说　　明　　书

信息提取模型的训练方法及系统、信息提取方法及系统

技术领域

本说明书涉及人工智能技术领域，尤其涉及一种信息提取模型的训练方法及系统、信息提取方法及系统。

背景技术

目前，很多场景存在对文档进行信息提取的需求。比如，以风控场景为例，为了有效防控风险交易，风控方需要交易参与方上传能够证明交易真实性的证明文档。上述证明文档包括但不限于工商证明文件、贸易合同以及跨境物流证明等。风控方可以基于上述证明文档进行交易验真（即验证交易的真实性），以便识别出风险交易。其中，在上述交易验真的过程中，需要对证明文档中的内容进行信息提取。

通常，可以预先训练得到信息提取模型，通过该信息提取模型对文档进行信息提取。然而，信息提取模型的训练过程需要大量的经过标注的样本文档。在实际应用中，采用常用语言（例如中文、英文等）描述的样本文档较为容易获得，而采用不常用语言描述的样本文档不容易获得。这使得训练后的信息提取模型只能适用于对常用语言的文档进行信息提取，而不能适用于对不常用语言的文档进行信息提取。

背景技术部分的内容仅仅是发明人个人所知晓的信息，并不代表上述信息在本公开申请日之前已经进入公共领域，也不代表其可以成为本公开的现有技术。

发明内容

本说明书提供一种信息提取模型的训练方法及系统、信息提取方法及系统，能够基

于常用语言描述的样本文档训练得到支持对常用语言和不常用语言的文档进行信息提取的信息提取模型。

为了达到上述目的，本说明书实施例采用如下技术方案：

第一方面，本说明书提供一种信息提取模型的训练方法，该方法可用于训练具有多语言信息提取能力的信息提取模型，所述方法包括：获得样本文档对应的原始文本信息、以及所述样本文档对应的信息提取标注信息，所述样本文档为采用第一语言集合中的至少一种语言描述的文档，所述原始文本信息包括所述样本文档中的 N 个原始文本单元，所述 N 为大于或等于 1 的整数；基于所述原始文本信息，生成所述样本文档对应的扩增文本信息，所述扩增文本信息中包括所述 N 个原始文本单元中的至少部分原始文本单元各自对应的至少一个扩增文本单元，所述扩增文本单元是通过将所述原始文本单元中的部分文本内容翻译为第二语言集合中的至少一种语言得到的；以及基于所述原始文本信息、所述扩增文本信息、以及所述信息提取标注信息，对所述信息提取模型进行训练。

第二方面，本说明书还提供一种信息提取方法，包括：获得目标文档；对所述目标文档进行文本识别得到目标文本信息，所述目标文本信息包括所述目标文档中的至少一个文本单元；通过信息提取模型对所述目标文本信息进行信息提取，得到所述目标文档中的关键信息，其中，所述信息提取模型是通过第一方面任一项所述的方法训练得到的。

第三方面，本说明书还提供一种信息提取模型的训练系统，包括至少一个存储介质，存储有至少一个指令集用于信息提取模型的训练；以及至少一个处理器，同所述至少一个存储介质通信连接，其中，当所述训练系统运行时，所述至少一个处理器读取所述至少一个指令集，并且根据所述至少一个指令集的指示：获得样本文档对应的原始文本信息、以及所述样本文档对应的信息提取标注信息，所述样本文档为采用第一语言集合中的至少一种语言描述的文档，所述原始文本信息包括所述样本文档中的 N 个原始文本单元，所述 N 为大于或等于 1 的整数；基于所述原始文本信息，生成所述样本文档对应的扩增文本信息，所述扩增文本信息中包括所述 N 个原始文本单元中的至少部分原始

文本单元各自对应的至少一个扩增文本单元，所述扩增文本单元是通过将所述原始文本单元中的部分文本内容翻译为第二语言集合中的至少一种语言得到的；以及基于所述原始文本信息、所述扩增文本信息、以及所述信息提取标注信息，对所述信息提取模型进行训练。

第四方面，本说明书还提供一种信息提取系统，包括：至少一个存储介质，存储有至少一个指令集用于信息提取；以及至少一个处理器，同所述至少一个存储介质通信连接，其中，当所述信息提取系统运行时，所述至少一个处理器读取所述至少一个指令集并实施第二方面中提供的信息提取方法。

第五方面，本说明书还提供一种计算机可读非易失性存储介质，其中，所述计算机可读非易失性存储介质中存储有至少一个指令集，所述至少一个指令集被至少一个处理器执行时，实现第一方面中提供的信息提取模型的训练方法，或者实现如第二方面所述的信息提取方法。

本说明书提供的信息提取模型的训练方法及系统、信息提取方法及系统的其他功能将在以下说明中部分列出。本说明书提供的信息提取模型的训练方法及系统、信息提取方法及系统的创造性方面可以通过实践或使用下面详细示例中所述的方法、装置和组合得到充分解释。

附图说明

为了更清楚地说明本说明书实施例中的技术方案，下面将对实施例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说明书的一些实施例，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他的附图。

图 1 示出了根据本说明书的实施例提供的一种信息提取的应用场景示意图；

图 2 示出了根据本说明书的实施例提供的一种计算设备的硬件结构图；

图 3 示出了根据本说明书的实施例提供的一种信息提取模型的训练方法的流程图；

图 4 示出了根据本说明书的实施例提供的一种扩增文本单元的生成过程示意图；

图 5A 示出了根据本说明书的实施例提供的一组原始文本单元和扩增文本单元的示意图；

图 5B 示出了根据本说明书的实施例提供的另一组原始文本单元和扩增文本单元的示意图；

图 5C 示出了根据本说明书的实施例提供的另一组原始文本单元和扩增文本单元的示意图；

图 6 示出了根据本说明书的实施例提供的信息提取模型的两个训练阶段的示意图；

图 7 示出了根据本说明书的实施例提供的一种信息提取方法的流程图。

具体实施方式

以下描述提供了本说明书的特定应用场景和要求，目的是使本领域技术人员能够制造和使用本说明书中的内容。对于本领域技术人员来说，对所公开的实施例的各种局部修改是显而易见的，并且在不脱离本说明书的精神和范围的情况下，可以将这里定义的一般原理应用于其他实施例和应用。因此，本说明书不限于所示的实施例，而是与权利要求一致的最宽范围。

这里使用的术语仅用于描述特定示例实施例的目的，而不是限制性的。比如，除非上下文另有明确说明，这里所使用的，单数形式“一”，“一个”和“该”也可以包括复数形式。当在本说明书中使用时，术语“包括”、“包含”和/或“含有”意思是指所关联的整数，步骤、操作、元素和/或组件存在，但不排除一个或多个其他特征、整数、步骤、操作、元素、组件和/或组的存在或在该系统/方法中可以添加其他特征、整数、步骤、操作、元素、组件和/或组。

考虑到以下描述，本说明书的这些特征和其他特征、以及结构的相关元件的操作和功能、以及部件的组合和制造的经济性可以得到明显提高。参考附图，所有这些形成本说明书的一部分。然而，应该清楚地理解，附图仅用于说明和描述的目的，并不旨在限制本说明书的范围。还应理解，附图未按比例绘制。

本说明书中使用的流程图示出了根据本说明书中的一些实施例的系统实现的操作。应该清楚地理解，流程图的操作可以不按顺序实现。相反，操作可以以反转顺序或同时实现。此外，可以向流程图添加一个或多个其他操作。可以从流程图中移除一个或多个操作。

为了方便描述，首先对本说明书后文中将会出现的术语进行解释。

LayoutXLM：LayoutXLM 是一种预训练模型。这种模型结合了文本和布局信息，以处理跨模态（如文本和图像）和跨语言的任务。LayoutXLM 在多种跨模态和跨语言的基准测试中表现出色，包括文档图像分类、表格识别、表格实体识别等任务。

多语言对齐任务：多语言对齐任务通常指的是将不同语言的文本或信息对齐或匹配的任务。例如，在机器翻译中，源语言和目标语言的句子或短语需要进行对齐，以便机器可以学习不同语言之间的对应关系。在跨语言信息检索中，查询和文档需要在不同语言之间进行对齐，以便找到相关的信息。

下面对本说明书的应用场景进行介绍。

本说明书提供了一种信息提取模型的训练方法和信息提取方法。该信息提取模型的训练方法和信息提取方法可以应用于交易验真场景。在该场景中，风控方可以预先训练得到信息提取模型。当交易参与方发起交易请求时，风控方可以要求交易参与方上传能够证明该交易的真实性的证明文档。风控方利用训练得到的信息提取模型，对上述证明文档进行信息提取得到关键信息，并根据关键信息对交易进行验真。若验真结果表征该交易为风险交易，则风控方拦截该交易。其中，上述的信息提取模型的训练过程可以采用本说明书提供的信息提取模型的训练方法，上述对证明文档进行信息提取的过程可以

采用本说明书提供的信息提取方法。

需要说明的是，上述交易验真场景只是本说明书提供的多个使用场景中的一个，本说明书提供的信息提取模型的训练方法和信息提取方法不仅可以应用于交易验真场景，还可以应用于需要从文档中提取信息的所有场景，比如，舆情分析场景、智能阅卷场景、病历分析场景、翻译场景，等等。本领域技术人员应当明白，当本说明书提供的信息提取模型的训练方法和信息提取方法应用于其他使用场景时，其实现方式及技术效果是类似的。

图 1 示出了根据本说明书的实施例提供的一种信息提取的应用场景示意图。

如图 1 所示，应用场景 100 可以包括训练设备 11 和执行设备 12。应用场景 100 涉及到模型训练阶段和模型应用阶段。其中，在模型训练阶段，训练设备 11 可以基于样本文档库训练得到信息提取模型。该训练后的信息提取模型可以被部署到执行设备 12 中，从而执行设备 12 可以利用信息提取模型对目标文档进行信息提取。

其中，样本文档库中可以包含使用常用语言描述的样本文档，以及每个样本文档对应的信息提取标注信息。样本文档库可以是设置在本地的数据库，也可以是设置在云端、服务器等位置的数据库。

训练设备 11 可以是具有一定计算能力的电子设备。训练设备 11 可以执行本说明书提供的信息提取模型的训练方法。训练设备 11 可以存储有执行本说明书描述的信息提取模型的训练方法的数据或指令，并可以执行或用于执行所述数据或指令。训练设备 11 可以包括具有数据信息处理功能的硬件设备和驱动该硬件设备工作所需必要的程序。

信息提取模型是经过训练设备 11 训练得到的具有信息提取能力的机器学习模型。本说明书对于信息提取模型的具体结构不做限定。在一些实施例中，信息提取模型可以采用 LayoutXLM 模型。

作为示例，训练设备 11 可以是服务器，例如本地服务器、云端服务器、机架式服务

器、刀片服务器。或者，也可以是台式计算机、笔记本电脑等。

执行设备 12 可以是具有一定计算能力的电子设备。执行设备 12 可以执行本说明书提供的信息提取方法。执行设备 12 可以存储有执行本说明书描述的信息提取方法的数据或指令，并可以执行或用于执行所述数据或指令。执行设备 12 可以包括具有数据信息处理功能的硬件设备和驱动该硬件设备工作所需必要的程序。

执行设备 12 上部署有已训练好的信息提取模型。执行设备 12 在获得到目标文档之后，可以采用上述信息提取模型对目标文档进行信息提取，从而得到目标文档中的关键信息。

在某个具体场景的系统架构中，执行设备 12 可以由该系统架构中的特定设备担任。例如，以交易验真场景为例，该场景的系统架构通常包括终端设备（对应于交易客户端）和服务器（对应于交易服务端）。

在一些实施例中，执行设备 12 可以由服务器担任。该情况下，交易参与方（例如用户或者商户）通过终端设备发起交易请求，并通过终端设备上传交易证明文档。终端设备可以将交易证明文档发送至服务器，由服务器对交易证明文档进行信息提取得到关键信息，进而基于关键信息进行交易验真。

在一些实施例中，执行设备 12 也可以由终端设备担任。该情况下，交易参与方通过终端设备发起交易请求，并通过终端设备上传交易证明文档。终端设备对交易证明文档进行信息提取得到关键信息，将上述关键信息发送至服务器，由服务器基于关键信息进行交易验真。

在一些实施例中，终端设备可以包括移动设备、平板电脑、笔记本电脑、机动车辆的内置设备或类似内容，或其任意组合。在一些实施例中，所述移动设备可包括智能家居设备、智能移动设备、虚拟现实设备、增强现实设备或类似设备，或其任意组合。在一些实施例中，所述智能家居装置可包括智能电视、台式电脑等，或任意组合。在一些实施例中，所述智能移动设备可包括智能手机、个人数字辅助、游戏设备、导航设备等，

或其任意组合。在一些实施例中，所述虚拟现实设备或增强现实设备可能包括虚拟现实头盔、虚拟现实眼镜、虚拟现实补丁、增强现实头盔、增强现实眼镜、增强现实补丁或类似内容，或其中的任何组合。例如，所述虚拟现实设备或所述增强现实设备可能包括谷歌眼镜、头戴式显示器、VR 等。在一些实施例中，所述机动车中的内置装置可包括车载计算机、车载电视等。在一些实施例中，终端设备可以包括图像采集设备，用于采集目标文档的视频或图像信息。在一些实施例中，所述图像采集设备可以是二维图像采集设备（比如 RGB 摄像头），也可以是二维图像采集设备（比如 RGB 摄像头）和深度图像采集设备（比如 3D 结构光摄像头、激光探测器，等等）。在一些实施例中，图像采集设备可以内置于终端设备中，例如智能手机、平板电脑内置的摄像头。在一些实施例中，图像采集设备还可以是与终端设备连接的外置设备，例如与台式电脑连接的扫描仪、高拍仪、摄像头等。

在一些实施例中，终端设备可以安装有一个或多个应用程序（APP）。所述 APP 包括但不限于：网页浏览器类 APP 程序、搜索类 APP 程序、聊天类 APP 程序、购物类 APP 程序、视频类 APP 程序、理财类 APP 程序、即时通信工具、邮箱客户端、社交平台软件等等。在一些实施例中，终端设备可以安装有目标 APP。所述目标 APP 能够根据终端设备采集的目标文档的视频或图像信息，对目标文档进行文本识别得到目标文本信息。然后，终端设备将目标文本信息输入信息提取模型，获取目标文档中的关键信息。

需要说明的是，上述训练设备 11 和执行设备 12 可以对应于相同的实体设备，也可以对应不同的实体设备，本说明书对此不做限定。

图 2 示出了根据本说明书的实施例提供的一种计算设备的硬件结构图。

在一些实施例中，计算设备 200 可以作为图 1 中的训练设备 11，执行本说明书提供的信息提取模型的训练方法。计算设备 200 还可以作为本说明书提供的信息提取模型的训练系统。在一些实施例中，计算设备 200 可以作为图 1 中的执行设备 12，执行本说明书提供的信息提取方法。计算设备 200 还可以作为本说明书提供的信息提取系统。

如图2所示，计算设备200可以包括至少一个存储介质230和至少一个处理器220。在一些实施例中，计算设备200还可以包括通信端口250和内部通信总线210。计算设备200还可以包括I/O组件260。

内部通信总线210可以连接不同的系统组件。例如，内部通信总线210可以连接存储介质230、处理器220、通信端口250和I/O组件260等。

I/O组件260支持计算设备200和其他组件之间的输入/输出。

通信端口250用于计算设备200同外界的数据通信。比如，通信端口250可以用于计算设备200同网络13之间的数据通信。通信端口250可以是有线通信端口也可以是无线通信端口。

存储介质230可以包括数据存储装置。所述数据存储装置可以是非暂时性存储介质，也可以是暂时性存储介质。比如，所述数据存储装置可以包括磁盘232、只读存储介质（ROM）234或随机存取存储介质（RAM）236中的一种或多种。存储介质230还包括存储在所述数据存储装置中的至少一个指令集。所述指令集可以包括计算机程序代码，所述计算机程序代码可以包括程序、例程、对象、组件、数据结构、过程、模块等等。

至少一个处理器220可以同至少一个存储介质230通信连接。当计算设备200运行时，至少一个处理器220读取所述至少一个指令集，并且根据所述至少一个指令集的指示，执行本说明书提供的信息提取模型的训练方法或信息提取方法。处理器220可以执行信息提取模型的训练方法或信息提取方法包含的步骤。处理器220可以是一个或多个处理器的形式，在一些实施例中，处理器220可以包括一个或多个硬件处理器，例如微控制器，微处理器，精简指令集计算机（RISC），专用集成电路（ASIC），特定于应用的指令集处理器（ASIP），中心处理单元（CPU），图形处理单元（GPU），物理处理单元（PPU），微控制器单元，数字信号处理器（DSP），现场可编程门阵列（FPGA），高级RISC机器（ARM），可编程逻辑器件（PLD），能够执行一个或多个功能的任何电路或处理器等，或其任何组合。

仅仅为了说明问题，附图中计算设备 200 仅示出了一个处理器 220。然而，应当注意，本说明书中计算设备 200 还可以包括多个处理器，因此，本说明书中披露的操作和/或方法步骤可以由一个处理器执行，也可以由多个处理器联合执行。例如，如果在本说明书中描述计算设备 200 的处理器 220 执行步骤 A 和步骤 B，则应该理解，步骤 A 和步骤 B 也可以由两个不同处理器 220 联合或分开执行（例如，第一处理器执行步骤 A，第二处理器执行步骤 B，或者第一和第二处理器共同执行步骤 A 和 B）。

图 3 示出了根据本说明书的实施例提供的一种信息提取模型的训练方法的流程图。如前所述，训练设备 11 可以执行本说明书的信息提取模型的训练方法。

如图 3 所示，信息提取模型的训练方法可以包括：

S310：获得样本文档对应的原始文本信息、以及样本文档对应的信息提取标注信息，样本文档为采用第一语言集合中的至少一种语言描述的文档，原始文本信息包括样本文档中的 N 个原始文本单元，N 为大于或等于 1 的整数。

第一语言集合可以为包含至少一种常用语言的集合。在本说明书中，常用语言可以理解为经常使用的语言。不常用语言可以理解为不经常使用的语言。能够理解，不同的地域或者环境对应的常用语言可能有所不同。在一些实施例中，第一语言合集中可以包括中文、英文中的至少一种。

当第一语言集合包括中文和英文时，样本文档是指采用中文、英文中的至少一种语言描述的文档。本说明书对于样本文档的文档类型不做限定。例如，样本文档可以是合同文件、个人身份证件、公司或组织的执照、许可证、物流单据、流水单据等。样本文档可以是实体文档、或实体文档的电子版、照片等。

样本文档对应的原始文本信息是指样本文档中所包含的文本信息。所述原始文本信息中可以包括多个原始文本单元。在本说明书中，一个原始文本单元是指能够表达一定语义的一个文本块。一个原始文本单元可以对应于样本文档中的一个字/单词、词语、短语、句子、段落等。本说明书对于原始文本信息中包含的原始文本单元的数量不做限定，

后文中以 N 个原始文本单元为例进行说明。

在一些实施例中，可以通过对样本文档进行文本识别得到所述 N 个原始文本单元。例如，可以对样本文档进行光学字符识别（Optical Character Recognition，OCR），得到所述 N 个原始文本单元。比如通过 OCR 从样本文档中识别得到的字/单词、词语、句子、段落等都可以作为原始文本单元。在一些实施例中，对样本文档进行文本识别还可以得到所述 N 个原始文本单元的布局信息。上述布局信息可以表征原始文本单元在样本文档中的位置、方向以及格式等。也就是说，所述原始文本信息可以包括 N 个原始文本单元、以及所述 N 个原始文本单元各自对应的布局信息。

在一些实施例中，当通过 OCR 从样本文档中识别得到一段英文文本“Account Name: HONGKONG ABC TECHNOLOGY CO., LIMITED”时，这段英文文本可以作为一个原始文本单元。

在另一些实施例中，上述英文文本中的各短语可以分别作为一个原始文本单元，比如上述英文文本包括 3 个原始文本单元“Account Name”、“HONGKONG ABC TECHNOLOGY”、“CO., LIMITED”。

需要说明的是，上述对原始文本单元的举例仅为示例，还可以将颗粒度设置的更细。例如，在一些实施例中，将可以将“HONGKONG ABC TECHNOLOGY”进一步分割，得到“HONGKONG”、“ABC”、“TECHNOLOGY”3 个原始文本单元。这样，上述英文文本一共包含 5 个原始文本单元。或者，在另一些实施例中，还可以将颗粒度设置的更粗。例如，上述英文文本共包括 2 个原始文本单元，分别为“Account Name”和“HONGKONG ABC TECHNOLOGY CO., LIMITED”。

在一些实施例中，样本文档对应的信息提取标注信息用于指示样本文档中关键信息。

其中，关键信息是指样本文档中需要关注的信息，例如，需要信息提取模型提取的信息。作为示例，以交易验真场景为例，关键信息可以根据文档的类型设置。例如，对

于个人身份证件，关键信息可以包括姓名、证件号、证件有效期等。对于物流单据，关键信息可以包括物流单号、发货地、收货地、发货人、收货人等。

作为示例，为了指示样本文档中关键信息，信息提取标注信息可以包括样本文档中对应于关键信息的原始文本单元的标识。或者，信息提取标注信息也可以包括样本文档中对应于关键信息的原始文本单元中的文本内容。

作为示例，当样本文档为物流单据时，可以对物流单据的图像进行 OCR 识别。可能得到以下原始文本单元：“1：物流单号：123456789”、“2：发货地 A 国 B 市”、“3：收货地 C 国 D 市”、“4：发货人：甲公司”、“5：收货人：乙公司”、“6：货物重量：xxxKG”、“7：货物尺寸：x*y*z”、“8：运输方式：快递”、“9：费用明细：单价 a 元、运费 b 元、保险 c 元”等。

其中，关键信息可以包括下述 5 个原始文本单元：“1：物流单号：123456789”、“2：发货地 A 国 B 市”、“3：收货地 C 国 D 市”、“4：发货人：甲公司”、“5：收货人：乙公司”。这个情况下，可以将上述 5 个原始文本单元的标识（即编号）“1、2、3、4、5”作为信息提取标注信息，指示这 5 个原始文本单元为关键信息。

其中，信息提取标注信息可以是通过人工标注方式得到的，也可以是通过语义识别算法自动标注的，本说明书对此不做限定。

在一些实施例中，可以预先收集多个样本文档，通过对多个样本文档进行文本识别得到各样本文档对应的原始文本信息，并通过对多个样本文档进行关键信息标注得到信息提取标注信息。上述各样本文档对应的原始文本信息、以及信息提取标注信息可以被存储到样本文档库中。从而，训练设备 11 在需要训练信息提取模型时，从样本文档库中获得样本文档对应的原始文本信息、以及信息提取标注信息。

S320：基于原始文本信息，生成样本文档对应的扩增文本信息，扩增文本信息中包括 N 个原始文本单元中的至少部分原始文本单元各自对应的至少一个扩增文本单元，扩增文本单元是通过将原始文本单元中的部分文本内容翻译为第二语言集合中的至少

一种语言得到的。

其中，第二语言集合可以包括至少一种不常用语言的集合。在一些实施例中，第二语言集合中的语言与第一语言集合中的语言不同。例如，第一语言合集包括中文和英文，则第二语言合集可以包括日文、韩文、法文、俄文等。在一些实施例中，第二语言集合中包括所有语言中除第一语言集合之外的其他语言。例如，第一语言集合包括中文和英文，则第二语言集合中包括所有语言中除中文和英文之外的其他语言。

在本说明书中，训练设备可以通过对原始文本信息中的至少部分原始文本单元进行翻译（翻译为第二语言集合中的语言）得到扩展文本信息。其中，上述翻译后的文本单元被称为扩增文本单元。扩增文本单元的含义可以参见前文对原始文本单元的解释，此处不做过多赘述。扩增文本单元与原始文本单元的不同之处在于，原始文本单元是从样本文档中识别得到的原始的文本单元，而扩增文本单元是通过对原始文本单元进行翻译后得到的文本单元。

在一些实施例中，扩增文本信息中包括 N 个原始文本单元中的 M 个原始文本单元各自对应的至少一个扩增文本单元，M 为大于或等于 1 的整数。也就是说，扩增文本信息是通过对原始文本信息中的 M 个原始文本单元进行翻译得到的。其中，M 个原始文本单元在 N 个原始文本单元中的占比小于或等于预设阈值。

作为示例，对于一个样本文档，预设阈值为 n%，则 M 可以为小于或等于 N*n%的整数。

例如，n%可以为 20%-30%。以 20%为例，N 当为 300 时，M 小于或等于 60。即，样本文档中包括 300 个原始文本单元时，可以选取其中最多 60 个原始文本单元进行翻译，得到对应的扩增文本单元。对占比小于或等于预设阈值的原始文本单元进行翻译，得到对应的扩增文本单元，可以保证翻译得到的扩增文本单元依然可以表达其本身的含义。而且，还能够使得扩增文本单元与原始文本单元在不同的语言合集上语义保持对齐，

进而使根据扩增文本单元进行预训练得到的信息提取模型，有助于提高模型的性能，使预训练得到的信息提取模型对多种语言的文档进行表征的能力更加优异。

下面结合图 4 对扩增文本单元的生成方式进行举例说明。

图 4 示出了根据本说明书的实施例提供的一种扩增文本单元的生成过程示意图。如图 4 所示，在一些实施例中，训练设备可以对原始文本单元进行分词处理，得到多个文本子单元。例如参见图 4，假设原始文本单元被划分为文本子单元 1、文本子单元 2、…、文本子单元 X。训练设备可以在多个文本子单元中确定至少一个目标文本子单元。例如，训练设备可以选择文本子单元 2 和文本子单元 4 作为目标文本子单元。进而，训练设备可以将上述至少一个目标文本子单元翻译为第二语言集合中的至少一种语言，得到扩增文本单元。例如结合图 4，训练设备对文本子单元 2 进行翻译得到文本子单元 2 对应的译文，对文本子单元 4 进行翻译得到文本子单元 4 对应的译文。进而，训练设备基于文本子单元 1、文本子单元 2 对应的译文、文本子单元 3、文本子单元 4 对应的译文、文本子单元 5、…、文本子单元 X 生成扩增文本单元。

通过将原始文本单元中的部分文本内容由第一语言集合中的语言翻译为第二语言集合中的语言，可以实现在已有的原始文本信息的基础上生成更多语言的扩增文本信息，从而样本文档对应有多种语言的文本信息，能够解决不常用语言的样本文档不容易获得的问题。进一步的，由于扩增文本单元是通过对原始文本单元中的一部分文本内容进行翻译得到的，原始文本单元与扩增文本单元中的各文本子单元的语义是对齐的，从而在后文描述的预训练过程中有助于信息提取模型执行多语言对齐任务，使得预训练的信息提取模型具有对多语言表征的能力。

在一些实施例中，在对原始文本单元进行分词处理时，训练设备可以基于规则进行分词、基于统计进行分词或者基于深度学习进行分词。例如，对于英文，可以基于英文语法规则进行分词，如通过英文文本中的空格和标点符号分词，得到分词结果（即多个文本子单元）。再例如，对于中文，可以使用分词算法统计大量文本数据，学习词语边

界，然后将中文文本输入分词算法，得到分词结果。其中，分词算法可以包括隐马尔科夫模型（Hidden Markov Model，HMM）和条件随机场（Conditional Random Field，CRF）等。

在一些实施例中，在对原始文本单元进行分词处理时，训练设备可以通过训练基于神经网络的模型实现分词。例如，可以使用基于双向长短期记忆网络（Bidirectional Long Short-Term Memory，BiLSTM）的序列标注模型进行训练，然后将原始文本单元输入BiLSTM 得到分词结果。

例如，对于原始文本单元“Account Name: HONGKONG ABC TECHNOLOGY CO., LIMITED”进行分词后，可以得到以下分词：“Account Name”、“HONGKONG”、“ABC”、“TECHNOLOGY”以及“CO., LIMITED”。

经过分词处理之后得到的每个分词即为原始文本单元的一个文本子单元。

在一些实施例中，在确定目标文本子单元时，训练设备可以根据多个文本子单元在样本文档中的出现频次、多个文本子单元的词性中的至少一项，在多个文本子单元中确定出至少一个目标文本子单元。

作为示例，训练设备可以将多个文本子单元中，出现次数大于或等于预设阈值的动词和/或名词确定为目标文本子单元。

参考上述示例，分词中存在名词“Account Name”和“LIMITED”。当样本文档中出现的“Account Name”和“CO., LIMITED”次数大于预设阈值时，则可以确定“Account Name”和“CO., LIMITED”为目标文本子单元。

例如，假设预设阈值为 10 次，当“Account Name”出现次数大于 10 次，“CO., LIMITED”出现次数大于 10 次时，即可以确定“Account Name”和“CO., LIMITED”为目标文本子单元。

另一些示例中，训练设备可以对多个文本子单元按照词频排序后，将排序处于预设名次之前的动词和/或名词确定为目标文本子单元。

参考上述示例，分词中存在名词“Account Name”和“CO., LIMITED”。当对样本文档中的文本子单元按照词频排序后，“Account Name”和“CO., LIMITED”在预设名次之前，则可以确定“Account Name”和“LIMITED”为目标文本子单元。

例如，假设预设名次是 10，“Account Name”排名第 3，“CO., LIMITED”排名第 5，即可以确定“Account Name”和“CO., LIMITED”为目标文本子单元。

在一些实施例中，训练设备直接可以在多个文本子单元中，随机确定至少一个文本子单元为目标文本子单元。

参考上述示例，可以在“Account Name”、“HONGKONG”、“ABC”、“TECHNOLOGY”、以及“CO., LIMITED”中随机确定至少一个文本子单元为目标文本子单元。

例如，可以根据多个文本子单元生成词语列表，每个文本子单元为列表中的一个词语。对列表进行编号，例如，0 号对应“Account Name”，1 号对应“HONGKONG”，2 号对应“ABC”，3 号对应“TECHNOLOGY”以及 4 号对应“CO., LIMITED”。然后，生成 1-5 个处于 0-4 之间不重复的随机数，并将随机数对应的文本子单元作为目标文本子单元。假设生成两个随机数“0；4”，则确定“Account Name”和“CO., LIMITED”为目标文本子单元。

名词和/或动作相较于其他词性的词语，一般在文档中具有重要语义且在翻译时不容易出现歧义。出现频次较高的词语，相较于出现频次较低的词语，在文档中具有重要语义。基于文本子单元在样本文档中的出现频次、多个文本子单元的词性中的至少一项，确定目标文本子单元，可以筛选出具有代表性、语义重要的词语，屏蔽掉语义不重要的词语，保证后续翻译时的准确度，进而提高训练得到的信息提取模型的准确度。

在一些实施例中，训练设备在原始文本单元中确定目标文本子单元后，可以将至少一个目标文本子单元翻译为第二语言集合中的至少一种语言，得到扩增文本单元。

作为示例，第二语言集合中可以包括多种语言，在将目标文本子单元进行翻译时，可以随机选取第二语言集合中的至少一种语言。例如，参考上文中第二语言集合的示例，可以将目标文本子单元翻译为日文、韩文、法文、俄文中的至少一种。

图 5A 示出了根据本说明书的实施例提供的一组原始文本单元和扩增文本单元的示意图。如图 5A 所示，作为示例，假设原始文本单元为“Account Name: HONGKONG ABC TECHNOLOGY CO., LIMITED”，确定的目标文本子单元为“Account Name”和“CO., LIMITED”，在第二语言集合确定翻译的语言为日文。则得到的扩增文本单元为“アカウント名: HONGKONG ABC TECHNOLOGY 有限会社”。

图 5B 和图 5C 示出了根据本说明书的实施例提供的另一组原始文本单元和扩增文本单元的示意图。如图 5B 和图 5C 所示，假设原始文本单元为“Account Name: HONGKONG ABC TECHNOLOGY CO., LIMITED”，确定的目标文本子单元为“Account Name”和“CO., LIMITED”。假设在第二语言集合确定翻译的语言为日文和韩文，则根据上述示例得到的扩增文本单元可以为图 5B 中示出的“アカウント名: HONGKONG ABC TECHNOLOGY 유한회사”，或者也可以为图 5C 中示出的“계정 이름: HONGKONG ABC TECHNOLOGY 有限会社”。

S330：基于原始文本信息、扩增文本信息、以及信息提取标注信息，对信息提取模型进行训练。

在本说明书中，信息提取模型的训练过程可以包括预训练和微调训练两个阶段，每个阶段可以指示信息提取模型执行不同的任务，以训练信息提取模型的不同能力。例如，预训练阶段可以通过在输出前通过头文件（head）指示信息提取模型执行多语言对齐任务，微调训练阶段可以在输出前通过头文件（head）指示信息提取模型执行信息提取任务。

下面结合图 6 对上述两个训练阶段进行说明。图 6 示出了根据本说明书的实施例提供的信息提取模型的两个训练阶段的示意图。

参见图 6，在预训练阶段，训练设备可以基于原始文本信息和扩增文本信息，对信息提取模型进行预训练。预训练过程可以执行多语言对齐任务，使得预训练后的信息提取模型具有对目标文本单元的表征能力，目标文本单元是以第一语言集合和第二语言集合中的任意语言描述的文本单元。

作为示例，信息提取模型可以是 LayoutXLM。在进行预训练时，训练设备可以将原始文本信息和扩增文本信息输入 LayoutXLM。然后，训练设备针对 LayoutXLM 模型执行多语言对齐任务并进行预训练。

在一些实施例中，训练设备可以根据原始文本信息和扩增文本信息，获取多个样本对。一个样本对中包括具有对应关系的一个原始文本单元和一个扩增文本单元。然后，训练设备可以根据多个样本对，通过对比学习的方式对信息提取模型进行预训练。

其中，一个样本对中的原始文本单元和扩增文本单元可以视为同一类，该样本对于其他样本对可以视为不同类。例如，假设多个样本对包括：“A1、A2”、“B1、B2”、“C1、C2”、“D1、D2”。A1、A2 为对应的原始文本单元（A1）和扩增文本单元（A2），A1、A2 为同类，A1 和 B1、B2、C1、C2、D1、D2 为不同类，A2 和 B1、B2、C1、C2、D1、D2 为不同类。相应的，B1、B2 为同类，B1 和 A1、A2、C1、C2、D1、D2 为不同类，B2 和 A1、A2、C1、C2、D1、D2 为不同类。

在一些实施例中，训练设备可以将多个样本对输入信息提取模型，得到多个特征向量组。其中，每个特征向量组中包括样本对中的两个文本单元各自对应的特征向量。

例如，训练设备将多个样本对输入信息提取模型，信息提取模型可以提取多个样本对的特征向量，得到多个特征向量组。例如，信息提取模型对样本对“A1、A2”进行特征提取后，可以得到特征向量组“a1、a2”，a1 为原始文本单元（A1）的特征向量，a2 为扩增文本单元（A2）的特征向量。类似的，对于上述多个样本对，可以提取得到以下特征向量组“b1、b2”、“c1、c2”、“d1、d2”。

在一些实施例中，训练设备在得到样本对对应的特征向量组之后，可以基于对比学习的策略对信息提取模型的参数进行更新。其中，对比学习的过程可以包括：以最小化同一特征向量组中的两个特征向量之间的差异、最大化不同特征向量组中的特征向量之间的差异为训练目标，对信息提取模型的参数进行更新。

在一些实施例中，在向量空间中，可以通过距离表征两个向量之间的差异，距离越近表示差异越小，距离越远表示差异越大。

作为示例，向量空间中的距离可以通过范数度量，常见的范数包括 L1 范数、L2 范数以及无穷范数。例如，L1 范数也称为曼哈顿范数或者绝对值和，它是向量中各个元素绝对值之和。L2 范数也被称为欧几里得范数，它是向量中各个元素平方和的平方根，即欧氏距离。无穷范数是向量中各个元素绝对值的最大值。以 L2 范数为例，两个向量的 L2 范数即两个向量之间的欧式距离。

通过最小化同一特征向量组中的两个特征向量之间的距离，最大化不同特征向量组中的特征向量之间的距离，可以拉进同一样本对中的两个文本单元在向量空间之中的距离，而同一样本对中的两个文本单元用于指示同一语义在不同语言下的表征。这样，可以将信息提取模型在常用语言（第一语言集合）的文档中学习得到的文档表征能力迁移到不常用语言（第二语言集合）上，实现多语言对齐，进而提高对多种语言的文档进行表征的能力。

继续参见图 6，在经过上述预训练阶段之后，训练设备可以基于原始文本信息和信息提取标注信息，对预训练后的信息提取模型进行微调训练，使得微调训练后的信息提取模型具有提取关键信息的能力。

作为示例，在进行微调训练时，训练设备可以将原始文本信息输入预训练后的 LayoutXLM。然后，通过预训练后的 LayoutXLM 模型执行信息提取任务并进行微调训练。

在一些实施例中，训练设备可以将多个原始文本单元输入预训练后的信息提取模型，以通过信息提取模型获得多个原始文本单元各自对应的特征向量，并基于所述多个原始文本单元各自对应的特征向量进行关键信息提取，得到信息提取结果。进而，训练设备可以基于监督学习的策略对信息提取模型的参数进行更新。例如，训练设备以最小化信息提取结果和信息提取标注信息之间的差异为目标，对信息提取模型的参数进行更新。

在一些实施例中，参见图 6，当信息提取模型采用 LayoutXLM 模型时，LayoutXLM 模型可以支持视觉与语言融合，即，LayoutXLM 模型既可以支持文本作为输入，还可以支持布局信息作为输入。因此，训练设备可以将多个原始文本单元、以及所述多个原始文本单元的布局信息输入预训练后的信息提取模型，以通过信息提取模型获得多个原始文本信息各自对应的特征向量，进而基于所述多个原始文本各自对应的特征向量进行关键信息提取，得到信息提取结果。进而，训练设备可以基于监督学习的测量对信息提取模型的参数进行更新。该训练方式中，在微调训练阶段，信息提取模型不仅可以基于原始文本单元从语言维度对文档进行理解，还可以基于布局信息从视觉维度对文档进行理解，能够进一步提高信息提取模型对文档的信息理解能力，从而提高信息提取性能。

综上所述，本说明书提供的信息提取模型的训练方法中，可以获得采用第一语言集合中的语言描述的样本文档对应的原始文本信息，将原始文本信息中的部分文本内容翻译为第二语言集合中的语言，得到扩增文本信息，然后根据原始文本信息、扩增文本信息、以及样本文档的信息提取标注信息，对信息提取模型进行训练。通过对原始文本信息进行翻译，可以得到含义相同但是包含不同语言的扩增文本信息，使用原始文本信息和扩增文本信息对信息提取模型进行训练，可以使训练得到的信息提取模型具有提取多种语言的信息的能力，即，不仅适用于对采用第一语言集合中的语言（例如常用语言）描述的文档进行信息提取，还适用于对采用第二语言集合中的语言（例如不常用语言）描述的文档进行信息提取，从而能够满足需要对多种语言的文档进行信息提取的需求。

本说明书另一方面提供一种信息提取方法，图 7 示出了根据本说明书的实施例提供

的一种信息提取方法的流程图。作为示例，图 1 中执行设备 12 可以执行本说明书的信息提取方法。

如图 7 所示，信息提取方法可以包括：

S710、获得目标文档。

在一些实施例中，目标文档可以是采用第一语言集合和第二语言集合中的任意一种或多种语言描述的文件、证件、单据等。目标文档可以是实体文档、或实体文档的电子版、照片等。

例如，在交易验真场景下，目标文档可以是交易参与方上传的用于证明目标交易的真实性的交易证明文档。执行设备 12 作为终端设备时，执行设备 12 可以根据交易参与方的上传操作获得目标文档。执行设备 12 作为服务器时，执行设备 12 可以从终端设备接收目标文档。

S720、对目标文档进行文本识别得到目标文本信息，目标文本信息包括目标文档中的至少一个文本单元。

其中，执行设备 12 对目标文档进行文本识别的过程可以参见前文描述的样本文档的文本识别过程，此处不做赘述。

在一些实施例中，目标文本信息除了包括目标文档中的至少一个文本单元之外，还可以包括每个文本单元的布局信息。布局信息可以表征文本单元在目标文档中的位置、方向以及格式等。

S730、通过信息提取模型对目标文本信息进行信息提取，得到目标文档中的关键信息。

其中，信息提取模型是通过上述信息提取模型的训练方法训练得到的。执行设备 12 中部署有信息提取模型。执行设备 12 可以将上述目标文本信息输入到信息提取模型，以通过信息提取模型进行信息提取，得到目标文档中的关键信息。

在交易验真场景中，目标文档中的关键信息可以用于对交易的真实性进行验证。

例如，风控方可以预先在服务器上配置通过上述信息提取模型训练方法训练得到的信息提取模型。当交易参与方发起交易请求时，交易参与方可以上传能够证明该交易的真实性的证明文档。例如，交易合同、物流单据等。

风控方可以通过信息提取模型获取证明文档中的关键信息，并根据关键信息验证该交易的真实性。

作为示例，交易参与方可以获取证明文档的电子版并上传至风控方，风控方将接收到的证明文档输入信息提取模型，并获取证明文档中的关键信息。

例如，当证明文档为物流单据时，获取到的关键信息可以包括发货人和收货人。当物流单据中的发货人与当前交易的收款方一致、且物流单据中的收货人与当前交易的付款方一致时，可以确定当前交易为真实交易。否则，可以确定当前交易不是真实交易。在一些情况下，从物流单据中提取的关键信息除了包括发货人、收货人之外，还可以包括物流单号。风控方可以基于该物流单号查询物流状态是否为妥投状态。在物流状态为妥投状态，且物流单据的发货人与当前交易的收款方一致、物流单据中的收货人与当前交易的付款方一致时，可以确定当前交易为真实交易。当上述三个条件中的任意一项不满足时，可以确定当前交易不是真实交易。进一步的，当风控方确定当前交易为真实交易时，可以放行当前交易。当风控方确定当前交易不是真实交易时，可以拦截当前交易。

需要说明的是，上述基于提取得到的关键信息对交易进行验真的方式仅为示范性描述，本说明书对此不做限定。在提取得到关键信息之后还可以采用其他方式进行交易验真。

本说明书另一方面提供一种计算机可读非暂时性存储介质，存储有至少一个指令集，用来进行信息提取模型的训练或信息提取。当所述至少一个指令集被处理器执行时，所述至少一个指令集指导所述处理器实施本说明书所述的信息提取模型的训练方法或信息提取方法的步骤。在一些可能的实施方式中，本说明书的各个方面还可以实现为一种程序产品的形式，其包括程序代码。当所述程序产品在计算设备 200 上运行时，所述程

序代码用于使计算设备200执行本说明书描述的信息提取模型的训练方法，或者信息提取方法的步骤。用于实现上述方法的程序产品可以采用便携式紧凑盘只读存储器（CD-ROM）包括程序代码，并可以在计算设备200上运行。然而，本说明书的程序产品不限于此，在本说明书中，可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统使用或者与其结合使用。所述程序产品可以采用一个或多个可读介质的任意组合。可读介质可以是可读信号介质或者可读存储介质。可读存储介质例如可以为但不限于电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。可读存储介质的更具体的例子包括：具有一个或多个导线的电连接、便携式盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM或闪存）、光纤、便携式紧凑盘只读存储器（CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。所述计算机可读存储介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了可读程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。可读存储介质还可以是可读存储介质以外的任何可读介质，该可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。可读存储介质上包含的程序代码可以用任何适当的介质传输，包括但不限于无线、有线、光缆、RF等等，或者上述的任意合适的组合。可以以一种或多种程序设计语言的任意组合来编写用于执行本说明书操作的程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、C++等，还包括常规的过程式程序设计语言—诸如“C”语言或类似的程序设计语言。程序代码可以完全地在计算设备200上执行、部分地在计算设备200上执行、作为一个独立的软件包执行、部分在计算设备200上部分在远程计算设备上执行、或者完全在远程计算设备上执行。

上述对本说明书特定实施例进行了描述。其他实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执

行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者是可能有利的。

综上所述，在阅读本详细公开内容之后，本领域技术人员可以明白，前述详细公开内容可以仅以示例的方式呈现，并且可以不是限制性的。尽管这里没有明确说明，本领域技术人员可以理解本说明书需求囊括对实施例的各种合理改变，改进和修改。这些改变，改进和修改旨在由本说明书提出，并且在本说明书的示例性实施例的精神和范围内。

此外，本说明书中的某些术语已被用于描述本说明书的实施例。例如，“一个实施例”，“实施例”和/或“一些实施例”意味着结合该实施例描述的特定特征，结构或特性可以包括在本说明书的至少一个实施例中。因此，可以强调并且应当理解，在本说明书的各个部分中对“实施例”或“一个实施例”或“替代实施例”的两个或更多个引用不一定都指代相同的实施例。此外，特定特征，结构或特性可以在本说明书的一个或多个实施例中适当地组合。

应当理解，在本说明书的实施例的前述描述中，为了帮助理解一个特征，出于简化本说明书的目的，本说明书将各种特征组合在单个实施例、附图或其描述中。然而，这并不是说这些特征的组合是必须的，本领域技术人员在阅读本说明书的时候完全有可能将其中一部分设备标注出来作为单独的实施例来理解。也就是说，本说明书中的实施例也可以理解为多个次级实施例的整合。而每个次级实施例的内容在少于单个前述公开实施例的所有特征的时候也是成立的。

本文引用的每个专利，专利申请，专利申请的出版物和其他材料，例如文章、书籍、说明书、出版物、文件、物品等，除了与本文件不一致或相冲突的、或者对权利要求的最宽范围具有限制性影响的文件，均可以通过引用结合于此，并用于现在或以后与本文件相关联的所有目的。此外，在任何材料对相关术语的描述、定义和/或使用与本文件对相关术语的描述、定义和/或使用之间存在任何不一致或冲突时，以本文件中的术语为准。

最后，应理解，本文公开的申请的实施方案是对本说明书的实施方案的原理的说明。其他修改后的实施例也在本说明书的范围内。因此，本说明书披露的实施例仅仅作为示例而非限制。本领域技术人员可以根据本说明书中的实施例采取替代配置来实现本说明书中的申请。因此，本说明书的实施例不限于申请中被精确地描述过的实施例。

权　利　要　求　书

1. 一种信息提取模型的训练方法，包括：

获得样本文档对应的原始文本信息、以及所述样本文档对应的信息提取标注信息，所述样本文档为采用第一语言集合中的至少一种语言描述的文档，所述原始文本信息包括所述样本文档中的N个原始文本单元，所述N为大于或等于1的整数；

基于所述原始文本信息，生成所述样本文档对应的扩增文本信息，所述扩增文本信息中包括所述 N 个原始文本单元中的至少部分原始文本单元各自对应的至少一个扩增文本单元，所述扩增文本单元是通过将所述原始文本单元中的部分文本内容翻译为第二语言集合中的至少一种语言得到的；以及

基于所述原始文本信息、所述扩增文本信息、以及所述信息提取标注信息，对所述信息提取模型进行训练。

2. 根据权利要求1所述的方法，其中，所述基于所述原始文本信息、所述扩增文本信息、以及所述信息提取标注信息，对所述信息提取模型进行训练，包括：

基于所述原始文本信息和所述扩增文本信息，对所述信息提取模型进行预训练，使得预训练后的所述信息提取模型具有对目标文本单元的表征能力，所述目标文本单元是以所述第一语言集合和所述第二语言集合中的任意语言描述的文本单元；以及

基于所述原始文本信息和所述信息提取标注信息，对所述信息提取模型进行微调训练，使得微调训练后的所述信息提取模型具有提取关键信息的能力。

3. 根据权利要求2所述的方法，其中，所述基于所述原始文本信息和所述扩增文本信息，对所述信息提取模型进行预训练，包括：

根据所述原始文本信息和所述扩增文本信息，获取多个样本对，一个样本对中包括具有对应关系的一个原始文本单元和一个扩增文本单元；以及

根据所述多个样本对，通过对比学习的方式对所述信息提取模型进行预训练。

4. 根据权利要求 3 所述的方法，其中，所述根据所述多个样本对，通过对比学习的方式对所述信息提取模型进行预训练，包括：

将所述多个样本对输入所述信息提取模型，得到多个特征向量组，一个特征向量组中包括一个样本对中的两个文本单元各自对应的特征向量；

以最小化同一特征向量组中的两个特征向量之间的差异、最大化不同特征向量组中的特征向量之间的差异为训练目标，对所述信息提取模型的参数进行更新。

5. 根据权利要求 2 所述的方法，其中，基于所述原始文本信息和所述信息提取标注信息，对所述信息提取模型进行微调训练，包括：

将所述 N 个原始文本单元输入预训练后的所述信息提取模型，以通过所述信息提取模型获得所述 N 个原始文本单元各自对应的特征向量，并基于所述 N 个原始文本单元各自对应的特征向量进行关键信息提取，得到信息提取结果；以及

以最小化所述信息提取结果和所述信息提取标注信息之间的差异为训练目标，对所述信息提取模型的参数进行更新。

6. 根据权利要求 1 所述的方法，其中，所述扩增文本单元是采用如下方式生成的：

对所述原始文本单元进行分词处理，得到多个文本子单元；

在所述多个文本子单元中确定至少一个目标文本子单元；以及

将所述原始文本单元中的所述至少一个目标文本子单元翻译为所述第二语言集合中的至少一种语言，得到所述扩增文本单元。

7. 根据权利要求 6 所述的方法，其中，所述在所述多个文本子单元中确定至少一个目标文本子单元，包括：

根据所述多个文本子单元在所述样本文档中的出现频次、所述多个文本子单元的词性中的至少一项，在所述多个文本子单元中确定出所述至少一个目标文本子单元。

8. 根据权利要求 7 所述的方法，其中，所述根据所述多个文本子单元在所述样本文档中的出现频次、所述多个文本子单元的词性中的至少一项，在所述多个文本子单元中确定出所述至少一个目标文本子单元，包括：

将所述多个文本子单元中，出现次数大于或等于预设阈值的动词和/或名词确定为所述目标文本子单元。

9. 根据权利要求 7 所述的方法，其中，所述根据所述多个文本子单元在所述样本文档中的出现频次、所述多个文本子单元的词性中的至少一项，在所述多个文本子单元中确定出所述至少一个目标文本子单元，包括：

对所述多个文本子单元按照词频排序后，将排序处于预设名次之前的动词和/或名词确定为所述目标文本子单元。

10. 根据权利要求 6 所述的方法，其中，所述在所述多个文本子单元中确定至少一个目标文本子单元，包括：

在所述多个文本子单元中，随机选择至少一个文本子单元作为所述目标文本子单元。

11. 根据权利要求 1 所述的方法，其中，所述扩增文本信息中包括所述 N 个原始文本单元中的 M 个原始文本单元各自对应的至少一个扩增文本单元，所述 M 为大于或等于 1 的整数，所述 M 个原始文本单元在所述 N 个原始文本单元中的占比小于或等于预设阈值。

12. 一种信息提取方法，包括：

获得目标文档；

对所述目标文档进行文本识别得到目标文本信息，所述目标文本信息包括所述目标文档中的至少一个文本单元；以及

通过信息提取模型对所述目标文本信息进行信息提取，得到所述目标文档中的关键信息，其中，所述信息提取模型是通过权利要求 1-11 任一项所述的方法训练得到的。

13. 一种信息提取模型的训练系统，包括：

至少一个存储介质，存储有至少一个指令集用于进行信息提取模型的训练；以及

至少一个处理器，同所述至少一个存储介质通信连接，其中，当所述训练系统运行时，所述至少一个处理器读取所述至少一个指令集，并且根据所述至少一个指令集的指示：

获得样本文档对应的原始文本信息、以及所述样本文档对应的信息提取标注信息，所述样本文档为采用第一语言集合中的至少一种语言描述的文档，所述原始文本信息包括所述样本文档中的 N 个原始文本单元，所述 N 为大于或等于 1 的整数，

基于所述原始文本信息，生成所述样本文档对应的扩增文本信息，所述扩增文本信息中包括所述 N 个原始文本单元中的至少部分原始文本单元各自对应的至少一个扩增文本单元，所述扩增文本单元是通过将所述原始文本单元中的至少部分文本内容翻译为第二语言集合中的至少一种语言得到的，以及

基于所述原始文本信息、所述扩增文本信息、以及所述信息提取标注信息，对所述信息提取模型进行训练。

14. 根据权利要求 13 所述的训练系统，其中，为了基于所述原始文本信息、所述扩增文本信息、以及所述信息提取标注信息，对所述信息提取模型进行训练，所述至少一个处理器：

基于所述原始文本信息和所述扩增文本信息，对所述信息提取模型进行预训练，使得预训练后的所述信息提取模型具有对目标文本单元的表征能力，所述目标文本单元是以所述第一语言集合和所述第二语言集合中的任意语言描述的文本单元；以及

基于所述原始文本信息和所述信息提取标注信息，对所述信息提取模型进行微调训练，使得微调训练后的所述信息提取模型具有提取关键信息的能力。

15. 根据权利要求 14 所述的训练系统，其中，为了基于所述原始文本信息和所述扩增文本信息，对所述信息提取模型进行预训练，所述至少一个处理器：

根据所述原始文本信息和所述扩增文本信息，获取多个样本对，一个样本对中包括具有对应关系的一个原始文本单元和一个扩增文本单元；以及

根据所述多个样本对，通过对比学习的方式对所述信息提取模型进行预训练。

16. 根据权利要求 15 所述的训练系统，其中，为了根据所述多个样本对，通过对比学习的方式对所述信息提取模型进行预训练，所述至少一个处理器：

将所述多个样本对输入所述信息提取模型，得到多个特征向量组，一个特征向量组中包括一个样本对中的两个文本单元各自对应的特征向量；

以最小化同一特征向量组中的两个特征向量之间的差异、最大化不同特征向量组中的特征向量之间的差异为训练目标，对所述信息提取模型的参数进行更新。

17. 根据权利要求 14 所述的训练系统，其中，为了基于所述原始文本信息和所述信息提取标注信息，对所述信息提取模型进行微调训练，所述至少一个处理器：

将所述 N 个原始文本单元输入预训练后的所述信息提取模型，以通过所述信息提取模型获得所述 N 个原始文本单元各自对应的特征向量，并基于所述 N 个原始文本单元各自对应的特征向量进行关键信息提取，得到信息提取结果；以及

以最小化所述信息提取结果和所述信息提取标注信息之间的差异为训练目标，对所述信息提取模型的参数进行更新。

18. 根据权利要求 14 所述的训练系统，其中，所述扩增文本单元是采用如下方式生成的：

对所述原始文本单元进行分词处理，得到多个文本子单元；

在所述多个文本子单元中确定至少一个目标文本子单元；以及

将所述原始文本单元中的所述至少一个目标文本子单元翻译为所述第二语言集合中的至少一种语言，得到所述扩增文本单元。

19. 一种信息提取系统，包括：

至少一个存储介质，存储有至少一个指令集用于信息提取；以及

至少一个处理器，同所述至少一个存储介质通信连接，其中，当所述信息提取系统运行时，所述至少一个处理器读取所述至少一个指令集并实施权利要求 12 中所述的信息提取方法。

20. 一种计算机可读非易失性存储介质，其中，所述计算机可读非易失性存储介质中存储有至少一个指令集，所述至少一个指令集被至少一个处理器执行时，实现如权利要求 1-11 中任一项所述的方法，或者实现如权利要求 12 所述的方法。