结果张量, 关系结果张量用于表征实施词的头位置在训练文本中的位置、实施词的尾位置在训练文本中的位置、实施词与关系词的关系。
- [0073] 作为举例, 如图4所示的关系结果张量图, 前两排为训练文本中与关系词存在主谓关系的实施词的位置, 第一排中1的位置表征实施词的头在训练文本中的位置。第一排中1的位置表征实施词的尾在训练文本中的位置。后两排为训练文本中与关系词存在动宾关系的实施词的位置。图4中动宾关系的两排中没有标注为1的位置, 则说明训练文本中不存在与关系词是动宾关系的实施词。
- [0074] 本申请实施例中, 通过关系结果张量表征实施词的信息, 可以使信息处理更方便, 在张量表中可以将实施词的位置和与关系词的关系进行表示, 减少了数据处理量, 提高了数据处理的速度。
- [0075] 如图5所示, 在一种可能的实现方式中, 步骤S104的实现过程可以包括:
- [0076] S1041, 基于所述头位置的向量, 得到第一交叉熵函数。
- [0077] 在本实施例中, 在训练文本输入待训练的关键信息提取模型之前可以对训练文本进行标注, 将关系词的头位置和尾位置标出, 本申请中可以将标出的关系词的头位置记作第一目标位置, 将标出的关系词的尾位置即为第二目标位置。另外, 还可以将训练文本中与关系词存在第一预设关系的词的位置标出, 本申请

中将标出的与关系词存在第一预设关系的实施词的位置记为第三目标位置，和/或将训练文本中与关系词存在第二预设关系的实施词的位置标出，本申请中将标出的与关系词存在第二预设关系的词的位置记为第四目标位置。

[0078] 在本实施例中，由于标出了关系词的第一目标位置，因此可以确定训练文本中关系词的头位置的真正向量。可以根据关系词的头位置的真正向量和关键信息提取模型输出的关系词的头位置的向量得到第一交叉熵函数。

[0079] S1042，基于所述尾位置的向量，得到第二交叉熵函数。

[0080] 在本实施例中，由于标出了关系词的第二目标位置，因此可以确定训练文本中关系词的尾位置的真正向量。因此可以根据关系词的尾位置的真正向量和关键信息提取模型输出的关系词的尾位置的向量得到第二交叉熵函数。

[0081] S1043，基于所述关系结果张量，得到第三交叉熵函数。

[0082] 在本实施例中，由于标出了实施词的第三目标位置和/或第四目标位置，在训练文本中存在第三目标位置，可以根据第三目标位置和第三目标位置的实施词与关系词的关系得到一真实关系张量。在训练文本中存在第四目标位置，可以根据第四目标位置和第四目标位置的实施词与关系词的关系得到一真实关系张量。在训练文本中存在第三目标位置和第四目标位置时，可以根据第三目标位置和第三目标位置的实施词与关系词的关系、第四目标位置和第四目标位置的实施词与关系词的关系得到一真实关系张量。

[0083] 在本实施例中，根据真实关系张量和关键信息提取模型输出的关系结果张量得到第三交叉熵函数。

[0084] S1044，基于所述第一交叉熵函数、所述第二交叉熵函数和所述第三交叉熵函数，得到所述交叉熵损失函数。

[0085] 在本实施例中，可以将第一交叉熵函数、第二交叉熵函数和第三交叉熵函数之和作为交叉熵损失函数。

[0086] 可选的，还可以第一交叉熵函数、第一交叉熵函数的权重、第二交叉熵函数、第二交叉熵函数的权重、第三交叉熵函数和第三交叉熵函数的权重得到交叉熵损失函数。

[0087] 本申请实施例中，根据第一交叉熵函数、第二交叉熵函数和第三交叉熵函数，

得到交叉熵损失函数，为关键信息提取模型中参数的更新提供了多源依据，可以使更新的参数更准确，进而使关键信息提取模型提取的文本的关键信息更准确。

[0088] 对应于上文实施例所述的关键信息提取模型的训练方法，图6示出了本申请实施例提供的关键信息的提取方法，参照图6，该提取方法包括：

[0089] S201，获取待处理文本。

[0090] S202，基于上述训练后的关键信息提取模型，得到所述待处理文本的关键信息。

[0091] 在本实施例中，训练后的关键信息提取模型为基于上述关键信息提取模型的训练方法训练得到的模型。

[0092] 在本实施例中，关键信息可以包括关键词以及关键词之间的关系。

[0093] 具体的，步骤S202的实现过程可以包括：

[0094] S2021，提取所述待处理文本中的关系词及所述关系词在所述待处理文本中的位置。

[0095] 在本实施例中，关系词在待处理文本中的位置可以包括关系词的头位置的向量和关系词的尾位置的向量。

[0096] S2022，基于所述关系词在所述待处理文本中的位置，得到所述待处理文本中的实施词及所述实施词与所述关系词的关系，所述关键信息包括所述关系词、所述实施词和所述实施词与所述关系词的关系。

[0097] 在本实施例中，计算关系词的头位置的向量和关系词的尾位置的向量的平均向量，基于平均向量确定实施词及实施词与关系词的关系。

[0098] 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，而不应对本申请实施例的实施过程构成任何限定。

[0099] 对应于上文实施例所述的关键信息提取模型的训练方法，本申请实施例提供了关键信息的提取装置，包括：

[0100] 文本获取模块，用于获取待处理文本；

[0101] 结果输出模块，用于基于所述训练后的关键信息提取模型，得到所述待处理文

本的关键信息，其中，所述训练后的关键信息提取模型为上述关键信息提取模型的训练方法得到的模型。

[0102] 在一种可能的实现方式中，结果输出模块还用于：

[0103] 提取所述待处理文本中的关系词及所述关系词在所述待处理文本中的位置；

[0104] 基于所述关系词在所述待处理文本中的位置，得到所述待处理文本中的实施词及所述实施词与所述关系词的关系，所述关键信息包括所述关系词、所述实施词和所述实施词与所述关系词的关系。

[0105] 在一种可能的实现方式中，所述关系词为词性为动词的词。

[0106] 在一种可能的实现方式中，所述实施词与所述关系词的关系包括主谓关系和/或动宾关系。

[0107] 本申请实施例还提供了一种终端设备，参见图7，该终端设备400可以包括：至少一个处理器410、存储器420以及存储在所述存储器420中并可在所述至少一个处理器410上运行的计算机程序，所述处理器410执行所述计算机程序时实现上述任意各个方法实施例中的步骤，例如图2所示实施例中的步骤S101至步骤S105或图6所示实施例中的步骤S201至步骤S202。或者，处理器410执行所述计算机程序时实现上述各装置实施例中各模块/单元的功能。

[0108] 示例性的，计算机程序可以被分割成一个或多个模块/单元，一个或者多个模块/单元被存储在存储器420中，并由处理器410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程序段，该程序段用于描述计算机程序在终端设备400中的执行过程。

[0109] 本领域技术人员可以理解，图7仅仅是终端设备的示例，并不构成对终端设备的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件，例如输入输出设备、网络接入设备、总线等。

[0110] 处理器410可以是中央处理单元(Central Processing Unit, CPU)，还可以是其他通用处理器、数字信号处理器(Digital Signal Processor, DSP)、专用集成电路(Application Specific Integrated Circuit, ASIC)、现成可编程门阵列(Field-Programmable Gate Array, FPGA)或者其他可编程逻辑器件、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。

[0111] 存储器420可以是终端设备的内部存储单元，也可以是终端设备的外部存储设备，例如插接式硬盘，智能存储卡，安全数字（Secure Digital，SD）卡，闪存卡（Flash Card）等。所述存储器420用于存储所述计算机程序以及终端设备所需的其它程序和数据。所述存储器420还可以用于暂时地存储已经输出或者将要输出的数据。

[0112] 总线可以是工业标准体系结构（Industry Standard Architecture，ISA）总线、外部设备互连总线或扩展工业标准体系结构（Extended Industry Standard Architecture，EISA）总线等。总线可以分为地址总线、数据总线、控制总线等。为便于表示，本申请附图中的总线并限定仅有一根总线或一种类型的总线。

[0113] 本申请实施例提供的关键信息提取模型的训练方法和关键信息的提取方法可以应用于计算机、平板电脑、笔记本电脑、上网本、个人数字助理（personal digital assistant，PDA）等终端设备上，本申请实施例对终端设备的具体类型不作任何限制。

[0114] 本申请实施例还提供了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计算机程序，所述计算机程序被处理器执行时实现可实现上述关键信息提取模型的训练方法或关键信息的提取方法各个实施例中的步骤。

[0115] 本申请实施例提供了一种计算机程序产品，当计算机程序产品在移动终端上运行时，使得移动终端执行时实现可实现上述关键信息提取模型的训练方法或关键信息的提取方法各个实施例中的步骤。

[0116] 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时，可以存储在一个计算机可读存储介质中。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。基于这样的理解，本申请实现上述实施例方法中的全部或部分流程，可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一计算机可读存储介质中，该计算机程序在被处理器执行时，可实现上述各个方法实施例的步骤。其中，所述计算机程序包括计算机程序代码，所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件

或某些中间形式等。所述计算机可读介质至少可以包括：能够将计算机程序代码携带到拍照装置/终端设备的任何实体或装置、记录介质、计算机存储器、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、电载波信号、电信信号以及软件分发介质。例如U盘、移动硬盘、磁碟或者光盘等。在某些司法管辖区，根据立法和专利实践，计算机可读介质不可以是电载波信号和电信信号。

[0117] 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。

[0118] 本领域普通技术人员可以意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。

[0119] 在本申请所提供的实施例中，应该理解到，所揭露的装置/网络设备和/或方法，可以通过其它的方式实现。例如，以上所描述的装置/网络设备实施例仅仅是示意性的，例如，所述模块或单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通讯连接可以是通过一些接口，装置或单元的间接耦合或通讯连接，可以是电性，机械或其它的形式。

[0120] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0121] 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，均应包含在本申请的保护范围之内。

权利要求书

- [权利要求 1] 一种关键信息提取模型的训练方法，其中，包括：
- 获取训练文本；
- 确定所述训练文本中的关系词在所述训练文本中的第一位置；
- 基于所述第一位置，确定所述训练文本中的实施词的第一信息，其中，所述实施词为与所述关系词存在第一预设关系的词和/或与所述关系词存在第二预设关系的词，所述第一信息包括所述实施词在所述训练文本中的第二位置及所述实施词与所述关系词的关系；
- 基于所述第一位置和所述第一信息，得到交叉熵损失函数；
- 基于所述交叉熵损失函数更新关键信息提取模型中的参数，得到训练后的关键信息提取模型。
- [权利要求 2] 如权利要求1所述的关键信息提取模型的训练方法，其中，所述确定所述训练文本中的关系词在所述训练文本中的第一位置，包括：
- 确定所述关系词在所述训练文本中的头位置和尾位置，所述第一位置包括所述头位置和所述尾位置，其中，所述关系词为词性为动词的词。
- [权利要求 3] 如权利要求2所述的关键信息提取模型的训练方法，其中，所述确定所述关系词在所述训练文本中的头位置和尾位置，包括：
- 确定所述关系词的头位置的向量和尾位置的向量。
- [权利要求 4] 如权利要求3所述的关键信息提取模型的训练方法，其中，所述基于所述第一位置，确定所述训练文本中的实施词的第一信息，包括：
- 计算所述头位置的向量和所述尾位置的向量的平均向量；
- 基于所述平均向量，确定所述实施词的第一信息，其中，所述第一预设关系主谓关系，和/或与所述第二预设关系为动宾关系。
- [权利要求 5] 如权利要求4所述的关键信息提取模型的训练方法，其中，在所述基于所述平均向量，确定所述实施词的第一信息之后，包括：
- 基于所述实施词的第一信息，得到所述实施词的关系结果张量。
- [权利要求 6] 如权利要求5所述的关键信息提取模型的训练方法，其中，所述基于

所述第一位置和所述第一信息，得到交叉熵损失函数，包括：

基于所述头位置的向量，得到第一交叉熵函数；

基于所述尾位置的向量，得到第二交叉熵函数；

基于所述关系结果张量，得到第三交叉熵函数；

基于所述第一交叉熵函数、所述第二交叉熵函数和所述第三交叉熵函数，得到所述交叉熵损失函数。

[权利要求 7] 如权利要求3所述的关键信息提取模型的训练方法，其中，所述确定所述关系词的头位置的向量和尾位置的向量，包括：

获得所述训练文本的文本向量；

对所述文本向量进行BIO标注，得到所述关系词的头位置的向量和尾位置的向量。

[权利要求 8] 如权利要求6所述的关键信息提取模型的训练方法，其中，所述基于所述第一交叉熵函数、所述第二交叉熵函数和所述第三交叉熵函数，得到所述交叉熵损失函数，包括：

将第一交叉熵函数、第二交叉熵函数和第三交叉熵函数之和作为交叉熵损失函数。

[权利要求 9] 如权利要求6所述的关键信息提取模型的训练方法，其中，所述基于所述第一交叉熵函数、所述第二交叉熵函数和所述第三交叉熵函数，得到所述交叉熵损失函数，包括：

基于第一交叉熵函数、第一交叉熵函数的权重、第二交叉熵函数、第二交叉熵函数的权重、第三交叉熵函数和第三交叉熵函数的权重得到交叉熵损失函数。

[权利要求 10] 如权利要求6所述的关键信息提取模型的训练方法，其中，所述基于所述头位置的向量，得到第一交叉熵函数，包括：

根据所述关系词的头位置的真实向量和所述关键信息提取模型输出的关系词的头位置的向量得到第一交叉熵函数。

[权利要求 11] 一种关键信息的提取方法，其中，应用于上述如权利要求1至10任一项所述关键信息提取模型的训练方法得到的训练后的关键信息提取模

型；

所述提取方法包括：

获取待处理文本；

基于所述训练后的关键信息提取模型，得到所述待处理文本的关键信息。

[权利要求 12] 如权利要求11所述的关键信息的提取方法，其中，所述基于所述训练后的关键信息提取模型，得到所述待处理文本的关键信息，包括：提取所述待处理文本中的关系词及所述关系词在所述待处理文本中的位置；

基于所述关系词在所述待处理文本中的位置，得到所述待处理文本中的实施词及所述实施词与所述关系词的关系，所述关键信息包括所述关系词、所述实施词和所述实施词与所述关系词的关系。

[权利要求 13] 如权利要求11所述的关键信息的提取方法，其中，所述关系词为词性为动词的词。

[权利要求 14] 如权利要求11所述的关键信息的提取方法，其中，所述实施词与所述关系词的关系包括主谓关系和/或动宾关系。

[权利要求 15] 一种关键信息的提取装置，其中，包括：
文本获取模块，用于获取待处理文本；
结果输出模块，用于基于所述训练后的关键信息提取模型，得到所述待处理文本的关键信息，其中，所述训练后的关键信息提取模型为上述如权利要求1至10任一项所述关键信息提取模型的训练方法得到的模型。

[权利要求 16] 如权利要求15所述的关键信息的提取装置，其中，所述结果输出模块还用于：
提取所述待处理文本中的关系词及所述关系词在所述待处理文本中的位置；
基于所述关系词在所述待处理文本中的位置，得到所述待处理文本中的实施词及所述实施词与所述关系词的关系，所述关键信息包括所述

关系词、所述实施词和所述实施词与所述关系词的关系。

- [权利要求 17] 一种终端设备，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计算机程序，其中，所述处理器执行所述计算机程序时实现如权利要求1至10任一项所述的关键信息提取模型的训练方法。
- [权利要求 18] 一种终端设备，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计算机程序，其中，所述处理器执行所述计算机程序时实现如权利要求11至14任一项所述的关键信息的提取方法。
- [权利要求 19] 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计算机程序，其中，所述计算机程序被处理器执行时实现如权利要求1至10任一项所述的关键信息提取模型的训练方法。
- [权利要求 20] 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计算机程序，其中，所述计算机程序被处理器执行时实现如权利要求11至14任一项所述的关键信息的提取方法。

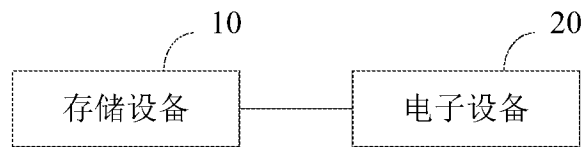


图 1

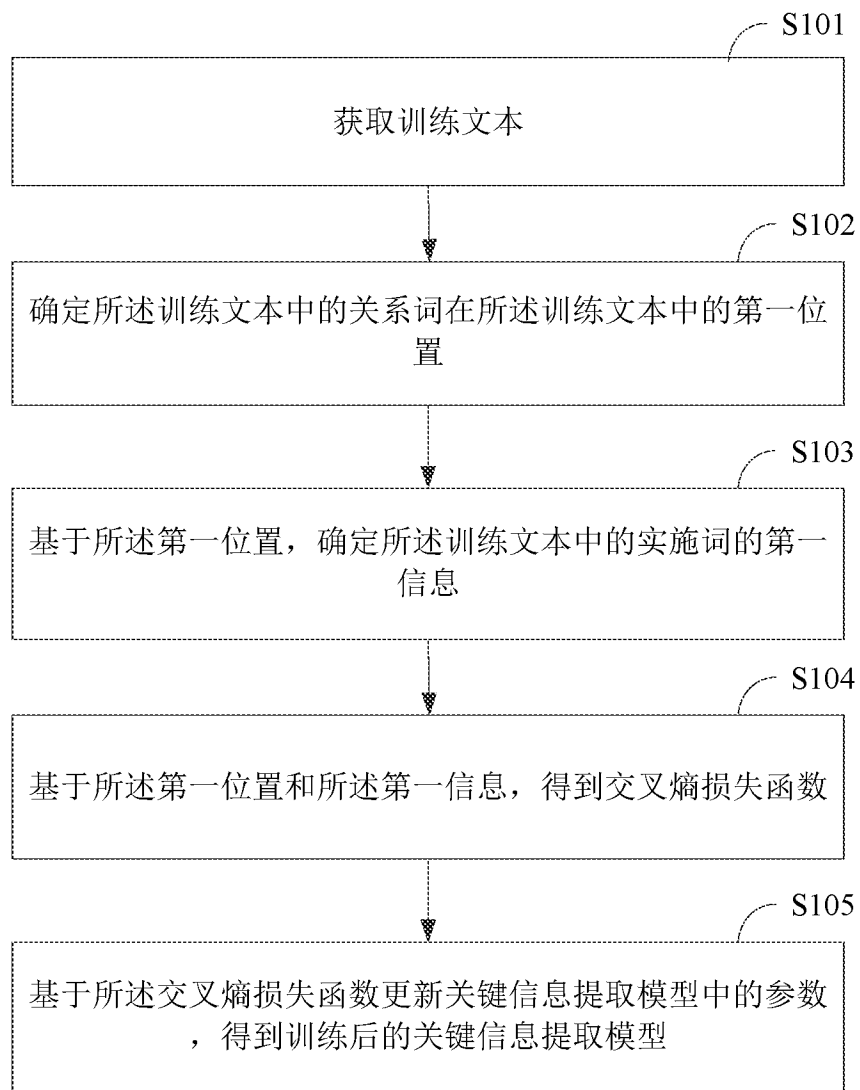


图 2

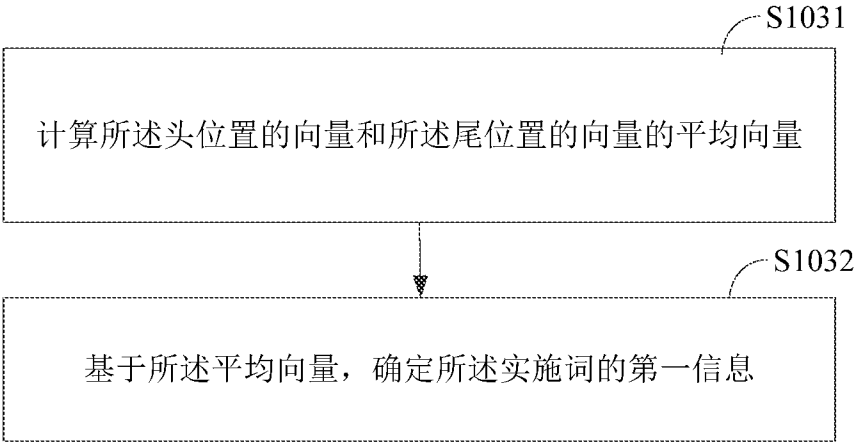


图 3

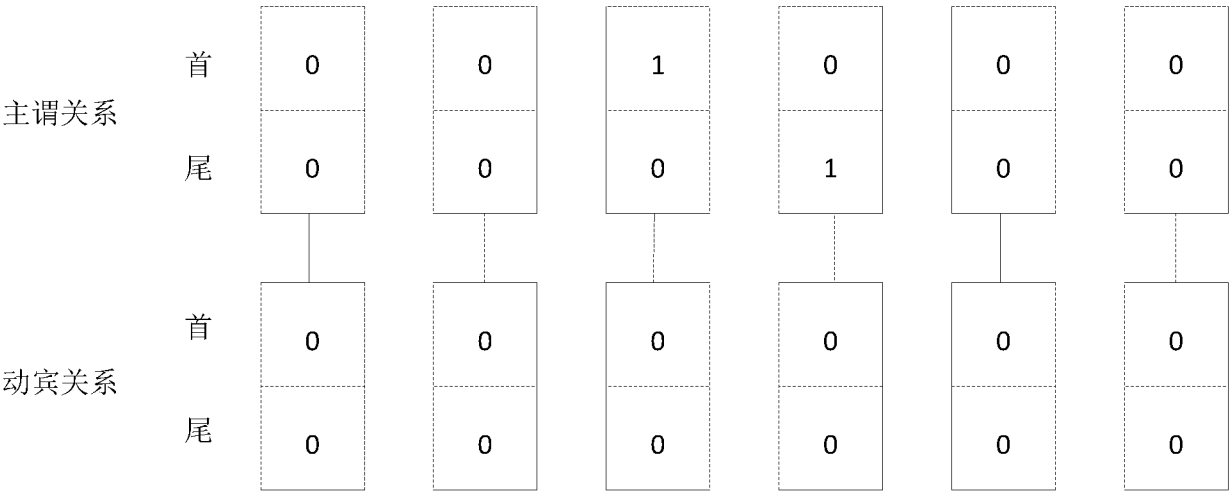


图 4

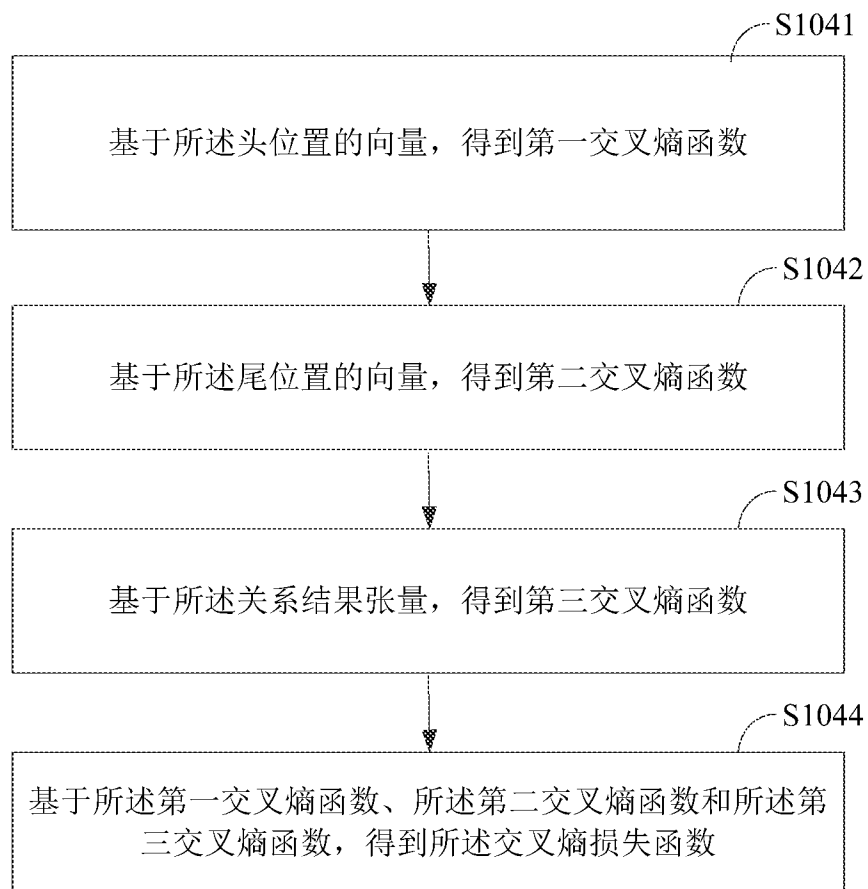


图 5

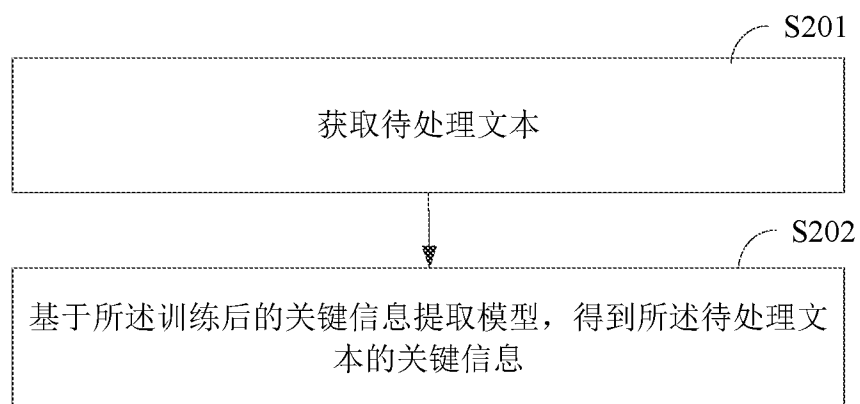


图 6

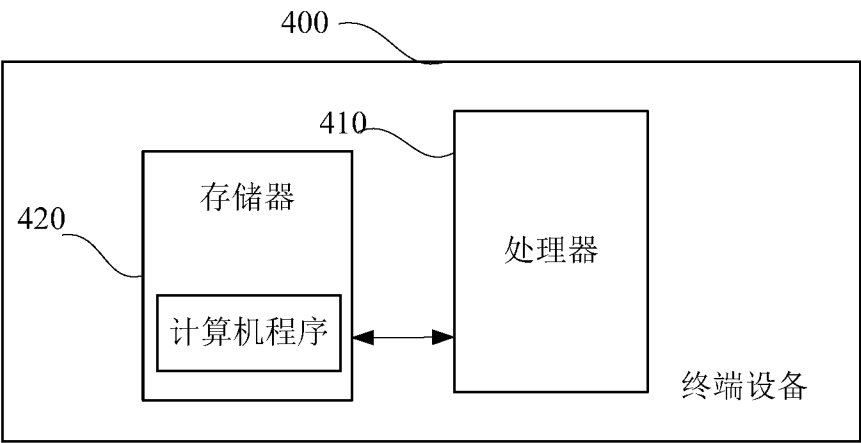


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/071360

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/33(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

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, EPODOC, WPI, IEEE, CNKI: 关键, 动词, 主语, 谓语, 提取, 抽取, 训练, 损失, 交叉熵, 位置, 向量, 权重, 头, 尾, key, verb, subject, predicate, extract, train, loss, cross entropy, position, vector, weight, head, end

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 113420120 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 21 September 2021 (2021-09-21) claims 1-10, and description, paragraphs [0033]-[0105]	1-20
X	CN 111291185 A (BOE TECHNOLOGY GROUP CO., LTD.) 16 June 2020 (2020-06-16) claims 1-10, and description, paragraphs [0050]-[0112]	1-20
A	CN 112270604 A (CHINA MERCHANTS BANK CO., LTD.) 26 January 2021 (2021-01-26) entire document	1-20
A	CN 112270196 A (PERFECT WORLD (BEIJING) SOFTWARE TECHNOLOGY DEVELOPMENT CO., LTD.) 26 January 2021 (2021-01-26) entire document	1-20
A	WO 2020224219 A1 (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 12 November 2020 (2020-11-12) 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

30 March 2022

Date of mailing of the international search report

08 April 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
Information on patent family members

International application No.

PCT/CN2022/071360

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	113420120	A	21 September 2021	None			
CN	111291185	A	16 June 2020	WO	2021147726	A1	29 July 2021
CN	112270604	A	26 January 2021	None			
CN	112270196	A	26 January 2021	None			
WO	2020224219	A1	12 November 2020	CN	110287961	A	27 September 2019

国际检索报告

国际申请号

PCT/CN2022/071360

A. 主题的分类

G06F 16/33 (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

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, EPDOC, WPI, IEEE, CNKI: 关键, 动词, 主语, 谓语, 提取, 抽取, 训练, 损失, 交叉熵, 位置, 向量, 权重, 头, 尾, key, verb, subject, predicate, extract, train, loss, cross entropy, position, vector, weight, head, end

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 113420120 A (平安科技深圳有限公司) 2021年9月21日 (2021 - 09 - 21) 权利要求1-10, 说明书第[0033]-[0105]段	1-20
X	CN 111291185 A (京东方科技集团股份有限公司) 2020年6月16日 (2020 - 06 - 16) 权利要求1-10, 说明书第[0050]-[0112]段	1-20
A	CN 112270604 A (招商银行股份有限公司) 2021年1月26日 (2021 - 01 - 26) 全文	1-20
A	CN 112270196 A (完美世界北京软件科技发展有限公司) 2021年1月26日 (2021 - 01 - 26) 全文	1-20
A	WO 2020224219 A1 (PING AN TECHNOLOGY SHENZHEN CO., LTD.) 2020年11月12日 (2020 - 11 - 12) 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

2022年3月30日

国际检索报告邮寄日期

2022年4月8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

唐娜

电话号码 86-(10)-53961405

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/071360

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	113420120	A	2021年9月21日	无			
CN	111291185	A	2020年6月16日	WO	2021147726	A1	2021年7月29日
CN	112270604	A	2021年1月26日	无			
CN	112270196	A	2021年1月26日	无			
WO	2020224219	A1	2020年11月12日	CN	110287961	A	2019年9月27日