文本切分方法、装置、电子设备及计算机可读存储介质

相关申请的交叉引用

本公开要求于2025年02月25日提交中国专利局的申请号为202510210103.9、名称为“文本切分方法、装置、电子设备及计算机可读存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本公开中。

技术领域

本公开涉及计算机技术领域，具体而言，涉及一种文本切分方法、装置、电子设备及计算机可读存储介质。

背景技术

随着大模型在自然语言处理领域的突破性进展，基于大模型的问答系统克服了传统问答系统无法有效处理多样化的用户提问、机械回复等缺陷，因其能够理解更复杂的语言结构，捕捉更加丰富的上下文信息，从而为用户提供更加精准、流畅且个性化的回答，与用户互动更加人性化、智能化。

虽然大模型提升了问答系统的回复质量，但由于其训练过程中对特定领域知识覆盖有限，会存在生成的答案虽然看似合理、逻辑通顺，但实际并不准确或与事实相悖的缺陷。为此，基于检索增强生成（Retrieval-Augmented Generation，简称 RAG）方法，通过将外部知识检索与生成式大模型相结合，为基于大模型的问答系统提供了强大的支持。RAG 方法通过在生成回答前首先从数据库中检索相关信息，将检索到的知识作为上下文输入模型，从而增强大模型的知识覆盖范围，提升其准确性和可靠性。

而数据库中的信息通常来自于外部文档，需要先将外部文档切分成多个片段，再将每一片段转换成向量存储至数据库中，外部文档切分的合理与否会影响 RAG 检索到的上下文中信息的完整性，进而影响 RAG 反馈的问答输出的准确性和可靠性。

发明内容

本公开目的之一为提供一种文本切分方法、装置、电子设备及计算机可读存储介质，其能够按照语义对文本进行合理切分，保证切分后的语义段的语义连贯性，进而能够提高 RAG 反馈的问答输出的准确性和可靠性。

为了实现上述目的，本公开实施例采用的技术方案如下：

第一方面，本公开提供一种文本切分方法，所述方法包括：

获取针对目标文本进行句子级别切分后得到的多个句子；

以所述多个句子中的第一个句子为起点、最后一个句子为终点，多次滑动预设滑动窗口，并在每次滑动后，将处于所述预设滑动窗口内的多个目标句子按照语义切分，得到至

少一个语义段；

对于每次切分得到的语义段，将本次切分得到的语义段和上次切分得到的语义段进行聚类，得到上次滑动后和本次滑动后所述预设滑动窗口内的每一目标句子所属的目标语义段，直至得到所述多个句子中每一句子所属的语义段。

在可选的实施方式中，所述将处于所述预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段的步骤包括：

将所述多个目标句子输入预先训练的文本切分模型，得到两两相邻的目标句子组成的句子对的语义连续度；

对于任一目标句子对，若所述目标句子对的语义连续度大于预设值，则将所述目标句子对中的两个目标句子归属为同一个语义段，否则，将所述目标句子对中的两个目标句子归属为不同的语义段；

依次将每一句子对作为所述目标句子对，得到所述多个目标句子语义切分后的至少一个语义段。

在可选的实施方式中，本次滑动后预设滑动窗口内的第一个目标句子与上次滑动预设滑动窗口内的最后一个目标句子相同，所述预设滑动窗口内两两相邻的目标句子组成句子对；

所述对于每次切分得到的语义段，将本次切分得到的语义段和上次切分得到的语义段进行聚类，得到上次滑动后和本次滑动后所述预设滑动窗口内的每一目标句子所属的目标语义段的步骤包括：

获取本次滑动后所述预设滑动窗口内的首个句子对的语义连续度；

计算所述首个句子对中的两个目标句子之间的语义相似度；

根据所述首个句子对中的两个目标句子之间的语义连续度和语义相似度，计算所述首个句子对的语义关联得分；

若所述语义关联得分大于预设分值，则将上次切分得到的最后一个语义段和本次切分得到的首个语义段合并为一个语义段，否则，将上次切分得到的最后一个语义段和本次切分得到的首个语义段作为不同的语义段。

在可选的实施方式中，所述计算所述首个句子对中的两个目标句子之间的语义相似度，包括：分别提取两个目标句子的特征向量；计算两个目标句子的特征向量的余弦相似度，并将计算出的余弦相似度作为语义相似度。

在可选的实施方式中，所述计算所述首个句子对中的两个目标句子之间的语义相似度，包括：将两个目标句子输入多模态多任务嵌入模型，以得到两个目标句子之间的语义相似度。

在可选的实施方式中，所述首个句子对的语义关联得分的算式为：语义关联得分=语义连续度×连续度权重+语义相似度×相似度权重。

在可选的实施方式中，所述将处于所述预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段，包括：获取两两相邻的目标句子组成的句子对的语义连续度；计算两两相邻的目标句子组成的句子对的语义相似度；根据两两相邻的目标句子组成的句子对中的两个目标句子之间的语义连续度和语义相似度，计算两两相邻的目标句子组成的句子对的语义关联得分；若语义关联得分大于预设分值，则将对应句子对中的目标句子归属为同一个语义段，否则，将其归属为不同的语义段。

在可选的实施方式中，所述对于每次切分得到的语义段，将本次切分得到的语义段和上次切分得到的语义段进行聚类，得到上次滑动后和本次滑动后所述预设滑动窗口内的每一目标句子所属的目标语义段的步骤之后，所述方法还包括：

获取预设限制条件，所述预设限制条件配置成限制检索增强生成 RAG 系统中的大模型输入的语义段；

根据所述预设限制条件，对每一所述目标语义段进行处理，得到每一所述目标语义段对应的参考语义段，并将每一所述参考语义段以向量化的形式存储至向量数据库，以在所述大模型接收到输入数据时，从所述向量数据库中检索出与所述输入数据相关的语义段向量，并基于所述输入数据及检索出的语义段向量进行反馈。

在可选的实施方式中，所述预设限制条件包括大模型输入的单个的语义段的最大长度和/或大模型输入的相邻两个语义段之间的重叠度。

在可选的实施方式中，所述预设限制条件为预设长度，所述目标语义段中所有句子的长度之和大于所述预设长度；所述根据所述预设限制条件，对每一所述目标语义段进行处理，得到每一所述目标语义段对应的参考语义段的步骤包括：

针对任一所述目标语义段，若所述目标语义段不满足所述预设限制条件，则根据所述预设限制条件对所述目标语义段再次进行切分，得到所述目标语义段对应的参考语义段；

若所述目标语义段满足所述预设限制条件，则将所述目标语义段作为其对应的参考语义段。

在可选的实施方式中，所述根据所述预设限制条件对所述目标语义段再次进行切分，得到所述目标语义段对应的参考语义段的步骤包括：

将所述目标语义段中的第一个句子作为参考句子；

从所述参考句子开始，从所述目标语义段中确定目标句子序列，其中，所述目标句子序列中的所有句子的长度之和最大且小于所述预设长度；

将所述目标句子序列中的所有句子归为一个参考语义段；

若所述目标句子序列中的最后一个句子非所述目标语义段的最后一个句子，则从所述目标语义段中重新确定参考句子，并返回所述从所述参考句子开始，从所述目标语义段中确定目标句子序列的步骤，直至得到所述目标语义段对应的所有参考语义段。

在可选的实施方式中，所述从所述目标语义段中重新确定参考句子的步骤包括：

从所述目标句子序列中确定第一关联句子；

从所述目标语义段中位于所述目标句子序列之后的句子中确定第二关联句子；

若所述第一关联句子的长度和所述第二关联句子的长度之和大于所述预设长度，则根据所述第二关联句子确定所述参考句子，否则根据所述第一关联句子确定所述参考句子。

第二方面，本公开提供一种文本切分装置，所述装置包括：

获取模块，配置成获取针对目标文本进行句子级别切分后得到的多个句子；

切分模块，配置成以所述多个句子中的第一个句子为起点、最后一个句子为终点，多次滑动预设滑动窗口，并在每次滑动后，将处于所述预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段；

聚类模块，配置成对于每次切分得到的语义段，将本次切分得到的语义段和上次切分得到的语义段进行聚类，得到上次滑动后和本次滑动后所述预设滑动窗口内的每一目标句子所属的目标语义段，直至得到所述多个句子中每一句子所属的语义段。

第三方面，本公开提供一种电子设备，包括处理器和存储器，所述存储器配置成存储程序，所述处理器配置成在执行所述程序时，实现如前述实施方式中任一项所述的文本切分方法。

第四方面，本公开提供一种计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现如前述实施方式中任一项所述文本切分方法。

相对于现有技术，公开本公开实施例所提供的文本切分方法、装置、电子设备及计算机可读存储介质，公开针对多个句子，滑动预设滑动窗口，依次对预设滑动窗口内的目标句子按照语义切分，保证预设滑动窗口内目标句子切分后的语义连贯性，并且将本次切分得到的语义段和上次切分得到的语义段进行聚类，保证相邻两次滑动后预设滑动窗口之间的句子的语义连贯性，最终实现了文本的合理切分，保证了切分后的语义段的语义连贯性，进而提高了 RAG 反馈的问答输出的准确性和可靠性。

附图说明

为了更清楚地说明本公开实施例的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，应当理解，以下附图仅示出了本公开的某些实施例，因此不应被看作是对范围的限定，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相关的附图。

图 1 为本公开实施例提供的 RAG 系统的框架图。

图 2 为本公开实施例提供的文本切分方法的流程示例图。

图 3 为本公开实施例提供的利用预设滑动窗口进行语义切分的过程示例图。

图 4 为本公开实施例提供的文本切分装置的方框示例图。

图 5 为本公开实施例提供的电子设备的方框示例图。

图标：10-电子设备；11-处理器；12-存储器；13-总线；100-文本切分装置；110-获取模块；120-切分模块；130-聚类模块。

具体实施方式

为使本公开实施例的目的、技术方案和优点更加清楚，下面将结合本公开实施例中的附图，对本公开实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本公开一部分实施例，而不是全部的实施例。通常在此处附图中描述和示出的本公开实施例的组件可以以各种不同的配置来布置和设计。

因此，以下对在附图中提供的本公开的实施例的详细描述并非旨在限制要求保护的本公开的范围，而是仅仅表示本公开的选定实施例。基于本公开中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本公开保护的范围。

应注意到：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步定义和解释。

在本公开的描述中，需要说明的是，若出现术语“上”、“下”、“内”、“外”等指示的方位或位置关系为基于附图所示的方位或位置关系，或者是该公开产品使用时惯常摆放的方位或位置关系，仅是为了便于描述本公开和简化描述，而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作，因此不能理解为对本公开的限制。

此外，若出现术语“第一”、“第二”等仅为区分描述，而不能理解为指示或暗示相对重要性。

需要说明的是，在不冲突的情况下，本公开的实施例中的特征可以相互结合。

现有 RAG 系统中，在生成回答前，首先从外部文档中检索相关信息，将检索到的知识作为上下文输入大模型，从而增强大模型的知识覆盖范围，提升其准确性和可靠性。请参照图 1，图 1 为本公开实施例提供的 RAG 系统的框架图，RAG 系统包括以下处理步骤：

（1）针对文档进行文本切分，得到切分后的文本段；

（2）将切分后的文本段进行向量化存储，生成向量数据库；

（3）在针对指定问题进行处理时，对向量数据库进行向量检索，从向量数据库中检索出与指定问题相关联的向量，将检索到的向量还原成对应的文本段（图中的语义段），作为指定问题的上下文，基于指定问题及其上下文构建 RAG 系统中大模型的提示词（Prompt）；

（4）根据提示词进行问答输出。

在RAG框架中，文档切分采用的切分策略通常是基于固定的切分粒度的方式，切分粒度通常与大模型的输入长度限制相关联，这种切分粒度固定的方式虽然能够满足大模型的输入要求，但不可避免地会出现语义割裂的问题，并且针对不同的大模型，其输入长度的限制也有差别，固定粒度的切分方式也无法灵活地适应不同场景的需求。

有鉴于此，本公开实施例提供文本切分方法、装置、电子设备及计算机可读存储介质，能够按照语义对文本进行合理切分，保证切分后的语义段的语义连贯性，下面将对其进行详细描述。

请参照图2，图2为本公开实施例提供的文本切分方法的流程示例图，该方法包括以下步骤：

步骤S101，获取针对目标文本进行句子级别切分后得到的多个句子。

在本公开实施例中，目标文本可以是从一个或多个文档中提取出的文本，文档可以是word文档、txt文档、pdf文档等，也可以是对图片进行文字识别后，从中提取出的文本。

在本公开实施例中，可以采用预先训练的模型对目标文本进行句子级别的切分，也可以根据目标文本中的标点符号及换行符号，对目标文本进行句子级别的划分。

步骤S102，以多个句子中的第一个句子为起点、最后一个句子为终点，多次滑动预设滑动窗口，并在每次滑动后，将处于预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段。

在本公开实施例中，预设滑动窗口的大小影响切分效率及切分精度，窗口大小代表每次需要进行语义切分的句子的数量，可以根据业务应用场景调节预设滑动窗口的大小，例如，预设滑动窗口大小为2。

在本公开实施例中，进行语义切分可以采用预先训练的语言模型进行语义切分，语言模型包括、但不限于BERT-NSP模型（一种使用了Transformer架构来处理自然语言处理的深度学习模型）、BGE 模型（一种文本嵌入模型，配置成将文本转换为低维稠密向量，以便进行高效的语义分析和检索）及SimCSE（Similarity Consensus Semantic Encoding）模型（一种配置成生成句子嵌入的模型）等。

步骤S103，对于每次切分得到的语义段，将本次切分得到的语义段和上次切分得到的语义段进行聚类，得到上次滑动后和本次滑动后预设滑动窗口内的每一目标句子所属的目标语义段，直至得到多个句子中每一句子所属的语义段。

在本公开实施例中，为了保证两次切分得到的语义段之间的语义连贯性，若上次切分的最后一个语义段和本次切分的第一个语义段之间是语义连贯的，需要将两个语义段合并为一个语义段，否则，将二者作为两个独立的语义段。例如，上次切分的最后一个语义段

1 包括句子 a、句子 b，本次切分的首个语义段 2 包括句子 b、句子 c 和句子 d，则语义段 1 和语义段 2 之间是语义连贯的，则语义段 1 和语义段 2 可以合并为一个语义段。

在本公开实施例中，作为一种可选的实现方式，可以滑动一次，进行一次聚类，也可以滑动多次后，再将每次滑动后语义切分得到的语义段进行逐次聚类，最终都能得到多个句子中每一句子所属的语义段。

本公开实施例提供的上述方法，依次对预设滑动窗口内的目标句子按照语义切分，保证预设滑动窗口内目标句子切分后的语义连贯性，并且将本次切分得到的语义段和上次切分得到的语义段进行聚类，保证相邻两次滑动后预设滑动窗口之间的句子的语义连贯性，实现了文本的合理切分。

在可选的实施方式中，为了将多个目标句子按照语义切分，可以通过确定两两相邻的目标句子之间的是否语义连续进行切分，一种可选的实现方式为：

首先，将多个目标句子输入预先训练的文本切分模型，得到两两相邻的目标句子组成的句子对的语义连续度；

在本公开实施例中，文本切分模型可以是基于 BERT-NSP、BGE 及 SimCSE（Similarity Consensus Semantic Encoding）的模型。以基于 BERT-NSP 的文本切分模型为例，可以在构造训练数据集过程中，根据文档内容对语句进行错位标注，即在处理文档时，识别并标注出那些与文档整体内容不一致或存在错误的句子，以基于错位标注对文档中的错误进行自动纠错、提高从文档中提取信息的准确性、对文档的整体质量和一致性进行准确评估等。

此外，在数据集中段落间的样本较为稀缺的情况下，还可以采用过采样策略，即通过增加少数类的样本数量来平衡各类别之间的样本数，常见的过采样策略包括但不限于随机过采样和在少数类的样本之间插值生成新的合成样本，以确保训练数据的样本均衡性，进而提高模型的泛化能力。

其次，对于任一目标句子对，若目标句子对的语义连续度大于预设值，则将目标句子对中的两个目标句子归属为同一个语义段，否则，将目标句子对中的两个目标句子归属为不同的语义段；

在本公开实施例中，作为一种可选的实现方式，通过文本切分模型得到的语义连续度可以是两个目标句子之间语义连续的概率值，若大于预设值，则认为这两个目标句子属于同一个语义段，否则，认为两个目标句子属于不同的语义段。作为另一种可选的实现方式，通过文本切分模型得到的语义连续度可以是由 0 和/或 1 组成的字符串，其中的 0 表示对应的两个目标句子之间语义不连续，1 表示对应的两个目标句子之间语义连续。例如，多个目标句子为句子 a、句子 b、句子 c 和句子 d，句子 a 和句子 b 组成句子对 1，句子 b 和句子 c 组成句子对 2，句子 c 和句子 d 组成句子对 3。文本切分模型输出的字符串为 011，则

其中的“0”表示句子对 1 的语义连续度，表征句子 a 和句子 b 之间语义不连续，第 1 个“1”表示句子对 2 的语义连续度，表征句子 b 和句子 c 之间语义连续，第 2 个“1”表示句子对 3 的语义连续度，即句子 c 和句子 d 之间语义连续。

最后，依次将每一句子对作为目标句子对，得到多个目标句子语义切分后的至少一个语义段。

在本公开实施例中，将多个目标句子中语义连续的句子划分为同一个语义段，将语义不连续的句子划分为不同的语义段，得到多个目标句子语义切分后的至少一个语义段。

在可选的实施方式中，为了保证相邻切分得到的语义段之间的语义连贯性，将上次滑动预设滑动窗口内的最后一个目标句子作为本次滑动后预设滑动窗口内的第一个目标句子，即本次滑动后预设滑动窗口内的第一个目标句子与上次滑动预设滑动窗口内的最后一个目标句子相同，基于此，一种将本次切分得到的语义段和上次切分得到的语义段进行聚类的实现方式为：

首先，获取本次滑动后预设滑动窗口内的首个句子对的语义连续度；

在本公开实施例中，首个句子对的语义连续度可以通过将首个句子对中的两个句子输入前文所述的文本切分模型后得到。

其次，计算首个句子对中的两个目标句子之间的语义相似度；

在本公开实施例中，一种可选的实现方式为：可以通过将两个目标句子输入 BGE、M3E（Multi-Modal Multi-Task Embedding，多模态多任务嵌入）等模型得到语义相似度，另一种实现方式可以为：通过分别提取两个目标句子的特征向量，再计算两者特征向量的余弦相似度，将计算出的余弦相似度作为语义相似度。

第三，根据首个句子对中的两个目标句子之间的语义连续度和语义相似度，计算首个句子对的语义关联得分；

在本公开实施例中，可以为语义连续度和语义相似度分别设置一个连续度权重和相似度权重，语义关联得分=语义连续度×连续度权重+语义相似度×相似度权重。语义连续度权重和语义相似度权重可以根据实际应用场景及各自对应的模型的可靠性或者与实际应用场景的匹配度进行设置，例如，连续度权重设置为 0.7，相似度权重设置为 0.3。

最后，若语义关联得分大于预设分值，则将上次切分得到的最后一个语义段和本次切分得到的首个语义段合并为一个语义段，否则，将上次切分得到的最后一个语义段和本次切分得到的首个语义段作为不同的语义段。

在本公开实施例中，预设分值可以根据实际需要设置，例如预设分值为 0.5。

需要说明的是，在语义切分效率允许的情况下，也可以利用语义关联得分在每次滑动后，将处于预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段，即将处

于预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段的另一种实现方式也可以是：

获取两两相邻的目标句子组成的句子对的语义连续度；

计算两两相邻的目标句子组成的句子对的语义相似度；

根据两两相邻的目标句子组成的句子对中的两个目标句子之间的语义连续度和语义相似度，计算两两相邻的目标句子组成的句子对的语义关联得分；

若语义关联得分大于预设分值，则将对应句子对中的目标句子归属为同一个语义段，否则，将其归属为不同的语义段。

进一步地，为了提高语义切分的效率，也可只对多个目标句子中语义连续度小于或者等于预设值的待确认语义对，计算待确认语义对的语义相似度，根据待确认语义对的语义连续度和语义相似度，计算待确认语义对的语义关联得分，若语义关联得分大于预设分值，则将待确认语义对中的目标句子归属为同一个语义段，否则，将其归属为不同的语义段。由于其只针对语义连续度小于或者等于预设值的待确认语义对，通过计算语义关联得分的方式确认其包括的目标句子是否归属为同一语义段，从一定程度上减少了计算量，与通过计算目标句子中每一句子对的语义关联得分的方式相比，从一定程度上提高了语义切分的效率。

为了更直观地展示滑动预设滑动窗口的过程中对多个句子进行语义切分的过程，请参照图 3，图 3 为本公开实施例提供的利用预设滑动窗口进行语义切分的过程示例图，针对预设滑动窗口内的句子 1、句子 2 和句子 3，进行语义切分，得到：语义段【句子 1，句子 2】和语义段【句子 3】，滑动预设滑动窗口，针对预设滑动窗口内的句子 3、句子 4 和句子 5，进行语义切分，得到：语义段【句子 3，句子 4，句子 5】，将本次得到的语义段【句子 3，句子 4，句子 5】和上次得到的语义段【句子 1，句子 2】和语义段【句子 3】进行聚类，得到语义段【句子 1，句子 2】和语义段【句子 3，句子 4，句子 5】。

为了展示本公开实施例提供的语义切分方式的效果，本公开实施例对人工标注的 19 段的句子进行三种切分方式的切分并对三种切分方式的切分结果进行评估和对比。方式一：固定字数切分；方式二：本公开实施例中利用预设滑动窗口和 BERT-NSP 模型切分；方式三：本公开实施例中利用预设滑动窗口和 BERT-NSP+M3E 模型切分。

方式一的切分结果：18 段，正确切分的段数为 2 个；P=11.11%、R=94.73%、F1=19.87%；

方式二的切分结果：16 段，正确切分的段数为 13 个，P=81.25%、R=84.21%、F1=82.70%；

方式三的切分结果：18 段，正确切分的段数为 16 个，P=88.88%、R=94.73%、F1=91.71%；

其中，P 表示精准率，计算方式为：正确切分出的段数目/切分段总数； R 表示召回率，计算方式为：正确切分出的段数目/应切分段总数；F1 表示切分结果得分，计算方式为：2×

精准率×召回率/(精准率+召回率)。

在本公开实施例中，为了能够使得切分后的语义段能够灵活地适配不同的大模型的输入需求，在得到上次滑动后和本次滑动后预设滑动窗口内的每一目标句子所属的目标语义段，还需要将目标语义段按照所需要的大模型的输入需求进行处理，本公开实施例提供一种处理方式：

首先，获取预设限制条件，预设限制条件配置成限制检索增强生成 RAG 系统中的大模型输入的语义段；

在本公开实施例中，预设限制条件可以是大模型输入的单个的语义段的最大长度，例如，每个语义段的最大长度不超过 500 字，也可以是大模型输入的相邻两个语义段之间的重叠度，例如，第一个语义段是从第 1 句到第 10 句，第二个语义段是从第 6 句到第 15 句。

需要说明的是，预设限制条件也可以将最长长度和重叠度相结合，即既限制大模型输入的单个语义段的长度，同时又限制相邻两个语义段之间的重叠度。

其次，根据预设限制条件，对每一目标语义段进行处理，得到每一目标语义段对应的参考语义段，并将每一参考语义段以向量化的形式存储至向量数据库，以在大模型接收到输入数据时，从向量数据库中检索出与输入数据相关的语义段向量，并基于输入数据及检索出的语义段向量进行反馈。

在本公开实施例中，参考语义段是符合预设限制条件的语义段。一个目标语义段可以对应一个参考语义段，一个目标语义段也可以对应多个参考语义段。当一个目标语义段对应多个参考语义段时，为了保证多个参考语义段之间的语义连贯性，多个参考语义段之间可以存在部分重叠。

在可选的实施方式中，对于输入长度受限的大模型而言，为了在满足大模型的长度要求的同时，又能尽量保持句子之间的语义连贯性，本公开实施例提供了一种对目标语义段进行处理，以得到满足大模型输入长度要求的参考语义段的实现方式：

首先，针对任一目标语义段，若目标语义段不满足预设限制条件，则根据预设限制条件对目标语义段再次进行切分，得到目标语义段对应的参考语义段；

其次，若目标语义段满足预设限制条件，则将目标语义段作为其对应的参考语义段。

在本公开实施例中，预设限制条件为预设长度，若目标语义段满足预设限制条件，意味着目标语义段中所有句子的长度之和小于或者等于预设长度，则可以不用对目标语义段进行处理，目标语义段就是参考语义段，否则，意味着目标语义段中所有句子的长度之和大于预设长度，则需要将目标语义段进行再次切分，使得切分后的语义段满足预设限制条件，一种再次切分的实现方式为：

（1）将目标语义段中的第一个句子作为参考句子；

（2）从参考句子开始，从目标语义段中确定目标句子序列，其中，目标句子序列中的所有句子的长度之和最大且小于预设长度；

（3）将目标句子序列中的所有句子归为一个参考语义段；

（4）若目标句子序列中的最后一个句子非目标语义段的最后一个句子，则从目标语义段中重新确定参考句子，并返回（2）的步骤，直至得到目标语义段对应的所有参考语义段。

在可选的实施方式中，为了在再次切分时尽可能地保证最终得到的参考语义段之间的语义连贯性，本公开实施例提供一种确定参考句子的实现方式：

首先，从目标句子序列中确定第一关联句子；

其次，从目标语义段中位于目标句子序列之后的句子中确定第二关联句子；

最后，若第一关联句子的长度和第二关联句子的长度之和大于预设长度，则根据第二关联句子确定参考句子，否则根据第一关联句子确定参考句子。

在本公开实施例中，根据第二关联句子确定参考句子的方式，可以是将第二关联句子的第一个句子作为参考句子，也可以是将第一关联句子中后若干个句子的第一个句子作为参考句子，只要保证第一关联句子的后若干个句子和第二关联句子的长度之和小于或者等于预设长度即可。

在本公开实施例中，在本公开实施例中，第一关联句子和第二关联句子的个数均可以是一个或者多个，根据切分时允许的最大重叠数而定。第一关联句子可以是位于目标句子序列中后预设数量的句子，作为一种实现方式，预设数量可以是固定的，根据切分时允许的最大重叠数而定，例如，预设数量设置为 2；作为另一种实现方式，预设数量也可以是可变的，只要能够满足第一关联句子的长度和第二关联句子的长度之和小于或者等于预设长度即可。

本公开实施例以第一关联句子的个数为 2、第二关联句子为 1 为例进行举例说明，例如，目标语义段包括的句子如下：

【句子 1，句子 2，句子 3，句子 4，句子 5，句子 6，句子 7】

将句子 1 作为参考句子，句子 1、句子 2、句子 3 长度之和最大且小于预设长度，则目标句子序列为：（句子 1、句子 2、句子 3），且句子 1、句子 2、句子 3 属于参考语义段 1；

将句子 2、句子 3 作为第一关联句子，将句子 4 作为第二关联句子，若句子 2、句子 3 和句子 4 的长度之和小于预设长度，则将句子 2 作为参考句子；

句子 2、句子 3、句子 4、句子 5 长度之和最大且小于预设长度，则目标句子序列为：（句子 2、句子 3、句子 4、句子 5），且句子 2、句子 3、句子 4、句子 5 属于参考语义段 2；

将句子 4、句子 5 作为第一关联句子，将句子 6 作为第二关联句子，句子 4、句子 5 和句子 6 的长度之和大于预设长度，则可以将句子 6 作为参考句子，若句子 5 和句子 6 的长

度之和小于预设长度，则也可以将句子 5 作为参考句子；本次示例以句子 5 作为参考句子；

句子 5、句子 6 和句子 7 的长度之和小于预设长度，则目标句子序列为（句子 5、句子 6、句子 7），且句子 5、句子 6、句子 7 属于参考语义段 3；

由此，目标语义段被再次切分成 3 个参考语义段：参考语义段 1~参考语义段 3。

在本公开实施例中，预设限制条件除了可以为预设长度，还可以为重叠数，以重叠数为 2，对于目标语义段【句子 1，句子 2，句子 3，句子 4，句子 5】，可以再次切分为（句子 1，句子 2，句子 3，句子 4）和（句子 3，句子 4，句子 5）两个参考语义段，两个参考语义段共同包含两个重叠的句子：句子 3 和句子 4，也可以再次切分为（句子 1，句子 2，句子 3）、（句子 2，句子 3，句子 4）和（句子 3，句子 4，句子 5），可以根据实际需要而定。

为了执行上述实施例及各个可能的实施方式中的相应步骤，下面分别给出一种文本切分装置 100 的实现方式。请参照图 4，图 4 为本公开实施例提供的文本切分装置的方框示意图，需要说明的是，本公开所提供的文本切分装置 100，其基本原理及产生的技术效果和对应的上述实施例相同，为简要描述，本公开实施例部分未提及指出。

文本切分装置 100 包括获取模块 110、切分模块 120、聚类模块 130。

获取模块 110，配置成获取针对目标文本进行句子级别切分后得到的多个句子。

切分模块 120，配置成以多个句子中的第一个句子为起点、最后一个句子为终点，多次滑动预设滑动窗口，并在每次滑动后，将处于预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段。

聚类模块 130，配置成对于每次切分得到的语义段，将本次切分得到的语义段和上次切分得到的语义段进行聚类，得到上次滑动后和本次滑动后预设滑动窗口内的每一目标句子所属的目标语义段，直至得到多个句子中每一句子所属的语义段。

在可选的实施方式中，切分模块 120 配置成：

将多个目标句子输入预先训练的文本切分模型，得到两两相邻的目标句子组成的句子对的语义连续度；

对于任一目标句子对，若目标句子对的语义连续度大于预设值，则将目标句子对中的两个目标句子归属为同一个语义段，否则，将目标句子对中的两个目标句子归属为不同的语义段；

依次将每一句子对作为目标句子对，得到多个目标句子语义切分后的至少一个语义段。

在可选的实施方式中，本次滑动后预设滑动窗口内的第一个目标句子与上次滑动预设滑动窗口内的最后一个目标句子相同，预设滑动窗口内两两相邻的目标句子组成句子对；

对于每次切分得到的语义段，聚类模块 130 配置成：

获取本次滑动后预设滑动窗口内的首个句子对的语义连续度；

计算首个句子对中的两个目标句子之间的语义相似度；

根据首个句子对中的两个目标句子之间的语义连续度和语义相似度，计算首个句子对的语义关联得分；

若语义关联得分大于预设分值，则将上次切分得到的最后一个语义段和本次切分得到的首个语义段合并为一个语义段，否则，将上次切分得到的最后一个语义段和本次切分得到的首个语义段作为不同的语义段。

在可选的实施方式中，聚类模块 130 还配置成：

获取预设限制条件，预设限制条件配置成限制检索增强生成 RAG 系统中的大模型输入的语义段；

根据预设限制条件，对每一目标语义段进行处理，得到每一目标语义段对应的参考语义段，并将每一参考语义段以向量化的形式存储至向量数据库，以在大模型接收到输入数据时，从向量数据库中检索出与输入数据相关的语义段向量，并基于输入数据及检索出的语义段向量进行反馈。

在可选的实施方式中，预设限制条件为预设长度，目标语义段中所有句子的长度之和大于预设长度；聚类模块 130 在配置成根据预设限制条件，对每一目标语义段进行处理，得到每一目标语义段对应的参考语义段时配置成：

针对任一目标语义段，若目标语义段不满足预设限制条件，则根据预设限制条件对目标语义段再次进行切分，得到目标语义段对应的参考语义段；

若目标语义段满足预设限制条件，则将目标语义段作为其对应的参考语义段。

在可选的实施方式中，聚类模块 130 在配置成根据预设限制条件对目标语义段再次进行切分，得到目标语义段对应的参考语义段时配置成：

将目标语义段中的第一个句子作为参考句子；

从参考句子开始，从目标语义段中确定目标句子序列，其中，目标句子序列中的所有句子的长度之和最大且小于预设长度；

将目标句子序列中的所有句子归为一个参考语义段；

若目标句子序列中的最后一个句子非目标语义段的最后一个句子，则从目标语义段中重新确定参考句子，并返回从参考句子开始，从目标语义段中确定目标句子序列的步骤，直至得到目标语义段对应的所有参考语义段。

在可选的实施方式中，聚类模块 130 在配置成从目标语义段中重新确定参考句子时配置成：

从目标句子序列中确定第一关联句子；

从目标语义段中位于目标句子序列之后的句子中确定第二关联句子；

若第一关联句子的长度和第二关联句子的长度之和大于预设长度，则根据第二关联句子确定参考句子，否则根据第一关联句子确定参考句子。

本公开实施例还提供了电子设备 10 的方框示意图，电子设备 10 实现前述实施例的文本切分方法，请参照图 5，图 5 为本公开实施例提供的电子设备 10 的方框示意图，电子设备 10 包括处理器 11、存储器 12 和总线 13，处理器 11 和存储器 12 通过总线 13 连接。

处理器 11 可以是一种集成电路芯片，具有信号的处理能力。在实现过程中，上述实施例的文本切分方法的各步骤可以通过处理器 11 中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器 11 可以是通用处理器，包括 CPU（Central Processing Unit, 中央处理器）、NP（Network Processor, 网络处理器）等；还可以是 DSP（Digital Signal Processor, 数字信号处理器）、ASIC（Application Specific Integrated Circuit, 专用集成电路）、FPGA（Field Programmable Logic Gate Array, 现场可编程逻辑门阵列）或者其它可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。

存储器 12 配置成存储实现文本切分方法的程序，该程序可以是以软件或固件（firmware）的形式存储于存储器 12 中或固化在电子设备 10 的 OS（Operating System,操作系统）中的软件功能模块。

处理器 11 在接收到执行指令后，执行程序以实现前述实施例的文本切分方法。

本公开实施例提供一种计算机可读存储介质，其上存储有计算机程序，该计算机程序被处理器执行时实现如前述所述的文本切分方法。

综上所述，本公开实施例提供了一种文本切分方法、装置、电子设备及计算机可读存储介质，方法包括：获取针对目标文本进行句子级别切分后得到的多个句子；以多个句子中的第一个句子为起点、最后一个句子为终点，多次滑动预设滑动窗口，并在每次滑动后，将处于预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段；对于每次切分得到的语义段，将本次切分得到的语义段和上次切分得到的语义段进行聚类，得到上次滑动后和本次滑动后预设滑动窗口内的每一目标句子所属的目标语义段，直至得到多个句子中每一句子所属的语义段。与现有技术相比，本公开实施例至少具有以下优势：（1）针对多个句子，滑动预设滑动窗口，依次对预设滑动窗口内的目标句子按照语义切分，保证预设滑动窗口内目标句子切分后的语义连贯性，并且将本次切分得到的语义段和上次切分得到的语义段进行聚类，保证相邻两次滑动后预设滑动窗口之间的句子的语义连贯性，最终实现了文本的合理切分，保证了切分后的语义段的语义连贯性，进而提高了 RAG 反馈的问答输出的准确性和可靠性；（2）利用相邻两次滑动后的预设滑动窗口的句子之间的语义连续度和语义相似度，准确地判断两者之间的语义连贯性，提高语义切分的合理性；（3）

由于保证了语义切分后的语义段内的句子间的语义连贯性，减少了后续对 RAG 的大模型的硬件资源要求，更具备通用性、经济性，达到降本增效的目的；（4）在切分的同时能够兼顾配置成限制 RAG 系统中的大模型输入的语义段的预设限制条件，根据预设限制条件对切分得到的语义段进行处理，使得最终的切分结果在满足 RAG 系统中的大模型的需求的同时又能最大程度地保证语义连贯性，达到能够灵活适应不同大模型要求的效果。

以上所述，仅为本公开的各种实施方式，但本公开的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本公开揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本公开的保护范围之内。因此，本公开的保护范围应所述以权利要求的保护范围为准。

工业实用性

本公开实施例所提供的文本切分方法、装置、电子设备及计算机可读存储介质，包括：获取针对目标文本进行句子级别切分后得到的多个句子；以多个句子中的第一个句子为起点、最后一个句子为终点，多次滑动预设滑动窗口，并在每次滑动后，将处于预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段；对于每次切分得到的语义段，将本次切分得到的语义段和上次切分得到的语义段进行聚类，得到上次滑动后和本次滑动后预设滑动窗口内的每一目标句子所属的目标语义段，直至得到多个句子中每一句子所属的语义段。本公开实施例具有以下优势：（1）针对多个句子，滑动预设滑动窗口，依次对预设滑动窗口内的目标句子按照语义切分，保证预设滑动窗口内目标句子切分后的语义连贯性，并且将本次切分得到的语义段和上次切分得到的语义段进行聚类，保证相邻两次滑动后预设滑动窗口之间的句子的语义连贯性，最终实现了文本的合理切分，保证了切分后的语义段的语义连贯性，进而提高了 RAG 反馈的问答输出的准确性和可靠性；（2）利用相邻两次滑动后的预设滑动窗口的句子之间的语义连续度和语义相似度，准确地判断两者之间的语义连贯性，提高语义切分的合理性；（3）由于保证了语义切分后的语义段内的句子间的语义连贯性，减少了后续对 RAG 的大模型的硬件资源要求，更具备通用性、经济性，达到降本增效的目的；（4）在切分的同时能够兼顾配置成限制 RAG 系统中的大模型输入的语义段的预设限制条件，根据预设限制条件对切分得到的语义段进行处理，使得最终的切分结果在满足 RAG 系统中的大模型的需求的同时又能最大程度地保证语义连贯性，达到能够灵活适应不同大模型要求的效果。

权利要求书

1.一种文本切分方法，其特征在于，所述方法包括：

获取针对目标文本进行句子级别切分后得到的多个句子；

以所述多个句子中的第一个句子为起点、最后一个句子为终点，多次滑动预设滑动窗口，并在每次滑动后，将处于所述预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段；

对于每次切分得到的语义段，将本次切分得到的语义段和上次切分得到的语义段进行聚类，得到上次滑动后和本次滑动后所述预设滑动窗口内的每一目标句子所属的目标语义段，直至得到所述多个句子中每一句子所属的语义段。

2.根据权利要求 1 所述的文本切分方法，其特征在于，所述将处于所述预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段的步骤包括：

将所述多个目标句子输入预先训练的文本切分模型，得到两两相邻的目标句子组成的句子对的语义连续度；

对于任一目标句子对，若所述目标句子对的语义连续度大于预设值，则将所述目标句子对中的两个目标句子归属为同一个语义段，否则，将所述目标句子对中的两个目标句子归属为不同的语义段；

依次将每一句子对作为所述目标句子对，得到所述多个目标句子语义切分后的至少一个语义段。

3.根据权利要求 1 所述的文本切分方法，其特征在于，本次滑动后预设滑动窗口内的第一个目标句子与上次滑动预设滑动窗口内的最后一个目标句子相同，所述预设滑动窗口内两两相邻的目标句子组成句子对；

所述对于每次切分得到的语义段，将本次切分得到的语义段和上次切分得到的语义段进行聚类，得到上次滑动后和本次滑动后所述预设滑动窗口内的每一目标句子所属的目标语义段的步骤包括：

获取本次滑动后所述预设滑动窗口内的首个句子对的语义连续度；

计算所述首个句子对中的两个目标句子之间的语义相似度；

根据所述首个句子对中的两个目标句子之间的语义连续度和语义相似度，计算所述首个句子对的语义关联得分；

若所述语义关联得分大于预设分值，则将上次切分得到的最后一个语义段和本次切分得到的首个语义段合并为一个语义段，否则，将上次切分得到的最后一个语义段和本次切分得到的首个语义段作为不同的语义段。

4.根据权利要求 3 所述的文本切分方法，其特征在于，所述计算所述首个句子对中的两

个目标句子之间的语义相似度，包括：

分别提取两个目标句子的特征向量；

计算两个目标句子的特征向量的余弦相似度，并将计算出的余弦相似度作为语义相似度。

5.根据权利要求 3 所述的文本切分方法，其特征在于，所述计算所述首个句子对中的两个目标句子之间的语义相似度，包括：

将两个目标句子输入多模态多任务嵌入模型，以得到两个目标句子之间的语义相似度。

6.根据权利要求 3 所述的文本切分方法，其特征在于，所述首个句子对的语义关联得分的算式为：

语义关联得分=语义连续度×连续度权重+语义相似度×相似度权重。

7.根据权利要求 1 所述的文本切分方法，其特征在于，所述将处于所述预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段，包括：

获取两两相邻的目标句子组成的句子对的语义连续度；

计算两两相邻的目标句子组成的句子对的语义相似度；

根据两两相邻的目标句子组成的句子对中的两个目标句子之间的语义连续度和语义相似度，计算两两相邻的目标句子组成的句子对的语义关联得分；

若语义关联得分大于预设分值，则将对应句子对中的目标句子归属为同一个语义段，否则，将其归属为不同的语义段。

8.根据权利要求 1-7 中任一项所述的文本切分方法，其特征在于，所述对于每次切分得到的语义段，将本次切分得到的语义段和上次切分得到的语义段进行聚类，得到上次滑动后和本次滑动后所述预设滑动窗口内的每一目标句子所属的目标语义段的步骤之后，所述方法还包括：

获取预设限制条件，所述预设限制条件配置成限制检索增强生成 RAG 系统中的大模型输入的语义段；

根据所述预设限制条件，对每一所述目标语义段进行处理，得到每一所述目标语义段对应的参考语义段，并将每一所述参考语义段以向量化的形式存储至向量数据库，以在所述大模型接收到输入数据时，从所述向量数据库中检索出与所述输入数据相关的语义段向量，并基于所述输入数据及检索出的语义段向量进行反馈。

9.根据权利要求 8 所述的文本切分方法，其特征在于，所述预设限制条件包括大模型输入的单个的语义段的最大长度和/或大模型输入的相邻两个语义段之间的重叠度。

10.根据权利要求 8 所述的文本切分方法，其特征在于，所述预设限制条件为预设长度，所述目标语义段中所有句子的长度之和大于所述预设长度；所述根据所述预设限制条件，

对每一所述目标语义段进行处理，得到每一所述目标语义段对应的参考语义段的步骤包括：

针对任一所述目标语义段，若所述目标语义段不满足所述预设限制条件，则根据所述预设限制条件对所述目标语义段再次进行切分，得到所述目标语义段对应的参考语义段；

若所述目标语义段满足所述预设限制条件，则将所述目标语义段作为其对应的参考语义段。

11.根据权利要求 10 所述的文本切分方法，其特征在于，所述根据所述预设限制条件对所述目标语义段再次进行切分，得到所述目标语义段对应的参考语义段的步骤包括：

将所述目标语义段中的第一个句子作为参考句子；

从所述参考句子开始，从所述目标语义段中确定目标句子序列，其中，所述目标句子序列中的所有句子的长度之和最大且小于所述预设长度；

将所述目标句子序列中的所有句子归为一个参考语义段；

若所述目标句子序列中的最后一个句子非所述目标语义段的最后一个句子，则从所述目标语义段中重新确定参考句子，并返回所述从所述参考句子开始，从所述目标语义段中确定目标句子序列的步骤，直至得到所述目标语义段对应的所有参考语义段。

12.根据权利要求 11 所述的文本切分方法，其特征在于，所述从所述目标语义段中重新确定参考句子的步骤包括：

从所述目标句子序列中确定第一关联句子；

从所述目标语义段中位于所述目标句子序列之后的句子中确定第二关联句子；

若所述第一关联句子的长度和所述第二关联句子的长度之和大于所述预设长度，则根据所述第二关联句子确定所述参考句子，否则根据所述第一关联句子确定所述参考句子。

13.一种文本切分装置，其特征在于，所述装置包括：

获取模块，配置成获取针对目标文本进行句子级别切分后得到的多个句子；

切分模块，配置成以所述多个句子中的第一个句子为起点、最后一个句子为终点，多次滑动预设滑动窗口，并在每次滑动后，将处于所述预设滑动窗口内的多个目标句子按照语义切分，得到至少一个语义段；

聚类模块，配置成对于每次切分得到的语义段，将本次切分得到的语义段和上次切分得到的语义段进行聚类，得到上次滑动后和本次滑动后所述预设滑动窗口内的每一目标句子所属的目标语义段，直至得到所述多个句子中每一句子所属的语义段。

14.一种电子设备，其特征在于，包括处理器和存储器，所述存储器配置成存储程序，所述处理器配置成在执行所述程序时，实现如权利要求 1-12 中任一项所述的文本切分方法。

15.一种计算机可读存储介质，其特征在于，其上存储有计算机程序，该计算机程序被处理器执行时实现如权利要求 1-12 中任一项所述文本切分方法。