

文本数据处理、文本表示、文本聚类的方法及设备

本申请要求于 2023 年 07 月 31 日提交中国专利局、申请号为 202310955488.2、申请名称为“文本数据处理、文本表示、文本聚类的方法及设备”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

本申请涉及计算机技术，尤其涉及一种文本数据处理、文本表示、文本聚类的方法及设备。

背景技术

文本表示作为自然语言处理领域的基础能力被广泛应用于下游的文本分类、聚类、检索、问答、推荐等任务中。高质量的文本表示对下游任务的效果至关重要。

传统的文本表示方法，如独热（one-hot）编码、词袋模型等，存在维度灾难和信息丢失等问题，难以准确地提取语义信息。随着深度学习技术的发展，基于深度学习模型的文本表示方法成为了热点，有效地解决了传统文本表示方法的问题，并在各种任务中取得了较好的效果。近些年来，大规模预训练语言模型（如 BERT 系列的模型、其他 LLM 等）的出现提供了更强的模型底座，基于预训练语言模型训练的文本表示模型的效果远优于传统的文本表示方法。但是，预训练语言模型作为一个通用型的模型，在预训练阶段的训练目标与文本表示模型的目标之间有着很大的差异。

为了提升文本表示质量，通常使用较大规模高质量人工标注的数据集，对预训练语言模型进行强监督训练，来提升模型的文本表示能力。然而人工标注数据耗时费力，难以获取大规模高质量的标注数据来训练模型，导致获得的文本表示模型的文本表示能力较差，文本表示质量较低，进而导致下游处理任务效果差。

发明内容

本申请提供一种文本数据处理、文本表示、文本聚类的方法及设备，用以解决文本表示模型的文本表示能力较差，文本表示质量较低，进而导致下游处理任务效果差的问题。

第一方面，本申请提供一种文本数据处理方法，包括：

将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；

采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练，得到文本表示模型，所述文本表示模型用于生成输入文本的文本表示。

第二方面，本申请提供一种文本表示方法，应用于服务器，包括：

接收端侧设备对文本表示模型的调用请求，所述调用请求包含待处理的文本信息；

将所述文本信息输入文本表示模型进行编码，得到所述文本信息的文本表示，其中所述文本表示模型是通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的；

向所述端侧设备输出所述文本信息的文本表示。

第三方面，本申请提供一种文本聚类方法，包括：

获取待聚类的多个文本片段；

将所述多个文本片段分别输入文本表示模型进行编码，生成各所述文本片段的文本表示，所述文本表示模型是通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的；

根据各所述文本片段的文本表示进行聚类处理，得到文本聚类结果。

第四方面，本申请提供一种文本数据处理装置，包括：

划分模块，用于将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；

训练模块，用于采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练，得到文本表示模型，所述文本表示模型用于生成输入文本的文本表示。

第五方面，本申请提供一种文本表示装置，应用于服务器，包括：

接收模块，用于接收端侧设备对文本表示模型的调用请求，所述调用请求包含待处理的文本信息；

编码模块，用于将所述文本信息输入文本表示模型进行编码，得到所述文本信息的文本表示，其中所述文本表示模型是通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的；

输出模块，用于向所述端侧设备输出所述文本信息的文本表示。

第六方面，本申请提供一种文本聚类装置，包括：

获取模块，用于获取待聚类的多个文本片段；

编码模块，用于将所述多个文本片段分别输入文本表示模型进行编码，生成各所述文本片段的文本表示，所述文本表示模型是通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的；

聚类模块，用于根据各所述文本片段的文本表示进行聚类处理，得到文本聚类结果。

第七方面，本申请提供一种电子设备，包括：处理器，以及与所述处理器通信连接的存储器；所述存储器存储计算机执行指令；所述处理器执行所述存储器存储的计算机执行指令，以实现前述任一方面所述的方法。

第八方面，本申请提供一种计算机可读存储介质，所述计算机可读存储介质中存储有计算机执行指令，当所述计算机执行指令被处理器执行时用于实现前述任一方面所述的方法。

第九方面，本申请提供一种计算机程序产品，包括计算机程序，所述计算机程序被计算机执行时，实现前述任一方面所述的方法。

本申请提供的文本数据处理、文本表示、文本聚类的方法及设备，通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集，采用配置的多阶段训练策略，使用多个不同数据等级的训练集，对预训练模型进行多阶段训练，得到文本表示模型，相较于基于全部训练数据对预训练模型进行一次性（一个阶段）的训练，基于多个不同数据等级的训练集分阶段地对预训练模型进行多阶段训练，更能提升预训练模型的表示能力，从而提升训练得到的文本表示模型的文本表示能力，提升文本表示质量，进而可提升下游文本处理任务的效果。

附图说明

此处的附图被并入说明书中并构成本说明书的一部分，示出了符合本申请的实施例，并与说明书一起用于解释本申请的原理。

图1为本申请所适用的一示例系统架构的示意图；

图2为本申请一示例性实施例提供的文本数据处理方法流程图；

图3为本申请一示例性实施例提供的文本数据处理方法的框架图；

图 4 为本申请另一示例性实施例提供的文本数据处理方法流程图；

图 5 为本申请一示例性实施例提供的构建弱监督训练集的流程图；

图 6 为本申请一示例性实施例提供的文本表示方法的流程图；

图 7 为本申请一示例性实施例提供的文本聚类方法流程图；

图 8 为本申请一示例性实施例提供的文本检索方法流程图；

图 9 为本申请实施例提供的一种电子设备的结构示意图。

通过上述附图，已示出本申请明确的实施例，后文中将有更详细的描述。这些附图和文字描述并不是为了通过任何方式限制本申请构思的范围，而是通过参考特定实施例为本领域技术人员说明本申请的概念。

具体实施方式

这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本申请相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本申请的一些方面相一致的装置和方法的例子。

首先对本申请所涉及的名词进行解释：

文本表示 (Text Embedding)：也即文本向量表示，是指将文本转换成向量形式的表示方法。这种表示方法通过对文本进行数学上的编码，将文本转换为一组连续数值，形成向量表示。

Transformer：是一种基于注意力机制的神经网络模型，用于处理自然语言处理任务。它包括多层编码器和解码器，其中编码器将输入序列转换为一系列隐藏状态，并通过注意力机制将重点放在输入序列的不同部分上。解码器使用类似的机制，但它还使用自注意力机制来确保输出序列的流畅性和一致性。

对比学习 (Contrastive Learning)：是一种机器学习方法，用于学习相似性和差异性。该方法通过将同类样本进行聚合，将不同类样本进行区分，从而学习相似性表示。

大规模预训练语言模型：在大规模语料库上对大规模语言模型 (Large Language Model, 简称 LLM) 进行预训练后得到的预训练模型，用于自然语言处理任务。这些模型一般包含多层神经网络，其输入是一个文本序列，输出是对该文本序列的语言表示。预训练意味着在特定任务之前，模型已经被训练并预先学会处理大量的语言数据。通过预先训练模型，使得模型可以捕捉到更加复杂的语言和语义规则，从而在各种自然语言处理任务中表现出

色，并减少对特定任务的大规模数据需求。常见的大规模预训练语言模型如 BERT，以及其他多种 LLM。

BERT：全称是 Bidirectional Encoder Representations from Transformer，是一种基于 Transformer 的双向编码器表示模型，是一个预训练的语言表征模型。

视觉问答任务：根据输入的图像和问题，从输入图像的视觉信息中确定问题的答案。

图像描述任务：生成输入图像的描述文本。

视觉蕴涵任务：预测输入图像和文本在语义上的相关性，即蕴涵、中性或矛盾。

指代表达与理解任务：根据输入文本定位输入图像中与输入文本对应的图像区域。

图像生成任务：基于输入的描述文本生成图像。

基于文本的情感分类任务：预测输入文本的情感分类信息。

文本摘要任务：生成输入文本的摘要信息。

多模态任务：是指输入输出数据涉及图像和文本等多种模态数据的下游任务，例如视觉问答任务、图像描述任务、视觉蕴涵任务、指代表达与理解任务、图像生成任务等。

多模态预训练模型：是指输入输出数据涉及图像和文本等多种模态数据的预训练模型，经过微调训练后可以应用于多模态任务处理。

大模型是指具有大规模模型参数的深度学习模型，通常包含上亿、上百亿、甚至上千亿的模型参数。大模型又可以称为基石模型/基础模型（Foundation Model），通过大规模无标注的语料进行大模型的预训练，产出亿级以上参数的预训练模型，这种模型能适应广泛的下游任务，模型具有较好的泛化能力，例如大规模语言模型（LLM）、多模态预训练模型（Multi-modal Pre-training Model）等。大模型可以广泛应用于自然语言处理（Natural Language Processing，简称 NLP）、计算机视觉等领域，具体可以应用于如视觉问答（Visual Question Answering，简称 VQA）、图像描述（Image Caption，简称 IC）、图像生成等计算机视觉领域任务，以及基于文本的情感分类、文本摘要生成、机器翻译等自然语言处理领域任务，大模型主要的应用场景包括数字助理、智能机器人、搜索、在线教育、办公软件、电子商务、智能设计等。

预训练语言模型作为一个通用型的模型，在预训练阶段的训练目标与文本表示模型的目标之间有着很大的差异。为了提升文本表示质量，需使用大规模高质量人工标注的数据集，对预训练语言模型进行强监督训练，来提升模型的文本表示能力。然而人工标注数据耗时费力且标注一致性较低，难以获取大规模高质量的标注数据来训练模型，因此基于预训练语言模型训练得到的文本表示模型的文本表示能力较差，文本表示质量较低，进而导致下游处理任务效果差。

本申请提供一种文本数据处理方法，通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集，采用配置的多阶段训练策略，使用多个不同数据等级的训练集，对预训练模型进行多阶段训练，得到文本表示模型，其中，训练集的数据等级表示训练集的质量优劣，也即表示训练集中训练数据的质量的优劣。较高数据等级的训练集的质量优于较低数据等级低的训练集的质量，使用较高数据等级的训练集进行模型训练的效果更好。本申请的方法，通过用于训练的文本数据划分成多个不同数据等级的训练集，基于多个不同数据等级的训练集，分阶段地对预训练模型进行多阶段训练，相较于基于全部训练数据对预训练模型进行一次性（一个阶段）的训练，能够提升预训练模型的表示能力，从而提升训练得到的文本表示模型的文本表示能力，提升文本表示质量，进而可提升下游文本处理任务的效果。

本申请实施例中，用于训练获得文本表示模型的预训练模型，是能够实现文本表示的机器学习模型，具体可以是预训练语言模型，如 BERT、或其他任意预训练的大规模语言模型 LLM，或基于大规模预训练语言模型所构建的模型，此处不做具体限定。

通过多阶段训练得到的文本表示模型，可以应用于如下任一种基于文本的处理任务：语义检索、聚类、推荐、分类、问答。在下游的文本处理任务中文本表示模型用于生成给定文本的文本表示。进一步地，基于给定文本的文本表示，实现语义检索、聚类、推荐、分类、问答等处理任务。

另外，文本表示模型还可以应用于基于多模态的处理任务中，实现任务中文本信息的编码，获得高质量的文本表示。基于高质量的文本表示，结合其他模态数据的表示，可以实现多模态的处理任务，提升多模态的处理任务的效果。

另外，文本表示模型还可以作为下游任务模型的底座模型，基于该底座模型的文本表示能力，构建用于执行下游任务的目标任务模型。

例如，将文本表示模型作为底座模型，在底座模型的最后一层之后叠加分类器，得到文本分类模型。文本表示模型作为文本分类模型中的文本表示模块。在文本分类任务中，通过文本分类模型中的文本表示模块（也即多阶段训练得到的文本表示模型）获得输入文本的文本表示，通过分类器基于输入文本的文本表示进行分类预测，得到输入文本的分类结果。另外，基于文本表示模型的底座模型，还可以构建应用于智能问答、多模态任务等其他下游任务的模型，此处不做具体限定。

图 1 为本申请所适用的一示例系统架构的示意图。如图 1 所示，该系统架构包括服务器和端侧设备。其中，服务器与端侧设备之间具有可通信的通信链路，能够实现服务器与端侧设备间的通信连接。

其中，服务器可以是部署在云端的服务器集群、或者本地具有计算能力的设备。服务器用于获取并存储预先配置的多阶段训练策略，以及多个不同数据等级的训练集。其中，多个不同数据等级的训练集是通过将用于训练的文本数据进行多等级划分得到的。服务器还负责采用配置的多阶段训练策略，使用多个不同数据等级的训练集，对预训练模型进行多阶段训练，得到文本表示模型。另外，服务器还可以负责构建大规模用于训练的文本数据，并基于预设规则将文本数据进行多等级划分，得到多个不同数据等级的训练集。进一步地，对于获得文本表示模型，服务器可以基于获得的文本表示模型对外提供文本表示服务；或者，服务器还可以向端侧设备提供文本表示模型或文本表示模型的模型参数，端侧设备可以获得文本表示模型，或者基于获得的文本表示模型的模型参数初始化预训练模型，获得文本表示模型。图 1 中用 m 表示不同数据等级的训练集的数量， m 为正整数。

端侧设备负责执行基于文本表示模型的文本表示能力实现的下游任务，如基于文本的语义检索、聚类、推荐、分类、问答等下游任务。端侧设备可以是运行下游任务系统的电子设备，具体可以为具有网络通信功能、运算功能以及信息显示功能的硬件设备，其包括但不限于智能手机、平板电脑、台式电脑、本地服务器、云端服务器等。

基于图 1 所示的系统架构，端侧设备在运行基于文本的处理任务过程中，需要获取文本信息的文本表示时，向服务器发送对文本表示模型的调用请求，该调用请求包含待处理的文本信息。服务器接调用请求，获取待处理的文本信息，将待处理的文本信息输入文本表示模型进行编码，得到文本信息的文本表示，并向端侧设备返回文本信息的文本表示。端侧设备接收服务器返回的文本信息的文本表示，根据文本信息的文本表示继续执行下游任务的处理逻辑，实现下游的处理任务。

另外，服务器还可以将文本表示模型（或文本表示模型的模型参数）提供给端侧设备，端侧设备可以在本地使用文本表示模型，基于文本表示模型的文本表示能力实现具体的下游任务，如基于文本的语义检索、聚类、推荐、分类、问答等任务。当然，文本表示模型还可以应用于其他自然语言处理和人工智能领域的任务，此处不做具体限定。

下面以具体地实施例对本申请的技术方案以及本申请的技术方案如何解决上述技术问题进行详细说明。下面这几个具体的实施例可以相互结合，对于相同或相似的概念或过程可能在某些实施例中不再赘述。下面将结合附图，对本申请的实施例进行描述。

图 2 为本申请一示例性实施例提供的文本数据处理方法流程图。本实施例的执行主体为前述系统架构中的服务器。如图 2 所示，该方法具体步骤如下：

步骤 S201、将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集。

其中，文本数据是预先获取的用于对预训练模型的文本表示能力进行训练的训练数据，

包括但不限于大规模/超大规模的弱监督数据、较小规模的强监督数据。其中，弱监督数据可以从公开数据集中获取，还可以通过各种合理途径/方式尽可能地搜集公开的文本语料，从公开文本语料中抽取关键文本构成文本数据，例如抽取内容上相关的两个文本片段，构成文本对，作为文本数据。在构建弱监督数据时，不限定文本语料的领域、数据类型、数据来源，可以尽可能多地获取各个领域、各种来源、各种数据类型的文本语料，以增加弱监督数据的规模。本实施例对于弱监督数据的获取方式不做具体限定。另外，本实施例中使用小规模强监督数据即可获得很好的训练效果，因此可以从已有公开数据集中获取经过人工标注的文本数据，来获得强监督数据集。

本实施例中，对于用于训练的大规模的文本数据，服务器对文本数据进行多等级划分，将文本数据划分成多个训练集，不同的训练集对应不同的数据等级。其中，训练集的数据等级表示训练集的质量优劣，也即表示训练集中训练数据的质量的优劣。较高数据等级的训练集的质量优于较低数据等级低的训练集的质量，使用较高数据等级的训练集进行模型训练的效果更好。

在一可选实施方式中，该步骤具体可以采用如下方式实现：

根据文本数据的来源数据的使用行为数据，计算文本数据的使用指标；根据文本数据的使用指标，将文本数据划分成多个不同数据等级的训练集。其中，数据等级高的训练集中文本数据的使用指标大于数据等级低的训练集中文本数据的使用指标。例如，对于任意两个不同数据等级的训练集，其中第一数据等级的训练集中文本数据的使用指标的最小值，大于另一第二数据等级的训练集中文本数据的使用指标的最大值，第一数据等级高于第二数据等级。

其中，文本数据的来源数据是指抽取获得文本数据的文本语料，来源数据的使用行为数据是指文本语料的使用行为数据。对于不同的文本语料，使用行为数据所包含的数据项可以不同。文本语料的使用行为数据包括但不限于：点击、浏览、转发、收藏、评论、答复、引用、被检索等行为数据，还可以包括其他能够体现文本语料的使用情况的数据。

文本数据的使用指标是体现文本数据的来源数据被使用的情况的指标，可以根据文本数据的来源数据的使用行为数据计算得到。对于一个来源数据的多个不同类型的使用行为数据，可以分别计算不同的统计指标，作为使用指标。例如，可以计算文本数据的来源数据的点击率、浏览次数、转发次数、收藏次数、评论数、被引用次数、被检索到的次数等等。对于一个文本数据的多个使用指标，可以为不同类型的使用指标设置权重系统，对多个使用指标加权求和/加权平均，得到综合指标。

可选地，根据预先配置数据等级的数量，以及各个数据等级对应的指标范围，根据各

个文本数据的综合指标在哪个数据等级对应的指标范围内，来将文本数据划分到对应数据等级的训练集。

可选地，还可以根据各个文本数据的综合指标，对文本数据进行排序，根据排序结果，将文本数据划分成多个不同的训练集，再依次设置各个训练集的数据等级。例如，根据排序结果，将文本数据等分或按一定比例划分成多份，形成多个训练集。

该实施方式中，服务器基于文本数据的使用指标，根据预先配置的划分规则，将文本数据划分成多个不同数据等级的训练集，具体划分规则可以根据实际需要进行配置和调整，此处不做具体限定。

在另一可选实施方式中，该步骤具体可以采用如下方式实现：

根据文本数据的来源数据的数据类型，以及预先配置各数据类型对应的数据等级，将文本数据划分到对应数据等级的训练集。例如，对于从知识图谱、学术文献、问答数据中抽取的文本数据，可以划分到较高等级的训练集；对于从网上发布的文章、评论数据中抽取到文本数据，可以划分到较低等级的训练集。另外，各数据类型对应的数据等级可以根据实际需要进行配置和调整，此处不做具体限定。

在另一可选实施方式中，该步骤具体可以采用如下方式实现：

根据文本数据是否经过标注，将未经标注的文本数据划分到弱监督训练集，将经过标注的文本数据划分到强监督训练集，其中，弱监督训练集的数据等级低于强监督训练集的数据等级。

在实际应用中，经过标注是训练数据是高质量的强监督数据，未经过标注的通常是相对质量较低的弱监督数据。本实施方式中，根据文本数据是否经过标注，可以将文本数据自然地划分成两个训练集，经过标注的是强监督训练集，未经过标注的是弱监督训练集。其中，对于文本数据的具体标注方式可以根据实际应用场景的需要进行配置，此处不做具体限定。

在另一可选实施方式中，在训练优化预训练模型的文本表示能力时，用于训练的文本数据可以包括多个文本对，文本对包括两个内容相关的文本。使用文本对对预训练模型进行对比学习训练，可以很好地提升预训练模型的文本表示能力。该步骤具体可以采用如下方式实现：根据各文本对的相关度，将文本数据划分成多个不同数据等级的训练集，数据等级高的训练集中文本对的相关度大于数据等级低的训练集中文本对的相关度。

其中，各文本对的相关度可以使用现有的任意一种文本相关性分析的算法得到，例如，使用 TF-IDF (Term Frequency-Inverse Document Frequency, 词频-逆文本频率) 算法、文本匹配算法等，此处不做具体限定。

可选地，根据预先配置数据等级的数量，以及各个数据等级对应的相关度范围，根据各个文本对的相关度在哪个数据等级对应的相关度范围内，来将文本对划分到对应数据等级的训练集。可选地，还可以根据各个文本对的相关度，对文本对进行排序，根据排序结果，将文本对划分成多个不同的训练集，再依次设置各个训练集的数据等级。例如，根据排序结果，将文本对等分或按一定比例划分成多份，形成多个训练集。

该实施方式中，服务器基于文本对的相关度，根据预先配置的划分规则，将文本对划分成多个不同数据等级的训练集，具体划分规则可以根据实际需要进行配置和调整，此处不做具体限定。

本实施例中，对于文本数据进行等级划分所使用的划分策略，可以根据实际需要进行配置和调整，并不限于前述几种实施方式中所提及的划分策略，此处不做具体限定。

步骤 S202、采用配置的多阶段训练策略，使用多个不同数据等级的训练集，对预训练模型进行多阶段训练，得到文本表示模型，文本表示模型用于生成输入文本的文本表示。

其中，用于训练获得文本表示模型的预训练模型，是能够实现文本表示的机器学习模型，具体可以是预训练语言模型，如 BERT、或其他任意预训练的大规模语言模型 LLM，或基于大规模预训练语言模型所构建的模型，此处不做具体限定。

示例性地，预训练模型可以是预训练的 BERT。在进行文本表示时，将文本输入 BERT 进行编码，将 BERT 最后一层输出的第一个分词（token）位置（即 CLS 位置）当作输入文本的文本表示。

示例性地，预训练模型可以是预训练语言模型的最后一层 transformer 结构的输出层之后叠加平均池化层（mean pooling）得到。在进行文本表示时，将文本输入预训练语言模型，在预训练语言模型的最后一层 transformer 结构的输出层的表示上进行平均池化操作后，得到输入文本的文本表示。

本实施例中，多阶段训练策略中配置了基于多个不同数据等级的训练集，对预训练模型进行多阶段训练的训练方式。

可选地，多阶段训练策略中可以不限定训练阶段的数量，根据划分的不同数据等级的训练集，按数据等级由低到高的顺序，依次使用各训练集对预训练模型进行阶段训练，得到文本表示模型。通过先使用大规模低等级的文本数据对预训练模型进行训练，可以逐阶段地提升预训练模型的文本表示能力和特性；后续使用高等级的文本数据对预训练模型进行更加精准地训练，可以进一步提升预训练模型的文本表示能力，获得更强大的文本表示模型。

示例性地，文本数据划分成大规模/超大规模的弱监督训练集和小规模的强监督训练集，

先使用大规模/超大规模的弱监督训练集，对预训练模型进行弱监督对比学习训练，得到一阶段模型，以大幅提升预训练模型的文本表示能力和特性；再使用小规模**的**强监督训练集，对一阶段模型进行强监督对比学习训练，进一步提升模型的文本表示能力，得到具有强大文本表示能力的文本表示模型。

可选地，多阶段训练策略可以配置的训练阶段数量，以及各训练阶段使用的训练集。配置的多阶段训练策略中，不同数据等级的训练集是数量与训练阶段数量可以相等也可以不相等。一个训练阶段可以使用一个或者多个训练集进行训练，一个训练集应用于零个、一个或者多个训练阶段。该步骤中，采用配置的多阶段训练策略，依次使用各阶段对应的训练集对预训练模型进行训练，得到文本表示模型。

示例性地，假设文本数据划分成了5个数据等级的训练集，按照数据等级由低到高依次为S1、S2、S3、S4、S5。根据实际应用的需要，可以在多阶段训练策略中配置3个训练阶段，一阶段使用S1和S2，二阶段使用S2、S3和S4，三阶段使用S5。

本实施例中，多阶段训练策略可以根据实际应用场景的需要，实际获取到的训练集的数据特性进行配置和调整，本实施例此处不做具体限定。

在一可选实施例中，可以由获取文本表示模型的用户自定义配置或更新多阶段训练策略。具体地，响应于多阶段训练策略的配置/更新请求，服务器通过训练策略配置界面输出多阶段训练策略，以及各训练集的数据等级和描述信息。其中，训练集的描述信息可以包含但不限于数据集的数据的来源信息、领域、数据的组成结构/格式、适用场景、数据特点等相关信息，能够指导用户更加合理正确地适用训练集。用户可以通过训练策略配置界面上显示训练集的描述信息了解各个训练集的特性和数据等级，并自定义配置或更新多阶段训练策略。

进一步地，响应于对多阶段训练策略中任一项配置信息的编辑操作，服务器配置/更新多阶段训练策略。其中，配置信息包括：训练阶段数量、各训练阶段使用的训练集。用户可以对多阶段训练策略中的训练阶段数量进行修改，也可以调整任一训练阶段具体所使用的训练集。

在一可选实施例中，在对预训练模型进行任一阶段训练之后，服务器可以通过前端界面输出当前的训练结果。当前的训练结果用于指导调整多阶段训练策略。训练结果可以是体现当前训练效果的信息，例如模型的各类性能指标（如准确率、召回率等）、模型性能的稳定性指标（如性能提升/下降的幅度等）。用户通过前端界面输出的当前的训练结果，可以根据实际应用需求，决定是否继续训练、或调整多阶段训练策略并继续训练、或停止训练等。

进一步地，在用户更新多阶段训练策略之后，响应于多阶段训练策略的更新，服务器使用更新后的多阶段训练策略对预训练模型进行后续阶段的训练，支持及时地调整多阶段训练策略，提升文本表示模型的训练效果，进而提升文本表示模型的文本表示能力。

本实施例的方法，通过用于训练的文本数据划分成多个不同数据等级的训练集，基于多个不同数据等级的训练集，分阶段地对预训练模型进行多阶段训练，相较于基于全部训练数据对预训练模型进行一次性（一个阶段）的训练，能够提升预训练模型的表示能力，从而提升训练得到的文本表示模型的文本表示能力，提升文本表示质量，进而可提升下游文本处理任务的效果。

在一可选实施例中，在使用多个不同数据等级的训练集，对预训练模型进行多个阶段的训练，得到文本表示模型之后，还可以使用垂直领域的专属训练集，对文本表示模型进行微调训练，得到适用于垂直领域的文本表示模型。

其中，垂直领域的专属训练集可以由用户根据文本表示模型实际应用的垂直领域和实际任务场景的数据进行构建。

示例性地，用户通过端侧设备构建垂直领域的专属训练集，并向服务器提交垂直领域的专属训练集。服务器接收端侧设备提交的垂直领域的专属训练集，使用专属训练集对文本表示模型进行训练，得到适用于垂直领域的文本表示模型。进一步地，服务器向端侧设备提供适用于垂直领域的文本表示模型。

例如，服务器向端侧设备提供适用于垂直领域的文本表示模型的下载地址，端侧设备根据下载地址下载适用于垂直领域的文本表示模型，并在本地运行垂直领域的文本表示模型。例如，服务器向端侧设备发送适用于垂直领域的文本表示模型的模型参数，端侧设备使用接收到的模型参数初始化本地的文本表示模型，得到适用于垂直领域的文本表示模型。例如，服务器向端侧设备提供垂直领域的文本表示模型的应用程序编程接口（API），端侧设备在需要使用文本表示模型时，通过调用文本表示模型的 API 使用文本表示模型对传入的文本进行编码，获得文本的文本表示。

在另一典型实施例中，用于训练的文本数据包括多个文本对，文本对包括两个内容相关的文本。根据文本对是否经过标注，将未经标注的文本对划分到弱监督训练集，将经过标注的强相关的文本对划分到强监督训练集。在进行文本对的标注时，对文本对的内容是否具有较强相关性进行标注，去除不相关的文本对，保留强相关的文本对。弱监督训练集包括大规模/超大规模未经标注的文本对，强监督训练集包括较小规模经标注的强相关的文本对。弱监督训练集的数据等级低于强监督训练集的数据等级。

图3为本实施例提供的文本数据处理方法的框架图，本实施例中，如图3所示，使用

的训练数据包括大规模的弱监督训练集、较小规模的强监督训练集。进行多阶段训练时，在一阶段的训练中，使用大规模的弱监督训练集对预训练模型进行弱监督对比学习，提升预训练模型的文本表示能力和领域泛化能力，得到一阶段模型；进一步地，在二阶段的训练中，使用一阶段模型进行初始化，使用强监督训练集对一阶段模型进行强监督对比学习，进一步增强模型的文本表示能力，训练完成后得到具有较强文本表示能力的模型，作为文本表示模型，可以大大提升文本表示模型的文本表示能力，提升文本表示质量。

图4为本申请实施例提供的文本数据处理方法流程图。本实施例的执行主体为前述系统架构中的服务器。如图4所示，该方法具体步骤如下：

步骤S401、获取用于实现文本表示的预训练模型。

本实施例中，用于训练获得文本表示模型的预训练模型，是能够实现文本表示的机器学习模型，具体可以是大规模预训练语言模型，如BERT、或其他任意预训练的大规模语言模型LLM，或基于大规模预训练语言模型所构建的模型，此处不做具体限定。

示例性地，预训练模型可以是预训练的BERT。在进行文本表示时，将文本输入BERT进行编码，将BERT最后一层输出的第一个分词(token)位置(即CLS位置)当作输入文本的文本表示。

示例性地，预训练模型可以是预训练语言模型的最后一层transformer结构的输出层之后叠加平均池化层(mean pooling)得到。在进行文本表示时，将文本输入预训练语言模型，在预训练语言模型的最后一层transformer结构的输出层的表示上进行平均池化操作后，得到输入文本的文本表示。

本实施例中，基于已有的大规模语言模型、预训练模型等，构建用于实现文本表示的预训练模型，具体可以采用现有技术中构建文本表示模型的方式实现，预训练模型可以是现有技术中任意结构的待训练的文本表示模型，此处不做具体限定。

本实施例中，预先构建大规模/超大规模的弱监督数据集，以及小规模强监督数据集。弱监督训练集可以从类似公开数据集中获取；或者，还可以通过各种合理途径/方式尽可能地搜集公开的文本语料，从文本语料中抽取内容上相关的两个文本片段，构成文本对。本实施例对于弱监督训练集的获取方式不做具体限定。在构建弱监督训练集时，不限定文本语料的领域、数据类型，可以尽可能多地获取各个领域、各种来源、各种数据类型的文本语料。

强监督训练集(也即监督数据集)包含多个文本对，一个文本对包含两个相关的文本。强监督训练集所包含的文本对间的相关性是经过人工标注的，文本对的内容具有较强的相关性。本实施例中使用小规模强监督训练集即可获得很好的训练效果，因此可以从已有

公开数据集中获取经过人工标注的文本对，即可获得强监督训练集。

本实施例中，文本对包括：第一文本和与第一文本相关的第二文本。任一文本对可以表示为 (s_i, s_i^+) ， s_i 表示第一文本， s_i^+ 表示与 s_i 内容上相关的第二文本。

在训练预训练模型以获得文本表示模型的过程中，进行两个阶段的对比学习训练。在一阶段训练中，使用大规模/超大规模的弱监督训练集，对预训练模型进行弱监督对比学习训练，得到一阶段模型，可以非常有效地提升模型的文本表示能力和领域泛化能力。在二阶段训练中，使用强监督训练集，对一阶段模型进行强监督对比学习训练，得到文本表示模型，可以进一步提升模型的文本表示能力，得到具有很好文本表示能力和特性的文本表示模型，可以提升文本表示的质量。

步骤 S402、使用弱监督训练集，对预训练模型进行对比学习训练，得到一阶段模型。

该步骤中，对预训练模型进行一阶段训练，使用弱监督训练集所包含的未进行相关性标注的文本对，对预训练模型进行对比学习训练，实现弱监督对比学习训练，得到一阶段模型。

具体地，在使用未进行相关性标注的文本对，对预训练模型进行对比学习训练时，基于弱监督训练集所包含的文本对，对预训练模型进行多次迭代训练，每次迭代过程中采样批量（batch）文本对，基于采样的批量文本对，进行一次模型参数的优化，当满足迭代停止条件时，完成一阶段训练，获得一阶段模型。

其中，每次迭代采样文本对的批量大小可以表示为 K，K 的值可以根据实际应用场景和经验值进行设置，例如，K 可以设置为 128、300、512 等，本实施例对此不做具体限定。另外，迭代停止条件可以是迭代次数达到预设次数阈值、模型的效果相对稳定等，具体可以根据实际为应用场景和经验进行设置，本实施例对此不做具体限定。

在一可选实施方式中，在一次迭代训练过程中，对于任一文本对 (s_i, s_i^+) 中的第一文本 s_i ，将与第一文本 s_i 属于同一文本对的第二文本 s_i^+ ，作为与第一文本 s_i 相关的正样本 s_i^+ ，将与第一文本 s_i 属于不同文本对的第二文本，作为与第一文本 s_i 不相关的负样本（表示为 $s_{i,1}^-, s_{i,2}^-, \dots$ ）。基于所构建的第一文本 s_i 的正样本 s_i^+ 和负样本 $s_{i,1}^-, s_{i,2}^-, \dots$ ，对预训练模型进行对比学习训练，得到一阶段模型。

例如，以采样批量大小为 128 为例，对于采样得到的 128 个文本对中的任一文本对 (s_i, s_i^+) ，对于第一文本 s_i ，与第一文本 s_i 属于同一文本对的第二文本只有一个，第一文本 s_i 正样本只有一个，即 s_i^+ 。与第一文本 s_i 属于不同文本对的第二文本有 127 个，第一文本 s_i 的负样本有 127 个，包括： $s_{i,1}^-, s_{i,2}^-, \dots, s_{i,127}^-$ 。

在另一可选实施例中，在一次迭代训练过程中，对于任一文本对 (s_i, s_i^+) 中的第一文本

s_i , 将与第一文本 s_i 属于同一文本对的第二文本 s_i^+ , 作为与第一文本 s_i 相关的正样本 s_i^+ 。将与第一文本 s_i 属于不同文本对的第一文本和第二文本, 均作为与第一文本 s_i 不相关的负样本 (表示为 $s_{i,1}^-, s_{i,2}^-, \dots$)。基于所构建的第一文本 s_i 的正样本 s_i^+ 和负样本 $s_{i,1}^-, s_{i,2}^-, \dots$, 对预训练模型进行对比学习训练, 得到一阶段模型。

例如, 以采样批量大小为 128 为例, 对于采样得到的 128 个文本对中的任一文本对 (s_i, s_i^+) , 对于第一文本 s_i , 与第一文本 s_i 属于同一文本对的第二文本只有一个, 第一文本 s_i 正样本只有一个, 即 s_i^+ 。与第一文本 s_i 属于不同文本对的第一文本和第二文本共有 254 (127 × 2) 个, 因此第一文本 s_i 的负样本有 254 个, 包括: $s_{i,1}^-, s_{i,2}^-, \dots, s_{i,254}^-$ 。

在另一可选实施例中, 在一次迭代训练过程中, 对于任一文本对 (s_i, s_i^+) 中的第一文本 s_i , 将与第一文本 s_i 属于同一文本对的第二文本 s_i^+ , 作为与第一文本 s_i 相关的正样本 s_i^+ 。而第一文本 s_i 的负样本可以从其他不包含该第一文本 s_i 的文本对的第一文本和/或第二文本中采样指定数量的文本得到。其中, 对于一个第一文本采样负样本的指定数量可以根据实际应用场景进行设置, 此处不做具体限定。

进一步地, 进行迭代训练时, 将弱监督训练集所包含的文本对的两个文本分别输入预训练模型, 通过预训练模型生成文本对的两个文本的文本表示, 可以得到弱监督训练集所包含的任意文本的文本表示。

可选地, 在使用预训练模型生成文本对的两个文本的文本表示时, 还可以基于预训练模型构建双塔模型, 双塔模型包含两个并列的分支, 两个分支使用相同的预训练模型。将文本对的两个文本分别输入双塔模型的两个分支, 可以得到两个文本的文本表示。

进一步地, 对于任一文本对中的第一文本 s_i , 基于所构建的第一文本 s_i 的正样本 s_i^+ 和负样本 $s_{i,1}^-, s_{i,2}^-, \dots$, 根据第一文本 s_i 的文本表示和正样本 s_i^+ 的文本表示, 计算第一文本 s_i 与正样本 s_i^+ 间的相关度, 可以表示为 $\text{sim}(s_i, s_i^+)$ 。根据第一文本 s_i 的文本表示和任一负样本 $s_{i,j}^-$ 的文本表示, 计算第一文本 s_i 与负样本 $s_{i,j}^-$ 间的相关度, 可以表示为 $\text{sim}(s_i, s_{i,j}^-)$ 。根据第一文本 s_i 与正样本 s_i^+ 间的相关度 $\text{sim}(s_i, s_i^+)$ 和第一文本 s_i 与负样本 $s_{i,j}^-$ 间的相关度 $\text{sim}(s_i, s_{i,j}^-)$, 计算对比学习损失函数。以对比学习损失函数为优化目标, 基于对比学习损失函数, 训练预训练模型的参数, 以增大第一文本 s_i 与正样本 s_i^+ 间的相关度 $\text{sim}(s_i, s_i^+)$, 减小第一文本 s_i 与负样本 $s_{i,j}^-$ 间的相关度 $\text{sim}(s_i, s_{i,j}^-)$, 实现预训练模型的对比学习训练, 提升预训练模型的文本表示能力和特性。

示例性地, 对比学习损失函数可以表示为:

$$L = \sum_i^K L(s_i, s_i^+, s_{i,1}^-, s_{i,2}^-, \dots, s_{i,n}^-) = \sum_i^K \frac{e^{\text{sim}(s_i, s_i^+)}}{e^{\text{sim}(s_i, s_i^+)} + \sum_{j=1}^n e^{\text{sim}(s_i, s_{i,j}^-)}} \quad \text{其中, } n \text{ 表示第一文本 } s_i \text{ 的}$$

负样本的数量, e 为自然常数。sim() 表示两个文本之间的相关度, 具体可以是两个文本的文本表示的余弦相似度。

示例性地, 对比学习损失函数还是:

$$L = \sum_i^K L(s_i, s_i^+, s_{i,1}^-, s_{i,2}^-, \dots, s_{i,n}^-) = \sum_i^K \frac{\text{sim}(s_i, s_i^+)}{\text{sim}(s_i, s_i^+) + \sum_{j=1}^n \text{sim}(s_i, s_{i,j}^-)},$$

或者, $L = \sum_i^K L(s_i, s_i^+, s_{i,1}^-, s_{i,2}^-, \dots, s_{i,n}^-) = -\sum_i^K \log \frac{e^{\text{sim}(s_i, s_i^+)}}{e^{\text{sim}(s_i, s_i^+)} + \sum_{j=1}^n e^{\text{sim}(s_i, s_{i,j}^-)}}$, 其中 $\log$ 为对数函数。

本实施例中, 对比学习损失函数还可以基于 $L = \sum_i^K L(s_i, s_i^+, s_{i,1}^-, s_{i,2}^-, \dots, s_{i,n}^-) = \sum_i^K \frac{\text{sim}(s_i, s_i^+)}{\text{sim}(s_i, s_i^+) + \sum_{j=1}^n \text{sim}(s_i, s_{i,j}^-)}$ 变化而来, 具体可以根据实际需要和经验进行设置, 此处不做具体限定。

在经过多次迭代训练后, 当满足迭代停止条件时, 完成一阶段训练, 获得一阶段模型。

步骤 S403、使用强监督训练集, 对一阶段模型进行对比学习训练, 得到文本表示模型。

该步骤中, 对一阶段模型进行二阶段训练, 使用强监督训练集所包含的经相关性标注的文本对, 对一阶段模型进行对比学习训练, 实现强监督对比学习训练, 得到文本表示模型。

具体地, 在使用强监督训练集所包含的文本对, 对一阶段模型进行对比学习训练时, 基于文本对, 对一阶段模型进行多次迭代训练, 每次迭代过程中采样批量 (batch) 文本对, 基于采样的批量文本对, 进行一次模型参数的优化, 当满足迭代停止条件时, 完成二阶段训练, 获得文本表示模型。

在一可选实施例中, 对于强监督训练集中任一文本对中的第一文本, 将与第一文本属于同一文本对的第二文本, 作为与第一文本相关的正样本, 将与第一文本属于不同文本对的第二文本, 作为与第一文本不相关的负样本, 对一阶段模型进行对比学习训练, 得到文本表示模型。

在另一可选实施例中, 对于强监督训练集中任一文本对中的第一文本, 将与第一文本属于同一文本对的第二文本, 作为与第一文本相关的正样本; 将与第一文本属于不同文本对的第一文本和第二文本, 作为与第一文本不相关的负样本, 对一阶段模型进行对比学习训练, 得到文本表示模型。

在另一可选实施例中, 对于强监督训练集中任一文本对中的第一文本, 将与第一文本属于同一文本对的第二文本, 作为与第一文本相关的正样本; 而第一文本的负样本可以从其他不包含该第一文本的文本对的第一文本和/或第二文本中采样指定数量的文本得到, 对一阶段模型进行对比学习训练, 得到文本表示模型。其中, 对于一个第一文本采样负样本的指定数量可以根据实际应用场景进行设置, 此处不做具体限定。

该步骤中二阶段进行对比学习训练的过程与前述一阶段的对比学习训练类似，不同之处在于所使用的训练集和初始化的模型不同：一阶段使用弱监督训练集，初始化使用预训练模型，使用弱监督训练集对预训练模型进行训练；二阶段使用强监督训练集，初始化使用经一阶段训练后得到的一阶段模型，使用强监督训练集对一阶段模型进行训练。本实施例此处对于二阶段的具体训练过程不再赘述，可参照一阶段训练流程的相关描述。

需要说明的是，一阶段和二阶段的训练过程类似，但考虑到一阶段和二阶段所使用的数据集的大小不同，可以为两个阶段配置不同的批量（batch）大小、迭代停止条件，还可以配置不同的对比损失函数，以提升两个阶段的训练效果，获得具有很好的文本表示能力和特性的文本表示模型。

在二阶段训练完成后即可获得文本表示模型。进一步地，对于获得文本表示模型，服务器可以基于获得的文本表示模型对外提供文本表示服务。基于图1所示的系统架构，端侧设备在运行基于文本的处理任务过程中，需要获取文本信息的文本表示时，向服务器发送对文本表示模型的调用请求，该调用请求包含待处理的文本信息。服务器接调用请求，获取待处理的文本信息，将待处理的文本信息输入文本表示模型进行编码，得到文本信息的文本表示，并向端侧设备返回文本信息的文本表示。端侧设备接收服务器返回的文本信息的文本表示，根据文本信息的文本表示继续执行下游任务的处理逻辑，实现下游的处理任务。

另外，服务器还可以将文本表示模型提供给端侧设备，端侧设备可以在本地使用文本表示模型，基于文本表示模型的文本表示能力实现具体的下游任务，如基于文本的语义检索、聚类、推荐、分类、问答等任务。

示例性地，服务器向端侧设备提供文本表示模型的下载地址信息，端侧设备可以根据下载地址信息从服务器下载文本表示模型。或者，服务器可以向端侧设备下发文本表示模型的模型参数。端侧设备根据接收到的文本表示模型的模型参数，初始化文本表示模型，得到具有强大文本表示能力的文本表示模型。进一步地，端侧设备可以在本地运行文本表示模型，并基于文本表示模型的强大的文本表示能力，实现具体的下游任务。

本申请的方法，通过构建大规模/超大规模未进行相关性标注的文本对，作为弱监督训练集，并获取较小规模经过相关性标注的文本对，作为强监督训练集。在获得文本表示模型时，在一阶段的训练中，使用弱监督训练集对预训练模型进行弱监督对比学习，提升预训练模型的文本表示能力和领域泛化能力，得到一阶段模型；进一步地，在二阶段的训练中，使用一阶段模型进行初始化，使用强监督训练集对一阶段模型进行强监督对比学习，进一步增强模型的文本表示能力，训练完成后得到具有较强文本表示能力的模型，作为文

本表示模型，可以大大提升文本表示模型的文本表示能力，提升文本表示质量，进而可提升下游文本处理任务的效果。

图5为本申请一示例性实施例提供的构建弱监督训练集的流程图。在前述实施例的基础上，本实施例对构建大规模/超大规模弱监督训练集的过程进行详细地说明。如图5所示，构建包含未进行相关性标注的文本对的弱监督训练集的具体流程如下：

步骤S501、搜集多种不同数据类型的文本语料。

在构建弱监督训练集时，首先搜集多种不同数据类型的文本语料，从文本语料中抽取内容相关的文本对。

在搜集文本语料时，不限定文本语料的领域、数据类型、数据来源，尽可能多地通过合理的途径和方式，搜集多个不同领域、不同数据源、不同数据类型的公开的文本语料，以尽可能获得更大规模的文本语料，扩大文本语料覆盖的知识领域范围。

服务器搜集文本语料时，可以搜集公开的数据集、网页数据、百科数据、社区问答数据等，还可以获取其他应用平台/系统生产或构建的数据，如构建的知识图谱、产生的评论数据、问答数据、检索记录、开源的代码等等。服务器还可以采用任何现有的搜集文本语料的方式搜集更多的文本语料，此处对于文本语料的搜集方式和数据来源不做具体限定。

所获取的不同数据源的文本语料可以有不同的数据类型，包括但不限于文章类、问答类、评论数据、知识图谱、代码数据、文本链接等数据类型。

步骤S502、根据文本语料的数据类型，从文本语料中抽取文本对，作为用于训练的文本数据。

本实施例中，针对不同的数据类型的文本语料，可以使用不同的文本对抽取方法，抽取文本对。

示例性地，对于搜集到的文章类文本语料，通常包括文章的标题、摘要、概要、主体内容等。可以从文章类的文本语料中抽取同一文本语料中的主题和内容文本；将同一文本语料中的主题和内容文本分别作为第一文本和第二文本，生成文本对。其中，文章类文本语料的主题可以是标题、摘要、概要等，内容文本是指主体内容的全部或片段。另外，对于文章的主体内容较长的文本语料，还可以将文章的主体内容划分成多个内容片段，文章的标题（或摘要、概要）分别与文章主体内容的各个内容片段构成文本对。

示例性地，对于搜集到的问答类文本语料，抽取同一文本语料中的问题和问题的响应文本；将问题和问题的响应文本分别作为第一文本和第二文本，生成文本对。其中，问答类文本语料通常具有输入（query）-响应（response）类似的结构，基于该结构，分别将输入（query）和响应（response）作为第一文本和第二文本，构建文本对。

示例性地，对于搜集到的评论数据，通常包括发布的主题文本（如原创文本、或摘录文本），以及针对该主题文本的评论。每一主题文本与该主题文本的每一评论的内容相关，可以抽取评论数据中原始发布的主题文本和主题文本的评论文本；将主题文本和主题文本的评论文本分别作为第一文本和第二文本，生成文本对。

示例性地，对于搜集到的知识图谱，通常包括实体和实体的描述信息。每一实体的名称与该实体的描述信息具有一定相关性，可以获取知识图谱所包含实体的实体名称和实体描述；将同一实体的实体名称和实体描述分别作为第一文本和第二文本，生成文本对。

示例性地，对于搜集到的代码数据，包括代码和代码的注释文本。注释文本与对应代码在内容上相关，可以获取代码数据中的注释文本和注释文本对应的代码；将注释文本和注释文本对应的代码分别作为第一文本和第二文本，生成文本对。

示例性地，文本链接类型的数据是指给定一段文本作为引语（或引文），该引语链接到另一段更长的主体文本（如文献）。对于文本链接类型的数据，抽取引语（或引文）与链接到的主体文本，组成文本对。另外，也可以将引语（或引文）与链接到的主体文本的多个片段分别组成文本对。

本实施例中，通过搜集多种不同数据类型的文本语料，从文本语料中抽取文本对，可以构建大规模未经标注的文本对，构成大规模弱监督训练集，该弱监督训练集不仅包含大规模的文本对，而且能够覆盖广泛的领域知识。通过本实施例的方法可以构建包含超过 10 亿个文本对的超大规模的弱监督训练集。基于大规模的弱监督训练集对预训练模型进行弱监督对比学习训练，可以有效提升模型的文本表示能力和特性，并提升模型在各领域的泛化能力。

在任一方法实施例中，所使用的强监督训练集，可以通过获取公开的监督数据集的方式获得。另外，服务器也可以构建强监督训练集。在构建强监督训练集的过程中，获得文本对的方式与构建弱监督训练集的方式类似，进一步对文本对的相关性进行标注，去除不相关或相关性较弱的文本对，即可得到经相关性标注的强相关的文本对，从而得到强监督训练集。本实施例需小规模强监督训练集即可实现较好的训练效果，因此可以大大减少标注数据量，从而减少数据标注的时间和人力成本。

需要说明的是，为了降低人工标注的难度，可以对基于问答类数据抽取的文本对（包含问题-响应）进行标注，可以降低人工标注的难度，在一定程度上提升标注的效率。

下面各个实施例中，对通过采用配置的多阶段训练策略，使用多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的文本表示模型的应用进行示例性地说明。

图 6 为本申请一示例性实施例提供的文本表示方法的流程图。前述系统架构中的服务

器在获得文本表示模型之后，可以在本地运行文本表示模型，或者将文本表示模型部署到另一服务器上。本实施例的执行主体是运行文本表示模型的服务器。如图6所示，该方法具体步骤如下：

步骤 S601、接收端侧设备对文本表示模型的调用请求，调用请求包含待处理的文本信息。

示例性地，服务器可以对外提供文本表示模型的应用程序编程接口（API）。端侧设备可以通过该 API 调用文本表示模型。

端侧设备在调用文本表示模型的 API 时，服务器接收到端侧设备对文本表示模型的调用请求之后，从调用请求中提取待处理的文本信息。

步骤 S602、将文本信息输入文本表示模型进行编码，得到文本信息的文本表示，其中文本表示模型是通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的。

在获取到待处理的文本信息之后，将文本信息输入文本表示模型，通过文本表示模型将文本信息编码为文本表示。

本实施例中所使用的文本表示模型，是通过前述方法实施例的方法，通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的，具有很好的文本表示能力和特性，可以提升文本表示的质量。获得文本表示模型的具体实现方式和技术效果参见前述实施例中相关内容，此处不再赘述。

步骤 S603、向端侧设备输出文本信息的文本表示。

在通过文本表示模型编码得到文本信息的文本表示之后，服务器将文本信息的文本表示返回给端侧设备。

本实施例中，运行有文本表示模型的服务器对外提供文本表示模型的调用接口。端侧设备在实现各类基于文本的处理任务的过程中，在需要获取文本信息的文本表示时，可以通过该接口调用文本表示模型，使用文本表示模型的文本表示能力。

本实施例中的方法可以应用于各类基于文本表示模型的文本表示能力实现的下游任务，如基于文本的语义检索、聚类、推荐、分类、问答等任务。当然，文本表示模型还可以应用于其他自然语言处理和人工智能领域的任务，此处不做具体限定。

在其他实施例中，前述系统架构中的服务器在获取到文本表示模型之后，还可以将文本表示模型提供给端侧设备。端侧设备可以在本地使用文本表示模型，基于文本表示模型

的文本表示能力实现具体的下游任务，如基于文本的语义检索、聚类、推荐、分类、问答等任务。当然，文本表示模型还可以应用于其他自然语言处理和人工智能领域的任务，此处不做具体限定。

图7为本申请一示例性实施例提供的文本聚类方法流程图。本实施例的执行主体可以是运行有文本表示模型的端侧设备或服务器。如图7所示，该方法的具体步骤如下：

步骤S701、获取待聚类的多个文本片段。

应用于不同的聚类场景时，根据实际需要聚类的对象，可以将该对象相关的文本作为待聚类的文本片段。具体可以是对对象的描述文本、对象包含的关键文本等，此处不做具体限定。

示例性地，以需要对应用平台上发布的文章进行聚类为例，可以将这些文章的标题作为待聚类的文本片段，对文章的标题进行聚类，将具有相似内容的标题聚到一类，从而将具有相似内容标题的文章聚到一类。

示例性地，以新闻分析为例，通过将新闻按照主题进行聚类，可以更好地了解社会热点新闻。将新闻的主题作为待聚类的文本片段，对新闻的主题进行聚类，将具有相似主题的新闻聚到一类。

步骤S702、将多个文本片段分别输入文本表示模型进行编码，生成各文本片段的文本表示，文本表示模型是通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的。

在获取到待聚类的多个文本片段之后，将各个文本片段分别输入文本表示模型，通过文本表示模型将输入的文本片段编码为文本表示，即可得到各个文本片段的文本表示。

本实施例中所使用的文本表示模型，是通过前述方法实施例的方法，通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的，具有很好的文本表示能力和特性，可以提升文本表示的质量。获得文本表示模型的具体实现方式和技术效果参见前述实施例中相关内容，此处不再赘述。

步骤S703、根据各文本片段的文本表示进行聚类处理，得到文本聚类结果。

在获取到待聚类的多个文本片段的文本表示之后，使用聚类算法，根据文本片段的文本表示间的相似度，对多个文本片段进行聚类，将语义相似的文本片段聚到同一类中，得到文本聚类结果。

其中，所使用的聚类算法可以根据实际应用场景的需求进行选择 and 配置，此处不做具

体限定。

本实施例中，将前述实施例获得的文本表示模型，应用于文本聚类任务中，通过该文本表示模型可以获得文本片段的高质量的文本表示，基于高质量的文本表示对文本片段进行聚类，可以提升文本聚类的准确度。

图8为本申请一示例性实施例提供的文本检索方法流程图。本实施例的执行主体可以是运行有文本表示模型的端侧设备或服务器。如图8所示，该方法的具体步骤如下：

步骤S801、响应于查询请求，获取查询文本。

本实施例中，查询请求是用户通过用户设备发送的请求，该请求包含用户输入的查询文本。接收到查询请求之后，从查询请求中解析得到查询文本。

步骤S802、将查询文本输入文本表示模型进行编码，生成查询文本的文本表示，其中文本表示模型是通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的。

在获取到用户输入的查询文本之后，将查询文本输入文本表示模型，通过文本表示模型将查询文本编码为文本表示，即可得到查询文本的文本表示。

本实施例中所使用的文本表示模型，是通过前述方法实施例的方法，通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的，具有很好的文本表示能力和特性，可以提升文本表示的质量。获得文本表示模型的具体实现方式和技术效果参见前述实施例中相关内容，此处不再赘述。

步骤S803、根据查询文本的文本表示和检索库中各候选文本的文本表示，在检索库中检索与查询文本语义匹配的候选文本，其中候选文本的文本表示是通过文本表示模型编码生成的。

本实施例中，可以使用文本表示模型预先构建检索库。检索库包括候选文本，以及候选文本的文本表示。在构建检索库时，将候选文本分别输入文本表示模型进行编码，得到候选文本的文本表示。

在进行文本检索时，根据查询文本的文本表示和检索库中各候选文本的文本表示，计算查询文本的文本表示与检索库中各候选文本的文本表示的相似度，作为查询文本与候选文本的语义匹配度，从检索库中检索到与查询文本的语义匹配度较高的至少一个候选文本，作为与查询文本语义匹配的候选文本。

其中，基于文本表示检索与查询文本语义匹配的候选文本，可以采用现有的任意一种

检索匹配策略，来确定与查询文本的语义匹配度较高的至少一个候选文本，具体检索匹配策略可以根据检索系统的实际需求进行配置，此处不做具体限定。

步骤 S804、输出与查询文本语义匹配的候选文本。

在检索到与查询文本语义匹配的候选文本之后，将检索结果输出，以将检索结果展示给用户。

本实施例中，将前述实施例获得的文本表示模型，应用于文本检索任务中，通过该文本表示模型可以获得查询文本和候选文本的高质量的文本表示，基于高质量的文本表示进行文本检索，可以提升文本检索的准确度。

对应于上述方法，本申请实施例还提供一种文本数据处理装置，包括：

划分模块，用于将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；

训练模块，用于采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练，得到文本表示模型，所述文本表示模型用于生成输入文本的文本表示。

本申请实施例还提供一种文本表示装置，应用于服务器，包括：

接收模块，用于接收端侧设备对文本表示模型的调用请求，所述调用请求包含待处理的文本信息；

编码模块，用于将所述文本信息输入文本表示模型进行编码，得到所述文本信息的文本表示，其中所述文本表示模型是通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的；

输出模块，用于向所述端侧设备输出所述文本信息的文本表示。

本申请实施例还提供一种文本聚类装置，包括：

获取模块，用于获取待聚类的多个文本片段；

编码模块，用于将所述多个文本片段分别输入文本表示模型进行编码，生成各所述文本片段的文本表示，所述文本表示模型是通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的；

聚类模块，用于根据各所述文本片段的文本表示进行聚类处理，得到文本聚类结果。

图 9 为本申请实施例提供的一种电子设备的结构示意图。如图 9 所示，该电子设备包括：存储器 901 和处理器 902。存储器 901，用于存储计算机执行指令，并可被配置为存储

其它各种数据以支持在电子设备上的操作。处理器 902，与存储器 901 通信连接，用于执行存储器 901 存储的计算机执行指令，以实现上述任一方法实施例所提供的技术方案，其具体功能和所能实现的技术效果类似，此处不再赘述。该电子设备可以是部署在云端的云服务器，也可以是本地的服务器或终端，图 9 中以该电子设备为云服务器为例进行示例，本实施例此处不做具体限定。

可选的，如图 9 所示，该电子设备还包括：防火墙 903、负载均衡器 904、通信组件 905、电源组件 906 等其它组件。图 9 中仅示意性给出部分组件，并不意味着电子设备只包括图 9 所示组件。

本申请实施例还提供一种计算机可读存储介质，计算机可读存储介质中存储有计算机执行指令，计算机执行指令被处理器执行时用于实现上述任一方法实施例所提供的技术方案，具体功能和所能实现的技术效果此处不再赘述。

本申请实施例还提供了一种计算机程序产品，计算机程序产品包括：计算机程序，计算机程序存储在可读存储介质中，电子设备的至少一个处理器可以从可读存储介质读取计算机程序，至少一个处理器执行计算机程序使得电子设备执行上述任一方法实施例所提供的技术方案，具体功能和所能实现的技术效果此处不再赘述。

本申请实施例提供一种芯片，包括：处理模块与通信接口，该处理模块能执行前述方法实施例中电子设备的技术方案。可选的，该芯片还包括存储模块（如，存储器），存储模块用于存储指令，处理模块用于执行存储模块存储的指令，并且对存储模块中存储的指令的执行使得处理模块执行前述任一方法实施例所提供的技术方案。

上述存储器可以是对象存储(Object Storage Service, OSS)。

上述存储器可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，如静态随机存取存储器 (SRAM)，电可擦除可编程只读存储器 (EEPROM)，可擦除可编程只读存储器 (EPROM)，可编程只读存储器 (PROM)，只读存储器 (ROM)，磁存储器，快闪存储器，磁盘或光盘。

上述通信组件被配置为便于通信组件所在设备和其他设备之间有线或无线方式的通信。通信组件所在设备可以接入基于通信标准的无线网络，如移动热点 (WiFi)，第二代移动通信系统 (2G)、第三代移动通信系统 (3G)、第四代移动通信系统 (4G)/长期演进 (LTE)、第五代移动通信系统 (5G) 等移动通信网络，或它们的组合。在一个示例性实施例中，通信组件经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中，通信组件还包括近场通信 (NFC) 模块，以促进短程通信。例如，在 NFC 模块可基于射频识别 (RFID) 技术，红外数据协会 (IrDA) 技术，超宽带 (UWB)

技术，蓝牙（BT）技术和其他技术来实现。

上述电源组件，为电源组件所在设备的各种组件提供电力。电源组件可以包括电源管理系统，一个或多个电源，及其他与为电源组件所在设备生成、管理和分配电力相关联的组件。

本领域内的技术人员应明白，本申请的实施例可提供为方法、系统、或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质（包括但不限于磁盘存储器、只读光盘存储器（CD-ROM）、光学存储器等）上实施的计算机程序产品的形式。

本申请是参照根据本申请实施例的方法、设备（系统）、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

在一个典型的配置中，计算设备包括一个或多个处理器（CPU）、输入/输出接口、网络接口和内存。

内存可能包括计算机可读介质中的非永久性存储器，随机存取存储器（RAM）和/或非易失性内存等形式，如只读存储器（ROM）或闪存（flash RAM）。内存是计算机可读介质的示例。

计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计

算机的存储介质的例子包括,但不限于相变内存 (PRAM)、静态随机存取存储器 (SRAM)、动态随机存取存储器 (DRAM)、其他类型的随机存取存储器 (RAM)、只读存储器 (ROM)、电可擦除可编程只读存储器 (EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器 (CD-ROM)、数字多功能光盘 (DVD) 或其他光学存储、磁盒式磁带, 磁带磁盘存储或其他磁性存储设备或任何其他非传输介质, 可用于存储可以被计算设备访问的信息。按照本文中的界定, 计算机可读介质不包括暂存电脑可读媒体 (transitory media), 如调制的数据信号和载波。

需要说明的是, 本申请所涉及的用户信息 (包括但不限于用户设备信息、用户属性信息等) 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。

另外, 在上述实施例及附图中的描述的一些流程中, 包含了按照特定顺序出现的多个操作, 但是应该清楚了解, 这些操作可以不按照其在本文中出现的顺序来执行或并行执行, 仅仅是用于区分开各个不同的操作, 序号本身不代表任何的执行顺序。另外, 这些流程可以包括更多或更少的操作, 并且这些操作可以按顺序执行或并行执行。需要说明的是, 本文中的“第一”、“第二”等描述, 是用于区分不同的消息、设备、模块等, 不代表先后顺序, 也不限定“第一”和“第二”是不同的类型。“多个”的含义是两个以上, 除非另有明确具体的限定。

本领域技术人员在考虑说明书及实践这里公开的内容后, 将容易想到本申请的其它实施方案。本申请旨在涵盖本申请的任何变型、用途或者适应性变化, 这些变型、用途或者适应性变化遵循本申请的一般性原理并包括本申请未公开的本技术领域中的公知常识或惯用技术手段。说明书和实施例仅被视为示例性的, 本申请的真正范围和精神由下面的权利要求书指出。

应当理解的是, 本申请并不局限于上面已经描述并在附图中示出的精确结构, 并且可以在不脱离其范围进行各种修改和改变。本申请的范围仅由所附的权利要求书来限制。

权 利 要 求 书

1、一种文本数据处理方法，其特征在于，包括：

将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；

采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练，得到文本表示模型，所述文本表示模型用于生成输入文本的文本表示。

2、根据权利要求1所述的方法，其特征在于，所述将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集，包括：

根据所述文本数据的来源数据的使用行为数据，计算所述文本数据的使用指标；

根据所述文本数据的使用指标，将所述文本数据划分成多个不同数据等级的训练集，数据等级高的训练集中文本数据的使用指标大于数据等级低的训练集中文本数据的使用指标。

3、根据权利要求1或2所述的方法，其特征在于，所述将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集，包括：

所述文本数据包括多个文本对，所述文本对包括两个内容相关的文本，

根据各所述文本对的相关度，将所述文本数据划分成多个不同数据等级的训练集，数据等级高的训练集中文本对的相关度大于数据等级低的训练集中文本对的相关度。

4、根据权利要求3所述的方法，其特征在于，还包括：

搜集多种不同数据类型的文本语料；

根据所述文本语料的数据类型，从所述文本语料中抽取文本对，作为用于训练的文本数据。

5、根据权利要求1-4中任一项所述的方法，其特征在于，所述将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集，包括：

根据所述文本数据是否经过标注，将未经标注的文本数据划分到弱监督训练集，将经过标注的文本数据划分到强监督训练集，其中，所述弱监督训练集的数据等级低于所述强监督训练集的数据等级。

6、根据权利要求1-5中任一项所述的方法，其特征在于，所述采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练，得到文本表示模型，包括：

所述多阶段训练策略包括配置的训练阶段数量，以及各训练阶段使用的训练集，

采用配置的多阶段训练策略，依次使用各阶段对应的训练集对预训练模型进行训练，

得到文本表示模型。

7、根据权利要求 1-6 中任一项所述的方法，其特征在于，所述采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练，得到文本表示模型，包括：

根据所述多个不同数据等级的训练集，按数据等级由低到高的顺序，依次使用各所述训练集对预训练模型进行阶段训练，得到文本表示模型。

8、根据权利要求 6 或 7 所述的方法，其特征在于，还包括：

所述文本数据包括多个文本对，所述文本对包括两个内容相关的文本，

对预训练模型进行任一阶段训练时，基于当前阶段所使用的训练集中的文本对，对于任一文本对中的第一文本，将与所述第一文本属于同一文本对的第二文本，作为与所述第一文本相关的正样本，将与所述第一文本属于不同文本对的第二文本，作为与所述第一文本不相关的负样本，对预训练模型进行对比学习训练。

9、根据权利要求 6 或 7 所述的方法，其特征在于，还包括：

响应于所述多阶段训练策略的配置/更新请求，通过训练策略配置界面输出所述多阶段训练策略，以及各训练集的数据等级和描述信息；

响应于对所述多阶段训练策略中任一项配置信息的编辑操作，配置/更新所述多阶段训练策略，所述配置信息包括：训练阶段数量、各训练阶段使用的训练集。

10、根据权利要求 9 所述的方法，其特征在于，还包括：

在对预训练模型进行任一阶段训练之后，通过前端界面输出当前的训练结果，所述当前的训练结果用于指导调整所述多阶段训练策略；

响应于所述多阶段训练策略的更新，使用更新后的多阶段训练策略对预训练模型进行后续阶段的训练。

11、根据权利要求 1-7 中任一项所述的方法，其特征在于，所述使用所述多个不同数据等级的训练集，对预训练模型进行多个阶段的训练，得到文本表示模型之后，还包括：

接收端侧设备提交的垂直领域的专属训练集；

使用所述专属训练集对所述文本表示模型进行训练，得到适用于所述垂直领域的文本表示模型；

向所述端侧设备提供所述适用于所述垂直领域的文本表示模型。

12、一种文本表示方法，其特征在于，应用于服务器，包括：

接收端侧设备对文本表示模型的调用请求，所述调用请求包含待处理的文本信息；

将所述文本信息输入文本表示模型进行编码，得到所述文本信息的文本表示，其中所

述文本表示模型是通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的；

向所述端侧设备输出所述文本信息的文本表示。

13、一种文本聚类方法，其特征在于，包括：

获取待聚类的多个文本片段；

将所述多个文本片段分别输入文本表示模型进行编码，生成各所述文本片段的文本表示，所述文本表示模型是通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的；

根据各所述文本片段的文本表示进行聚类处理，得到文本聚类结果。

14、一种文本数据处理装置，其特征在于，包括：

划分模块，用于将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；

训练模块，用于采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练，得到文本表示模型，所述文本表示模型用于生成输入文本的文本表示。

15、一种文本表示装置，其特征在于，应用于服务器，包括：

接收模块，用于接收端侧设备对文本表示模型的调用请求，所述调用请求包含待处理的文本信息；

编码模块，用于将所述文本信息输入文本表示模型进行编码，得到所述文本信息的文本表示，其中所述文本表示模型是通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的；

输出模块，用于向所述端侧设备输出所述文本信息的文本表示。

16、一种文本聚类装置，其特征在于，包括：

获取模块，用于获取待聚类的多个文本片段；

编码模块，用于将所述多个文本片段分别输入文本表示模型进行编码，生成各所述文本片段的文本表示，所述文本表示模型是通过将用于训练的文本数据进行等级划分，获得多个不同数据等级的训练集；采用配置的多阶段训练策略，使用所述多个不同数据等级的训练集，对预训练模型进行多阶段训练得到的；

聚类模块，用于根据各所述文本片段的文本表示进行聚类处理，得到文本聚类结果。

17、一种电子设备，其特征在于，包括：处理器，以及与所述处理器通信连接的存储器；

所述存储器存储计算机执行指令；

所述处理器执行所述存储器存储的计算机执行指令，以实现如权利要求 1-13 中任一项所述的方法。

18、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质中存储有计算机执行指令，当所述计算机执行指令被处理器执行时用于实现权利要求 1-13 中任一项所述的方法。

19、一种计算机程序产品，其特征在于，包括计算机程序，所述计算机程序被计算机执行时，实现如权利要求 1-13 中任一项所述的方法。