

语言模型训练方法及装置、信息处理方法及装置

本申请要求于 2024 年 8 月 28 日提交中国专利局、申请号为 202411191440.X、发明名称为“语言模型训练方法及装置、信息处理方法及装置”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

本说明书实施例涉及机器学习技术领域，特别涉及语言模型训练方法及装置、信息处理方法及装置。

背景技术

10

随着计算机技术和互联网技术的发展，大语言模型在越来越多的场景中得以应用，而为了满足用户多样化的需求，大语言模型的能力也逐渐增强。当前业务中部署的大语言模型，主要通过在大规模文档数据上进行预训练来实现。这些模型在进行文档的预测任务时，通常仅利用文档前缀的上下文信息。然而，随着模型规模的扩大和应用场景的多样化，大语言模型在处理复杂的跨文档推理、指令执行以及阅读理解任务时，仍然面临诸多挑战；例如，模型可能无法准确跟随指令、在条件文档上的推理能力有限、以及在上文学习中的高方差等问题。现有技术中，为了避免这一问题，通常会采用标准预训练方法或者基于检索的预训练方法来提高模型理解能力，但是，这两种方法在模型预训练阶段，无法保证上下文的连贯性，同时会存在数据重复问题，进而导致模型在复杂任务上的表现不佳，以及存在过拟合问题。因此亟需一种有效的方案以解决上述问题。

15

20

发明内容

鉴于此，本说明书实施例提供了一种语言模型训练方法。本说明书一个或者多个实施例同时涉及一种信息处理方法，一种语言模型训练装置，一种信息处理装置，一种计算设备，一种计算机可读存储介质以及一种计算机程序产品，以解决现有技术中存在的技术缺陷。

根据本说明书实施例的第一方面，提供了一种语言模型训练方法，包括：

25

获取初始语言模型的文档集合，并在所述文档集合中筛选关联文档组成关联文档集合；

根据所述关联文档集合中的关联文档以及关联文档之间的相似度，构建所述关联文档集合对应的文档关系图；

在所述文档关系图中确定文档拼接路径，并按照所述文档拼接路径在所述关联文档集合中提取目标关联文档进行拼接处理，获得目标上下文信息；

30

基于所述目标上下文信息对所述初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型。

可选地，所述在所述文档集合中筛选关联文档组成关联文档集合，包括：

利用文本检索模型构建所述文档集合中各个初始文档对应的文档嵌入向量；

计算各个初始文档对应的文档嵌入向量之间的余弦相似度；

35

根据余弦相似度计算结果在所述文档集合中筛选关联文档，组成关联文档集合。

可选地，所述利用文本检索模型构建所述文档集合中各个初始文档对应的文档嵌入向量步骤执行之后，还包括：

利用预设的量化器对各个初始文档对应的文档嵌入向量进行量化处理，并根据量化后的文档嵌入向量构建文档索引；

5 按照预设的搜索策略和所述文档索引针对所述文档集合进行搜索，根据搜索结果确定关联文档并组成关联文档集合。

可选地，所述根据所述关联文档集合中的关联文档以及关联文档之间的相似度，构建所述关联文档集合对应的文档关系图，包括：

10 根据所述关联文档集合中的关联文档构建图节点，以及根据所述关联文档集合中关联文档之间的相似度构建图边；

基于所述图边建立所述图节点之间的连通关系，根据连通关系建立结果生成所述关联文档集合对应的文档关系图。

可选地，所述在所述文档关系图中确定文档拼接路径，包括：

15 在所述文档关系图中确定第 i 图节点，并构建所述第 i 图节点对应的初始文档路径，其中， i 为正整数， i 从 1 开始取值，且 i 最大值为所述文档关系图中的图节点数量；

在所述第 i 图节点连接的邻居图节点中选择第 j 图节点，并基于所述第 j 图节点将所述初始文档路径更新为中间文档路径，其中， j 为正整数， j 从 1 开始取值，且 j 最大值为邻居图节点数量；

判断所述第 i 图节点连接的邻居图节点中是否存在未被访问的邻居图节点；

若否， j 自增 1，并执行在所述第 i 图节点连接的邻居图节点中选择第 j 图节点的步骤；

20 若是， i 自增 1，并执行在所述文档关系图中确定第 i 图节点的步骤，直至 i 等于所述图节点数量的情况下，获得多个中间文档路径，在所述多个中间文档路径中提取满足文档拼接条件的中间文档路径作为文档拼接路径。

可选地，所述按照所述文档拼接路径在所述关联文档集合中提取目标关联文档进行拼接处理，获得目标上下文信息，包括：

25 确定所述文档拼接路径中包含的拼接图节点，基于所述拼接图节点在所述关联文档集合中提取目标关联文档；

将所述目标关联文档按照所述文档拼接路径的路径方向信息进行拼接，获得匹配所述初始语言模型输入条件的目标上下文信息。

30 可选地，所述基于所述目标上下文信息对所述初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型，包括：

将所述目标上下文信息输入至所述初始语言模型进行处理，获得预测文本信息；

基于所述目标上下文信息对应的基准文本信息和所述预测文本信息，对所述初始语言模型进行优化；检测优化后的初始语言模型是否满足训练停止条件；

若否，在所述文档集合中筛选候选关联文档组成候选关联文档集合，并将所述候选关联文档集合作

为所述关联文档集合，并执行根据所述关联文档集合中的关联文档以及关联文档之间的相似度，构建所述关联文档集合对应的文档关系图步骤；

若是，将优化后的初始语言模型作为目标语言模型。

根据本说明书实施例的第二方面，提供了一种信息处理方法，包括：

5 获取客户端提交的问题信息，基于所述问题信息在业务知识库中查询多个答案文档；

针对所述多个答案文档进行拼接，并将拼接结果和所述问题信息输入至目标语言模型进行处理，获得答案信息，所述目标语言模型通过上述方法训练获得；

将所述答案信息反馈至所述客户端。

根据本说明书实施例的第三方面，提供了一种语言模型训练装置，包括：

10 获取模块，被配置为获取初始语言模型的文档集合，并在所述文档集合中筛选关联文档组成关联文档集合；

构建模块，被配置为根据所述关联文档集合中的关联文档以及关联文档之间的相似度，构建所述关联文档集合对应的文档关系图；

15 拼接模块，被配置为在所述文档关系图中确定文档拼接路径，并按照所述文档拼接路径在所述关联文档集合中提取目标关联文档进行拼接处理，获得目标上下文信息；

训练模块，被配置为基于所述目标上下文信息对所述初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型。

根据本说明书实施例的第四方面，提供了一种信息处理装置，包括：

20 获取信息模块，被配置为获取客户端提交的问题信息，基于所述问题信息在业务知识库中查询多个答案文档；

模型预测模块，被配置为针对所述多个答案文档进行拼接，并将拼接结果和所述问题信息输入至目标语言模型进行处理，获得答案信息，所述目标语言模型通过上述方法训练获得；

反馈信息模块，被配置为将所述答案信息反馈至所述客户端。

根据本说明书实施例的第五方面，提供了一种计算设备，包括：

25 存储器 and 处理器；

所述存储器用于存储计算机可执行指令，所述处理器用于执行所述计算机可执行指令，该计算机可执行指令被处理器执行时实现上述语言模型训练方法或信息处理方法的步骤。

根据本说明书实施例的第六方面，提供了一种计算机可读存储介质，其存储有计算机可执行指令，该指令被处理器执行时实现上述语言模型训练方法或信息处理方法的步骤。

30 根据本说明书实施例的第七方面，提供了一种计算机程序产品，包括计算机程序或指令，该计算机程序或指令被处理器执行时实现上述语言模型训练方法或信息处理方法的步骤。

本实施例提供的语言模型训练方法，为了能够提高语言模型在处理复杂上下文和跨文档推理时性能，可以在获取到初始语言模型的文档集合后，在文档集合中筛选具有相似性的关联文档组成关联文档集合；此后，为了加强模型对跨文档文本的理解能力，可以根据关联文档集合中的关联文档以及关联文档之间

的相似度，构建关联文档集合对应的文档关系图；通过文档关系图体现关联文档之间的相似关系。在此基础上，可以在文档关系图中确定文档拼接路径，并按照文档拼接路径在关联文档集合中提取目标关联文档进行拼接处理，获得目标上下文信息；最后即可基于目标上下文信息对初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型。实现模型可以在相关文档序列上进行预训练，使得语言模型能够在文档边界之外进行阅读和推理，从而有效的提升模型在复杂上下文任务中的性能。

附图说明

- 图 1 是本说明书一个实施例提供的一种语言模型训练方法的示意图；
- 图 2 是本说明书一个实施例提供的一种语言模型训练方法的流程图；
- 图 3 是本说明书一个实施例提供的一种信息处理方法的流程图；
- 图 4 是本说明书一个实施例提供的一种信息处理方法的处理过程流程图；
- 图 5 是本说明书一个实施例提供的一种语言模型训练装置的结构示意图；
- 图 6 是本说明书一个实施例提供的一种信息处理装置的结构示意图；
- 图 7 是本说明书一个实施例提供的一种计算设备的结构框图。

具体实施方式

在下面的描述中阐述了很多具体细节以便于充分理解本说明书。但是本说明书能够以很多不同于在此描述的其它方式来实施，本领域技术人员可以在不违背本说明书内涵的情况下做类似推广，因此本说明书不受下面公开的具体实施的限制。

在本说明书一个或多个实施例使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本说明书一个或多个实施例。在本说明书一个或多个实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本说明书一个或多个实施例使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

应当理解，尽管在本说明书一个或多个实施例中可能采用术语第一、第二等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本说明书一个或多个实施例范围的情况下，第一也可以被称为第二，类似地，第二也可以被称为第一。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

此外，需要说明的是，本说明书一个或多个实施例所涉及的用户信息(包括但不限于用户设备信息、用户个人信息等)和数据(包括但不限于用于分析的数据、存储的数据、展示的数据等)，均为经用户授权或者经过各方充分授权的信息和数据，并且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准，并提供有相应的操作入口，供用户选择授权或者拒绝。

首先，对本说明书一个或多个实施例涉及的名词术语进行解释。

大语言模型：(Large Language Model，简称 LLM)是自然语言处理(NLP)领域中一类重要技术，它通过使用大量文本数据训练的深度学习模型，能够生成自然语言文本或理解语言文本的含义。

贪心算法(Greedy Algorithm)是一种在每一步选择中都采取在当前状态下最好或最优(即最有利)

的选择，从而希望导致结果是全局最好或最优的算法。贪心算法所做的每一个选择都是在某种意义下的局部最优选择。

在本说明书中，提供了一种语言模型训练方法。本说明书一个或者多个实施例同时涉及一种信息处理方法，一种语言模型训练装置，一种信息处理装置，一种计算设备，一种计算机可读存储介质以及一种计算机程序产品，在下面的实施例中逐一进行详细说明。

实际应用中，为了使语言模型可以提供更好的预测性能，通常会采用标准预训练方法（Standard Pretraining）和基于检索的预训练方法（kNNPretraining）对其进行预训练。其中，标准预训练方法通过随机打乱文档，将其拼接成固定长度的上下文窗口，进行语言建模训练。这种方法虽然简单，但无法保证上下文的连贯性，导致模型在处理复杂任务时表现不佳。而基于检索的预训练方法则是将每个文档及其检索到的相关文档放在同一个上下文中进行训练。这种方法可以在一定程度上提升模型的语义理解能力，但存在数据重复问题，可能导致模型过拟合。因此亟需一种有效的方案以解决上述问题。

参见图 1 所示的示意图，本实施例提供的语言模型训练方法，为了能够提高语言模型在处理复杂上下文和跨文档推理时性能，可以在获取到初始语言模型的文档集合后，在文档集合中筛选具有相似性的关联文档组成关联文档集合；此后，为了加强模型对跨文档文本的理解能力，可以根据关联文档集合中的关联文档以及关联文档之间的相似度，构建关联文档集合对应的文档关系图；通过文档关系图体现关联文档之间的相似关系。在此基础上，可以在文档关系图中确定文档拼接路径，并按照文档拼接路径在关联文档集合中提取目标关联文档进行拼接处理，获得目标上下文信息；最后即可基于目标上下文信息对初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型。实现模型可以在相关文档序列上进行预训练，使得语言模型能够在文档边界之外进行阅读和推理，从而有效的提升模型在复杂上下文任务中的性能。

参见图 2，图 2 示出了根据本说明书一个实施例提供的一种语言模型训练方法的流程图，具体包括以下步骤。

步骤 S202，获取初始语言模型的文档集合，并在所述文档集合中筛选关联文档组成关联文档集合。本实施例提供的语言模型训练方法可以应用于任意结构的大语言模型的预训练阶段，比如语言模型可以是 LLaMA 模型（Large Language Model Meta AI，大规模语言模型）等，本实施例在此不做任何限定。

具体的，初始语言模型具体是指当前阶段需要进行预训练的语言模型，其可以是部署在问答交互场景中的语言模型，在用户提交问题信息后，其可以在知识库中查询相关文档，并根据相关文档生成匹配问题信息的答案信息，从而向用户反馈满足用户问答需求的答案信息。相应的，文档集合具体是指对初始语言模型在预训练阶段所需要使用的文档组成的集合，该集合中包含的文档可以来自于不同的数据源，且可以对应不同的业务领域，比如数学知识领域、电影知识领域、交易知识领域等，本实施例在此不做任何限定。相应的，关联文档具体是指从文档集合中筛选出具有相似关系的文档，用于组成关联文档集合，可以理解为，关联文档集合中包含的关联文档均具有相似性，用于后续实现对语言模型的预训练，使其学习到在文档边界之外进行阅读和推理的能力。

基于此，为了提高语言模型在处理复杂上下文和跨文档推理时性能，可以在获取到初始语言模型的文档集合后，在文档集合中筛选具有相似性的关联文档组成关联文档集合；此后，为了加强模型对跨文档文本的理解能力，可以根据关联文档集合中的关联文档以及关联文档之间的相似度，构建关联文档集合对应的文档关系图；通过文档关系图体现关联文档之间的相似关系。在此基础上，可以在文档关系图中确定文档拼接路径，并按照文档拼接路径在关联文档集合中提取目标关联文档进行拼接处理，获得目标上下文信息；最后即可基于目标上下文信息对初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型即可。

进一步的，在获取到初始语言模型的文档集合后，为了能够选择相似性较高的文档组成关联文档集合，从而提高模型在跨文档上的理解能力，可以通过计算余弦相似度的方式筛选关联文档。本实施例中，具体实现方式如下：

利用文本检索模型构建所述文档集合中各个初始文档对应的文档嵌入向量；计算各个初始文档对应的文档嵌入向量之间的余弦相似度；根据余弦相似度计算结果在所述文档集合中筛选关联文档，组成关联文档集合。

具体的，文本检索模型具体是指能够针对各个初始文档构建嵌入向量的模型，其可以是 *Contriever* 模型，也可以选择其他能够构建嵌入向量的模型，本实施例在此不做任何限定。相应的，文档嵌入向量具体是指利用文本检索模型对初始文档进行 *embedding* 处理后得到的向量表达。相应的，余弦相似度具体是指根据文档嵌入向量对各个初始文本进行相似度计算后得到的值。

基于此，在确定需要对初始语言模型进行预训练的文档集合后，为了能够筛选出文档相似且不重复的文档用于模型预训练，可以通过计算余弦相似度的方式完成关联文档集合的构建。此过程中，可以先利用文本检索模型构建文档集合中各个初始文档对应的文档嵌入向量；在此基础上，可以计算各个初始文档对应的文档嵌入向量之间的余弦相似度；此后可以将余弦相似度计算结果与预设相似度阈值进行比较，选择大于相似度阈值的文档作为关联文档，并组成关联文档集合，可以保证关联文档集合中包含的关联文档都是相似性更高的文档，从而可以提升模型的上下文理解和推理能力。

具体实施时，在采用 *Contriever* 模型对文档集合中的文档进行 *embedding* 处理时，可以先通过 *Contriever* 模型将每个文档 d_i 映射为嵌入向量 $E(d_i)$ ，之后通过使用余弦相似度计算文档间的相似度，其可以通过公式 $s(d_i, d_j) = \cos(E(d_i), E(d_j))$ 实现，其中， d_i 和 d_j 表示两个需要计算相似度的文档。

综上，通过采用计算余弦相似度的方式筛选相似文档组成关联文档集合，实现后续可以在模型预训练阶段，使用相似文档完成预训练处理，从而使得模型可以在跨文档处理文本的基础上也具备较强的预测能力。

此外，为了提高关联文档集合的构建效率，可以采用近似最近邻搜索和反转文件索引的方式完成关联文档集合的构建。本实施例中，具体实现方式如下：

利用预设的量化器对各个初始文档对应的文档嵌入向量进行量化处理，并根据量化后的文档嵌入向量构建文档索引；按照预设的搜索策略和所述文档索引针对所述文档集合进行搜索，根据搜索结果确定关联文档并组成关联文档集合。

具体的，量化器具体是指对每个初始文档的文档嵌入向量进行量化处理的单元。相应的，文档索引具体是指根据量化后的文档嵌入向量构建用于搜索相近文档的索引信息。相应的，搜索策略具体是指用于设定搜索相近 k 个文档的策略。

基于此，为了能够实现高效检索目的，从而快速确定相似度更高的文档组成关联文档集合，可以先利用预设的量化器对各个初始文档对应的文档嵌入向量进行量化处理，以实现根据量化后的文档嵌入向量构建文档索引；在此基础上，可以按照预设的搜索策略和文档索引针对文档集合进行搜索，进而根据搜索结果确定相似性较高的关联文档组成关联文档集合，以便后续用于对初始语言模型进行预训练使用。

实际应用中，在采用近似最近邻搜索和反转文件索引（FAISS）进行关联文档集合构建时，可以先构建 FAISS 索引，通过使用产品量化（product quantization）方法，将文档嵌入向量进行量化和索引。

此后进行相似度检索，此过程中，可以在 FAISS 索引中进行近似最近邻搜索，快速找到每个文档的前 k 个最相似文档，进而组成关联文档集合使用。

具体实施时，首先，可以进行文档嵌入向量的产品量化；具体的，配置 PQ 量化器：使用 FAISS 库中的 `faiss.ProductQuantizer` 类来配置产品量化器。需要指定量化中心的维度（`dsub`）和子量化器的数量（`M`），这样可以将整个向量空间划分为多个子空间，每个子空间独立进行量化。训练 PQ 量化器：使用一组训练数据（可以是整个数据集的一部分）来训练 PQ 量化器。训练过程会找到每个子空间的最优量化中心。量化文档嵌入：将文档的嵌入向量通过训练好的 PQ 量化器进行量化，得到量化后的编码。

其次，构建 FAISS 索引；具体的，选择索引类型：基于产品量化的索引在 FAISS 中通常使用 IVFPQ（Inverted File with Product Quantization）索引类型。这种索引结合了倒排索引（IVF）和产品量化（PQ），以实现高效的近似最近邻搜索。创建索引：使用 FAISS 的 `faiss.IndexIVFPQ` 类来创建索引。需要指定量化器（即步骤一中训练好的 PQ 量化器）、倒排列表的聚类中心数量（`nlist`）以及每个聚类中心下保留的最近邻数量。添加数据到索引：将量化后的文档嵌入向量添加到创建的 FAISS 索引中。

最后，进行近似最近邻搜索；具体的，设置搜索参数：在搜索之前，根据需要设置 `nprobe` 参数，它决定了搜索时在每个聚类中心下要探索的最近邻数量，从而影响搜索速度和准确性。执行搜索：使用 FAISS 的搜索函数（如 `index.search`）对给定查询向量执行近似最近邻搜索。查询向量可以是任何新的文档嵌入向量，该向量同样需要被 PQ 量化器量化。获取结果：搜索函数将返回每个查询向量的前 k 个最相似文档的索引和距离。从而构建出关联文档集合用于初始语言模型的预训练使用。

举例说明，确定预训练语言模型-LLaMA 模型对应的文档集合，文档集合中包含 m 个文档。通过采用 `Contriever` 模型对文档集合中包含的各个文档进行 `embedding`，得到每个文档对应的嵌入向量 E，此后通过计算每个文档对应嵌入向量 E 之间的余弦相似度，根据余弦相似度计算结果，在 m 个文档筛选出具有相似性的 n 个文档组成关联文档集合（ $n < m$ ），例如 n 个文档可以是关联图像处理知识的文档 1，文档 2... 文档 n。用于后续可以对 LLaMA 模型进行预训练，使得该模型可以具有更好的上下文理解和推理能力。

综上，通过构建索引搜索相似文档的方式进行检索，可以提高关联文档构建效率，从而节省模型预训练时间。

步骤 S204, 根据所述关联文档集合中的关联文档以及关联文档之间的相似度, 构建所述关联文档集合对应的文档关系图。

具体的, 在上述确定具有相关性的关联文档组成的关联文档集合后, 为了能够更进一步的选择高相关文档用于模型预训练, 可以通过构建文档关系图的方式体现文档之间的关联性强弱, 进而通过文档关系图筛选高关联文档用于模型预训练使用, 因此可以根据关联文档集合中的关联文档以及关联文档之间的相似度, 构建关联文档集合对应的文档关系图, 以便于后续使用。其中, 文档关系图具体是指将关联文档作为图节点, 以及将关联文档之间的相似度作为图边, 构建出的无向加权图。

进一步的, 在构建文档关系图时, 考虑到文档关系图将在后续作为模型输入上下文信息的构建基础, 因此需要充分利用文档及文档之间的相似性完成构建。本实施例中, 具体实现方式如下:

根据所述关联文档集合中的关联文档构建图节点, 以及根据所述关联文档集合中关联文档之间的相似度构建图边; 基于所述图边建立所述图节点之间的连通关系, 根据连通关系建立结果生成所述关联文档集合对应的文档关系图。

基于此, 在构建文档关系图时, 可以先根据关联文档集合中的关联文档绘制图节点, 以及根据关联文档集合中关联文档之间的相似度绘制图边, 并且在绘制图边时, 边的权重由关联文档之间的相似度决定, 进而可以结合图边建立图节点之间的连通关系, 在此基础上, 即可根据连通关系建立结果生成关联文档集合对应的文档关系图, 以便后续进行高相关文档的筛选和拼接使用。

沿用上例, 在得到 n 个关联图像处理知识的文档后, 可以基于各个文档绘制 n 个图节点, 同时基于 n 个文档中各个文档之间的相似度, 构建图边, 在此基础上, 基于图边对 n 个图节点进行连接, 即可得到无向加权图, 以便后续通过分析该无向加权图, 从而筛选出设定数量的文档进行拼接, 进而对 LLaMA 模型进行预训练。

综上, 通过构建文档关系图的方式体现关联文档之间的相关度, 方便后续在此基础上筛选出高相关的文档进行拼接使用, 从而提高初始语言模型的预训练效果。

步骤 S206, 在所述文档关系图中确定文档拼接路径, 并按照所述文档拼接路径在所述关联文档集合中提取目标关联文档进行拼接处理, 获得目标上下文信息。

具体的, 在上述构建出文档关系图后, 由于文档关系图能够体现出各个关联文档之间的相关度强弱, 因此可以通过遍历文档关系图, 以在文档关系图中找到文档拼接路径, 通过文档拼接路径可以确定高相关关联文档, 此时即可按照文档拼接路径在关联文档集合中提取目标关联文档进行拼接处理, 进而得到能够输入模型进行处理的目标上下文信息, 实现在此基础上, 使用高相关文档对语言模型进行训练, 使得高模型可以从相关文档序列中学习上下文理解和推理能力。

其中, 文档拼接路径具体是指对文档关系图进行遍历, 根据遍历结果从中找到具有相关度更高的文档对应节点组成的链路, 该路径由多个图节点组成, 且各个图节点对应的关联文档为高相关性文档。相应的, 目标关联文档即为文档拼接路径所涉及到的关联文档。相应的, 目标上下文信息具体是指对文档拼接路径所关联的目标关联文档进行拼接后得到的文本。

进一步的, 在文档关系图中确定文档拼接路径时, 考虑到该路径将决定目标上下文信息中相似文档

之间的关联性，为了保证目标上下文信息是采用强关联的文档构建得到，从而可以提高模型跨文档的理解和推理能力，可以采用贪心算法确定文档拼接路径。本实施例中，具体实现方式如下：

在所述文档关系图中确定第 i 图节点，并构建所述第 i 图节点对应的初始文档路径，其中， i 为正整数， i 从 1 开始取值，且 i 最大值为所述文档关系图中的图节点数量；在所述第 i 图节点连接的邻居图节点中选择第 j 图节点，并基于所述第 j 图节点将所述初始文档路径更新为中间文档路径，其中， j 为正整数， j 从 1 开始取值，且 j 最大值为邻居图节点数量；判断所述第 i 图节点连接的邻居图节点中是否存在未被访问的邻居图节点；若否， j 自增 1，并执行在所述第 i 图节点连接的邻居图节点中选择第 j 图节点的步骤；若是， i 自增 1，并执行在所述文档关系图中确定第 i 图节点的步骤，直至 i 等于所述图节点数量的情况下，获得多个中间文档路径，在所述多个中间文档路径中提取满足文档拼接条件的中间文档路径作为文档拼接路径。

具体的，第 i 图节点具体是指文档关系图中边数量最少得图节点，初始文档路径具体是指将第 i 图节点作为路径起点构建的初始路径；相应的，第 j 图节点具体是指第 i 图节点的邻居节点中未被访问的邻居节点。相应的，中间文档路径具体是指将第 j 图节点加入初始文档路径后得到的路径。

基于此，在进行文档拼接路径确定时，为了能够从文档关系图中挖掘出高相关文档对应节点组成的路径，此时可以从图维度，在文档关系图中确定最低度的第 i 图节点，并构建第 i 图节点对应的初始文档路径，其中， i 为正整数， i 从 1 开始取值，且 i 最大值为文档关系图中的图节点数量。

而后可以在第 i 图节点连接的邻居图节点中选择未被访问的第 j 图节点，并将第 j 图节点添加到初始文档路径中，进而得到中间文档路径，其中， j 为正整数， j 从 1 开始取值，且 j 最大值为邻居图节点数量。

在此基础上，可以判断第 i 图节点连接的邻居图节点中是否存在未被访问的邻居图节点；若否，说明第 i 图节点对应的文档路径还可以扩展，因此可以 j 自增 1，并执行在第 i 图节点连接的邻居图节点中选择第 j 图节点的步骤；若是，说明第 i 图节点的扩展路径已经达到上限，因此可以选择除第 i 图节点之外的其他最低度节点再进行路径构建，因此可以 i 自增 1，并执行在所述文档关系图中确定第 i 图节点的步骤。直至 i 等于所述图节点数量的情况下，说明文档关系图中的全部图节点对应的路径都构建完成，因此可以获得多个中间文档路径，而后可以在多个中间文档路径中提取满足文档拼接条件的中间文档路径作为文档拼接路径。具体实施时，在选择文档拼接路径时可以选择最长中间文档路径作为文档拼接路径。

实际应用中，考虑到文档关系图为无向加权图，而为了能够在此基础上充分挖掘无向加权图，进而通过分析图的方式快速且精准的确定目标关联文档，可以采用贪心算法解决最大旅行商问题的构思挖掘最大旅行商路径（文档拼接路径）。也就是说，可以先初始化路径：从最低度的节点（与其他节点边连接数量最小的节点）开始，构建初始路径；之后再行路径扩展：通过选择当前节点的最高权重且未访问的邻居节点，将其加入到初始路径中，进而实现初始路径的更新；以此类推，当所有邻居节点均已访问后，则完成当前路径构建。此时可以选择下一个最低度节点，继续进行路径扩展，直至全部节点都完成路径构建后，即可从构建好的路径中确定最大旅行商路径，即文档拼接路径，在此基础上，即可根据

文档拼接路径选择文档进行拼接，以得到用于模型预训练使用的目标上下文信息。

综上，通过采用贪心算法解决最大旅行商路径问题，可以实现在文档关系图上最大化相关文档的连续访问，从而提高输入上下文的连贯性。

更进一步的，在得到文档拼接路径后，由于路径不仅能够体现所需文档内容，还具有方向性，因此可以按照方向信息对目标关联文档进行拼接。本实施例中，具体实现方式如下：

确定所述文档拼接路径中包含的拼接图节点，基于所述拼接图节点在所述关联文档集合中提取目标关联文档；将所述目标关联文档按照所述文档拼接路径的路径方向信息进行拼接，获得匹配所述初始语言模型输入条件的目标上下文信息。

具体的，拼接图节点具体是指文档拼接路中包含的图节点。相应的，路径方向信息具体是指表征文档拼接前后关系的信息。基于此，在得到文档拼接路径后，可以确定文档拼接路径中包含的拼接图节点；在此基础上，可以基于拼接图节点在关联文档集合中提取目标关联文档；而后为了保证拼接结果的连贯性，可以将目标关联文档按照文档拼接路径的路径方向信息进行拼接，进而可以根据拼接结果获得匹配初始语言模型输入条件的目标上下文信息，以实现后续用于预训练初始语言模型使用。

沿用上例，在得到基于 n 个文档构建的无向加权图，进一步的，可以采用贪心算法对无向加权图进行处理，进而挖掘出最大权重路径。具体的，可以先在无向加权图中选择最低度的图节点，而后将其作为起始节点构建初始路径。在此基础上，可以选择该图节点的最高权重且未被访问的邻居节点，将其加入到初始路径中，构建出中间路径。以此类推，直至所有的邻居节点都被访问后，再选择除该图节点之外且最低度的图节点再进行路径挖掘。最后通过遍历图中的每个节点，即可生成最大权重路径，其包含文档 1，文档 2，文档 3 和文档 5，在此基础上，沿最大权重路径对文档 1，文档 2，文档 3 和文档 5 进行拼接，可以得到固定大小的输入上下文，以便后续进行模型预训练使用。

综上，由于文档拼接路径是基于文档关系图挖掘得到，因此基于该路径进行文档拼接生成上下文信息时，可以保证上下文的连贯性，同时也能够保证各个目标关联文档的唯一性，从而可以解决数据重复问题，进而提高模型训练效果。

步骤 S208，基于所述目标上下文信息对所述初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型。

具体的，在上述通过拼接目标关联文档得到目标上下文信息后，进一步的，由于目标上下文信息是基于高相关的关联文档拼接而成，因此基于目标上下文信息对初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型后，可以有效提高目标语言模型的上下文理解能力和推理能力，从而显著提升目标语言模型在复杂任务重的表现。

进一步的，在模型训练过程中，为了能够得到满足部署需求的目标语言模型，可以在训练阶段进行模型检测。本实施例中，具体实现方式如下：

将所述目标上下文信息输入至所述初始语言模型进行处理，获得预测文本信息；基于所述目标上下文信息对应的基准文本信息和所述预测文本信息，对所述初始语言模型进行优化；检测优化后的初始语言模型是否满足训练停止条件；若否，在所述文档集合中筛选候选关联文档组成候选关联文档集合，并

将所述候选关联文档集合作为所述关联文档集合，并执行根据所述关联文档集合中的关联文档以及关联文档之间的相似度，构建所述关联文档集合对应的文档关系图步骤；若是，将优化后的初始语言模型作为目标语言模型。

具体的，预测文本信息具体是指初始语言模型对目标上下文信息进行法处理后，得到的预测信息。

5 相应的，基准文本信息具体是指目标上下文信息对应的真实信息。结合预测文本信息和基准文本信息对模型进行优化，可以使得模型学习到训练目的所需求的知识，进而满足业务部署需求。

基于此，在得到目标上下文信息后，可以将目标上下文信息输入至初始语言模型进行处理，根据处理结果可以预测文本信息；此后可以先获取目标上下文信息对应的基准文本信息，进一步的再结合基准文本信息和预测文本信息对初始语言模型进行优化；优化完成之后可以检测优化后的初始语言模型是否
10 满足训练停止条件；若否，说明当前优化后的初始语言模型还不能够满足部署条件，因此可以在文档集合中筛选候选关联文档组成候选关联文档集合，并将候选关联文档集合作为所述关联文档集合，并执行根据所述关联文档集合中的关联文档以及关联文档之间的相似度，构建所述关联文档集合对应的文档关系图步骤，直至获得满足条件的模型后，即可将优化后的初始语言模型作为目标语言模型，以部署在业务场景下使用。

15 实际应用中，在利用生成的目标上下文信息对初始语言模型进行预训练时，其中，被训练的初始语言模型可以采用与现有预训练管道相同的模型架构，并且可以使用 AdamW 优化器和余弦学习率调度完成模型优化；还可以通过增加训练批次中的上下文信息长度，提升模型的上下文理解和推理能力。

沿用上例，在得到文档 1，文档 2，文档 3 和文档 5 拼接后的上下文信息后，可以基于上下文信息对 LLaMA 模型进行预训练。实现训练好的 LLaMA 模型可以具备更好的跨文档理解和推理能力。比如
20 用户提交问题{图像锐化处理包含哪些步骤}，通过将该问题输入至训练好的 LLaMA 模型后，此时 LLaMA 模型可以先确定关联该问题的多个文档，通过将多个文档进行拼接后，将结合用户提出的问题生成答案{1.选择图像；2.选择锐化方法；3.应用锐化滤镜...}，最后再将答案反馈给用户即可。

本实施例提供的语言模型训练方法，为了能够提高语言模型在处理复杂上下文和跨文档推理时性能，可以在获取到初始语言模型的文档集合后，在文档集合中筛选具有相似性的关联文档组成关联文档集合；
25 此后，为了加强模型对跨文档文本的理解能力，可以根据关联文档集合中的关联文档以及关联文档之间的相似度，构建关联文档集合对应的文档关系图；通过文档关系图体现关联文档之间的相似关系。在此基础上，可以在文档关系图中确定文档拼接路径，并按照文档拼接路径在关联文档集合中提取目标关联文档进行拼接处理，获得目标上下文信息；最后即可基于目标上下文信息对初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型。实现模型可以在相关文档序列上进行预训练，使得语言模型
30 能够在文档边界之外进行阅读和推理，从而有效的提升模型在复杂上下文任务中的性能。

参见图 3，图 3 示出了根据本说明书一个实施例提供的一种信息处理方法的流程图，具体包括以下步骤。

步骤 S302，获取客户端提交的问题信息，基于所述问题信息在业务知识库中查询多个答案文档；

步骤 S304，针对所述多个答案文档进行拼接，并将拼接结果和所述问题信息输入至目标语言模型进

行处理，获得答案信息，所述目标语言模型通过上述方法训练获得；

步骤 S306，将所述答案信息反馈至所述客户端。

本实施例提供的信息处理方法中关于目标语言模型的描述，均可参见上述实施例中的描述，本实施例在此不做过多赘述。

5 下述结合附图 4，以本说明书提供的信息处理方法和语言模型训练方法在问答交互场景中的应用为例，对所述信息处理方法和语言模型训练方法进行进一步说明。其中，图 4 示出了本说明书一个实施例提供的一种信息处理方法的处理过程流程图，具体包括以下步骤。

步骤 S402，获取初始语言模型的文档集合，利用文本检索模型构建文档集合中各个初始文档对应的文档嵌入向量。

10 步骤 S404，计算各个初始文档对应的文档嵌入向量之间的余弦相似度。

步骤 S406，根据余弦相似度计算结果在文档集合中筛选关联文档，组成关联文档集合。

步骤 S408，根据关联文档集合中的关联文档构建图节点，以及根据关联文档集合中关联文档之间的相似度构建图边。

15 步骤 S410，基于图边建立所述图节点之间的连通关系，根据连通关系建立结果生成关联文档集合对应的文档关系图。

步骤 S412，采用贪心算法在文档关系图中确定文档拼接路径。

步骤 S414，确定文档拼接路径中包含的拼接图节点，基于拼接图节点在关联文档集合中提取目标关联文档。

20 步骤 S416，将目标关联文档按照文档拼接路径的路径方向信息进行拼接，获得匹配初始语言模型输入条件的目标上下文信息。

步骤 S418，基于目标上下文信息对初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型，并针对目标语言模型进行部署。

步骤 S420，在接收到客户端提交的问题信息的情况下，基于问题信息在业务知识库中查询多个答案文档；

25 步骤 S422，针对多个答案文档进行拼接，并将拼接结果和问题信息输入至目标语言模型进行处理，获得答案信息。

步骤 S424，将答案信息反馈至客户端。

30 综上所述，为了能够提高语言模型在处理复杂上下文和跨文档推理时性能，可以在获取到初始语言模型的文档集合后，在文档集合中筛选具有相似性的关联文档组成关联文档集合；此后，为了加强模型对跨文档文本的理解能力，可以根据关联文档集合中的关联文档以及关联文档之间的相似度，构建关联文档集合对应的文档关系图；通过文档关系图体现关联文档之间的相似关系。在此基础上，可以在文档关系图中确定文档拼接路径，并按照文档拼接路径在关联文档集合中提取目标关联文档进行拼接处理，获得目标上下文信息；最后即可基于目标上下文信息对初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型。实现模型可以在相关文档序列上进行预训练，使得语言模型能够在文档边界之外

进行阅读和推理，从而有效的提升模型在复杂上下文任务中的性能。

与上述方法实施例相对应，本说明书还提供了语言模型训练装置实施例，图 5 示出了本说明书一个实施例提供的一种语言模型训练装置的结构示意图。如图 5 所示，该装置包括：

5 获取模块 502，被配置为获取初始语言模型的文档集合，并在所述文档集合中筛选关联文档组成关联文档集合；

构建模块 504，被配置为根据所述关联文档集合中的关联文档以及关联文档之间的相似度，构建所述关联文档集合对应的文档关系图；

拼接模块 506，被配置为在所述文档关系图中确定文档拼接路径，并按照所述文档拼接路径在所述关联文档集合中提取目标关联文档进行拼接处理，获得目标上下文信息；

10 训练模块 508，被配置为基于所述目标上下文信息对所述初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型。

一个可选的实施例中，所述获取模块 502 进一步被配置为：

利用文本检索模型构建所述文档集合中各个初始文档对应的文档嵌入向量；计算各个初始文档对应的文档嵌入向量之间的余弦相似度；根据余弦相似度计算结果在所述文档集合中筛选关联文档，组成关
15 联文档集合。

一个可选的实施例中，所述获取模块 502 进一步被配置为：

利用预设的量化器对各个初始文档对应的文档嵌入向量进行量化处理，并根据量化后的文档嵌入向量构建文档索引；按照预设的搜索策略和所述文档索引针对所述文档集合进行搜索，根据搜索结果确定关联文档并组成关联文档集合。

20 一个可选的实施例中，所述构建模块 504 进一步被配置为：

根据所述关联文档集合中的关联文档构建图节点，以及根据所述关联文档集合中关联文档之间的相似度构建图边；基于所述图边建立所述图节点之间的连通关系，根据连通关系建立结果生成所述关联文档集合对应的文档关系图。

一个可选的实施例中，所述拼接模块 506 进一步被配置为：

25 在所述文档关系图中确定第 i 图节点，并构建所述第 i 图节点对应的初始文档路径，其中， i 为正整数， i 从 1 开始取值，且 i 最大值为所述文档关系图中的图节点数量；在所述第 i 图节点连接的邻居图节点中选择第 j 图节点，并基于所述第 j 图节点将所述初始文档路径更新为中间文档路径，其中， j 为正整数， j 从 1 开始取值，且 j 最大值为邻居图节点数量；判断所述第 i 图节点连接的邻居图节点中是否存在未被访问的邻居图节点；若否， j 自增 1，并执行在所述第 i 图节点连接的邻居图节点中选择第 j 图节点的步骤；若是， i 自增 1，并执行在所述文档关系图中确定第 i 图节点的步骤，直至 i 等于所述图节点数量的情况下，获得多个中间文档路径，在所述多个中间文档路径中提取满足文档拼接条件的中间文档路
30 径作为文档拼接路径。

一个可选的实施例中，所述拼接模块 506 进一步被配置为：

确定所述文档拼接路径中包含的拼接图节点，基于所述拼接图节点在所述关联文档集合中提取目标

关联文档；将所述目标关联文档按照所述文档拼接路径的路径方向信息进行拼接，获得匹配所述初始语言模型输入条件的目标上下文信息。

一个可选的实施例中，所述训练模块 508 进一步被配置为：

将所述目标上下文信息输入至所述初始语言模型进行处理，获得预测文本信息；基于所述目标上下文信息对应的基准文本信息和所述预测文本信息，对所述初始语言模型进行优化；检测优化后的初始语言模型是否满足训练停止条件；若否，在所述文档集合中筛选候选关联文档组成候选关联文档集合，并将所述候选关联文档集合作为所述关联文档集合，并执行根据所述关联文档集合中的关联文档以及关联文档之间的相似度，构建所述关联文档集合对应的文档关系图步骤；若是，将优化后的初始语言模型作为目标语言模型。

上述为本实施例的一种语言模型训练装置的示意性方案。需要说明的是，该语言模型训练装置的技术方案与上述的语言模型训练方法的技术方案属于同一构思，语言模型训练装置的技术方案未详细描述的细节内容，均可以参见上述语言模型训练方法的技术方案的描述。

与上述方法实施例相对应，本说明书还提供了信息处理装置实施例，图 6 示出了本说明书一个实施例提供的一种信息处理装置的结构示意图。如图 6 所示，该装置包括：

获取信息模块 602，被配置为获取客户端提交的问题信息，基于所述问题信息在业务知识库中查询多个答案文档；

模型预测模块 604，被配置为针对所述多个答案文档进行拼接，并将拼接结果和所述问题信息输入至目标语言模型进行处理，获得答案信息，所述目标语言模型通过上述方法训练获得；

反馈信息模块 606，被配置为将所述答案信息反馈至所述客户端。

上述为本实施例的一种信息处理装置的示意性方案。需要说明的是，该信息处理装置的技术方案与上述的信息处理方法的技术方案属于同一构思，信息处理装置的技术方案未详细描述的细节内容，均可以参见上述信息处理方法的技术方案的描述。

图 7 示出了根据本说明书一个实施例提供的一种计算设备 700 的结构框图。该计算设备 700 的部件包括但不限于存储器 710 和处理器 720。处理器 720 与存储器 710 通过总线 730 相连接，数据库 750 用于保存数据。

计算设备 700 还包括接入设备 740，接入设备 740 使得计算设备 700 能够经由一个或多个网络 760 通信。这些网络的示例包括公用交换电话网（PSTN, Public Switched Telephone Network）、局域网（LAN, Local Area Network）、广域网（WAN, Wide Area Network）、个域网（PAN, Personal Area Network）或诸如因特网的通信网络的组合。接入设备 740 可以包括有线或无线的任何类型的网络接口（例如，网络接口卡（NIC, network interface controller））中的一个或多个，诸如 IEEE802.11 无线局域网（WLAN, Wireless Local Area Network）无线接口、全球微波互联接入（Wi-MAX, Worldwide Interoperability for Microwave Access）接口、以太网接口、通用串行总线（USB, Universal Serial Bus）接口、蜂窝网络接口、蓝牙接口、近场通信（NFC, Near Field Communication）。

在本说明书的一个实施例中，计算设备 700 的上述部件以及图 7 中未示出的其他部件也可以彼此相

连接，例如通过总线。应当理解，图 7 所示的计算设备结构框图仅仅是出于示例的目的，而不是对本说明书范围的限制。本领域技术人员可以根据需要，增添或替换其他部件。

计算设备 700 可以是任何类型的静止或移动计算设备，包括移动计算机或移动计算设备（例如，平板电脑计算机、个人数字助理、膝上型计算机、笔记本计算机、上网本等）、移动电话（例如，智能手机）、可佩戴的计算设备（例如，智能手表、智能眼镜等）或其他类型的移动设备，或者诸如台式计算机或个
5 人计算机 (PC, Personal Computer) 的静止计算设备。计算设备 700 还可以是移动式或静止式的服务器。

其中，处理器 720 用于执行如下计算机可执行指令，该计算机可执行指令被处理器执行时实现上述语言模型训练方法或信息处理方法的步骤。

上述为本实施例的一种计算设备的示意性方案。需要说明的是，该计算设备的技术方案与上述的语言模型训练方法或信息处理方法的技术方案属于同一构思，计算设备的技术方案未详细描述的细节内容，
10 均可以参见上述语言模型训练方法或信息处理方法的技术方案的描述。

本说明书一实施例还提供一种计算机可读存储介质，其存储有计算机可执行指令，该计算机可执行指令被处理器执行时实现上述语言模型训练方法或信息处理方法的步骤。

上述为本实施例的一种计算机可读存储介质的示意性方案。需要说明的是，该存储介质的技术方案与上述的语言模型训练方法或信息处理方法的技术方案属于同一构思，存储介质的技术方案未详细描述
15 的细节内容，均可以参见上述语言模型训练方法或信息处理方法的技术方案的描述。

本说明书一实施例还提供一种计算机程序产品，包括计算机程序或指令，该计算机程序或指令被处理器执行时实现上述语言模型训练方法或信息处理方法的步骤。

上述为本实施例的一种计算机程序产品的示意性方案。需要说明的是，该计算机程序产品的技术方案与上述的语言模型训练方法或信息处理方法的技术方案属于同一构思，计算机程序产品的技术方案未
20 详细描述的细节内容，均可以参见上述语言模型训练方法或信息处理方法的技术方案的描述。

上述对本说明书特定实施例进行了描述。其它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的特定顺序或者连续顺序才能实现期望的结果。在某些
25 实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

所述计算机指令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可以包括：能够携带所述计算机程序代码的任何实体或装置、记录介质、U 盘、移动硬盘、磁碟、光盘、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等。需要说明的是，所述计算机可读介质包含的内容可以根据专利实践的要求进行适当的增减，例如在某些地区，
30 根据专利实践，计算机可读介质不包括电载波信号和电信信号。

需要说明的是，对于前述的各方法实施例，为了简便描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本说明书实施例并不受所描述的动作顺序的限制，因为依据本说明书实施例，某些步骤可以采用其它顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的

实施例均属于优选实施例，所涉及的动作和模块并不一定都是本说明书实施例所必须的。

在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其它实施例的相关描述。

5 以上公开的本说明书优选实施例只是用于帮助阐述本说明书。可选实施例并没有详尽叙述所有的细节，也不限制该发明仅为所述的具体实施方式。显然，根据本说明书实施例的内容，可作很多的修改和变化。本说明书选取并具体描述这些实施例，是为了更好地解释本说明书实施例的原理和实际应用，从而使所属技术领域技术人员能很好地理解和利用本说明书。

权 利 要 求

1.一种语言模型训练方法，其特征在于，包括：
获取初始语言模型的文档集合，并在所述文档集合中筛选关联文档组成关联文档集合；
根据所述关联文档集合中的关联文档以及关联文档之间的相似度，构建所述关联文档集合对应的文
5 档关系图；
在所述文档关系图中确定文档拼接路径，并按照所述文档拼接路径在所述关联文档集合中提取目标
关联文档进行拼接处理，获得目标上下文信息；
基于所述目标上下文信息对所述初始语言模型进行训练，直至获得满足训练停止条件的目标语言模
型。

10 2.根据权利要求1所述的语言模型训练方法，其特征在于，所述在所述文档集合中筛选关联文档组
成关联文档集合，包括：
利用文本检索模型构建所述文档集合中各个初始文档对应的文档嵌入向量；
计算各个初始文档对应的文档嵌入向量之间的余弦相似度；
根据余弦相似度计算结果在所述文档集合中筛选关联文档，组成关联文档集合。

15 3.根据权利要求2所述的语言模型训练方法，其特征在于，所述利用文本检索模型构建所述文档集
合中各个初始文档对应的文档嵌入向量步骤执行之后，还包括：
利用预设的量化器对各个初始文档对应的文档嵌入向量进行量化处理，并根据量化后的文档嵌入向
量构建文档索引；
按照预设的搜索策略和所述文档索引针对所述文档集合进行搜索，根据搜索结果确定关联文档并组
20 成关联文档集合。

4.根据权利要求1所述的语言模型训练方法，其特征在于，所述根据所述关联文档集合中的关联文
档以及关联文档之间的相似度，构建所述关联文档集合对应的文档关系图，包括：
根据所述关联文档集合中的关联文档构建图节点，以及根据所述关联文档集合中关联文档之间的相
似度构建图边；
25 基于所述图边建立所述图节点之间的连通关系，根据连通关系建立结果生成所述关联文档集合对应
的文档关系图。

5.根据权利要求1所述的语言模型训练方法，其特征在于，所述在所述文档关系图中确定文档拼接
路径，包括：
在所述文档关系图中确定第*i*图节点，并构建所述第*i*图节点对应的初始文档路径，其中，*i*为正整
30 数，*i*从1开始取值，且*i*最大值为所述文档关系图中的图节点数量；
在所述第*i*图节点连接的邻居图节点中选择第*j*图节点，并基于所述第*j*图节点将所述初始文档路
径更新为中间文档路径，其中，*j*为正整数，*j*从1开始取值，且*j*最大值为邻居图节点数量；
判断所述第*i*图节点连接的邻居图节点中是否存在未被访问的邻居图节点；
若否，*j*自增1，并执行在所述第*i*图节点连接的邻居图节点中选择第*j*图节点的步骤；
35 若是，*i*自增1，并执行在所述文档关系图中确定第*i*图节点的步骤，直至*i*等于所述图节点数量的

情况下，获得多个中间文档路径，在所述多个中间文档路径中提取满足文档拼接条件的中间文档路径作为文档拼接路径。

6.根据权利要求1所述的语言模型训练方法，其特征在于，所述按照所述文档拼接路径在所述关联文档集合中提取目标关联文档进行拼接处理，获得目标上下文信息，包括：

5 确定所述文档拼接路径中包含的拼接图节点，基于所述拼接图节点在所述关联文档集合中提取目标关联文档；

将所述目标关联文档按照所述文档拼接路径的路径方向信息进行拼接，获得匹配所述初始语言模型输入条件的目标上下文信息。

7.根据权利要求1所述的语言模型训练方法，其特征在于，所述基于所述目标上下文信息对所述初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型，包括：

10 将所述目标上下文信息输入至所述初始语言模型进行处理，获得预测文本信息；
基于所述目标上下文信息对应的基准文本信息和所述预测文本信息，对所述初始语言模型进行优化；
检测优化后的初始语言模型是否满足训练停止条件；

15 若否，在所述文档集合中筛选候选关联文档组成候选关联文档集合，并将所述候选关联文档集合作为所述关联文档集合，并执行根据所述关联文档集合中的关联文档以及关联文档之间的相似度，构建所述关联文档集合对应的文档关系图步骤；

若是，将优化后的初始语言模型作为目标语言模型。

8.一种信息处理方法，其特征在于，包括：
获取客户端提交的问题信息，基于所述问题信息在业务知识库中查询多个答案文档；
20 针对所述多个答案文档进行拼接，并将拼接结果和所述问题信息输入至目标语言模型进行处理，获得答案信息，所述目标语言模型通过权利要求1至7任一项所述方法训练获得；
将所述答案信息反馈至所述客户端。

9.一种语言模型训练装置，其特征在于，包括：
获取模块，被配置为获取初始语言模型的文档集合，并在所述文档集合中筛选关联文档组成关联文档集合；
25 构建模块，被配置为根据所述关联文档集合中的关联文档以及关联文档之间的相似度，构建所述关联文档集合对应的文档关系图；

拼接模块，被配置为在所述文档关系图中确定文档拼接路径，并按照所述文档拼接路径在所述关联文档集合中提取目标关联文档进行拼接处理，获得目标上下文信息；

30 训练模块，被配置为基于所述目标上下文信息对所述初始语言模型进行训练，直至获得满足训练停止条件的目标语言模型。

10.一种信息处理装置，其特征在于，包括：
获取信息模块，被配置为获取客户端提交的问题信息，基于所述问题信息在业务知识库中查询多个答案文档；

模型预测模块,被配置为针对所述多个答案文档进行拼接,并将拼接结果和所述问题信息输入至目标语言模型进行处理,获得答案信息,所述目标语言模型通过权利要求 1 至 7 任一项所述方法训练获得;
反馈信息模块,被配置为将所述答案信息反馈至所述客户端。

11.一种计算设备,其特征在于,包括:

5 存储器和处理器;

所述存储器用于存储计算机可执行指令,所述处理器用于执行所述计算机可执行指令,该计算机可执行指令被处理器执行时实现权利要求 1 至 8 任意一项所述方法的步骤。

12.一种计算机可读存储介质,其特征在于,其存储有计算机可执行指令,该计算机可执行指令被处理器执行时实现权利要求 1 至 8 任意一项所述方法的步骤。

10 13.一种计算机程序产品,其特征在于,包括计算机程序或指令,该计算机程序或指令被处理器执行时实现权利要求 1 至 8 任意一项所述方法的步骤。