

文本合并判断模型的训练方法和文本合并判断方法

技术领域

[01]本说明书涉及自然语言处理技术领域，尤其涉及一种文本合并判断模型的训练方法和文本合并判断方法。

背景技术

[02]通常情况下，将一个长文本分割成多个句子，可以通过“。”、“！”、“？”乃至“，”来分割。但由于文本的生成环境非常复杂，录入的文本可能存在错误使用分割情况。例如，用户通过移动终端的触摸屏输入文本，但错误地使用了分割符号、大量使用空格和错误使用分行的情况，又例如，用户通过语音输入文本，但语音录入环境的条件恶劣或用户录入时不正常地停顿，都会导致语音输入的文本存在分割错误的情况。因此，判断两个句子也即两个短文本能否进行合并一直是人工智能自然语言处理领域的基础任务之一，是文本查重、智能问答等上层应用的基础支撑技术。

发明内容

[03]本说明书实施例提供了一种文本合并判断方法、装置、存储介质及电子设备，可以训练文本合并判断模型，提高文本合并模型的鲁棒性，以及提高通过文本合并判断模型判断两个文本是否可以合并的准确性。所述技术方案如下：第一方面，本说明书实施例提供了一种文本合并判断模型的训练方法，所述方法包括：获取至少一个正样本组，以及获取至少一个负样本组，所述正样本组包括两个不可以合并的文本，所述负样本组包括两个可以合并的文本；通过所述至少一个正样本组和所述至少一个负样本组训练所述文本合并判断模型，直至所述文本合并判断模型收敛。

[04]第二方面，本说明书实施例提供了一种文本合并判断的方法，所述方法包括：获取两个待检测文本；将所述两个待检测文本输入至文本合并判断模型中，得到所述两个待检测文本是否可以合并的判断结果；其中，所述文本合并判断模型为采用第一方面所述的文本合并判断模型的训练方法训练得到的模型。

[05]第三方面，本说明书实施例提供了一种文本合并判断模型的训练装置，所述方法包括：样本获取模块，用于获取至少一个正样本组，以及获取至少一个负样本组，所述正样本组包括两个不可以合并的文本，所述负样本组包括两个可以合并的文本；模型训练

模块，用于通过所述至少一个正样本组和所述至少一个负样本组训练所述文本合并判断模型，直至所述文本合并判断模型收敛。

[06]第四方面，本说明书实施例提供一种文本合并判断的装置，所述装置包括：文本获取模块，用于获取两个待检测文本；结果获取模块，用于将所述两个待检测文本输入至文本合并判断模型中，得到所述两个待检测文本是否可以合并的判断结果；其中，所述文本合并判断模型为采用第一方面所述的文本合并判断模型的训练方法训练得到的模型。

[07]第五方面，本说明书实施例提供一种计算机存储介质，所述计算机存储介质存储有多条指令，所述指令适于由处理器加载并执行上述的方法步骤。

[08]第六方面，本说明书实施例提供一种计算机程序产品，所述计算机程序产品存储有多条指令，所述指令适于由处理器加载并执行上述的方法步骤。

[09]第七方面，本说明书实施例提供一种电子设备，可包括：处理器和存储器；其中，所述存储器存储有计算机程序，所述计算机程序适于由所述处理器加载并执行上述的方法步骤。

[10]本说明书一些实施例提供的技术方案带来的有益效果至少包括：本说明书实施例合理构建至少一个正样本组和负样本组，正样本组包括不可以合并的文本，负样本组包括可以合并的文本，通过至少一个正负样本组使文本合并判断模型可以自监督式地学习两个文本中是否存在可合并的关系，直至文本合并判断模型收敛，从而提高文本合并判断模型的训练效率，以及通过至少一个正负样本对使文本合并判断模型进行多轮训练，以使训练完成的文本合并判断模型具有较好的抗干扰性和鲁棒性，执行判断两个文本是否合并的任务的准确性较高，从而得到具有完整的语义的合并文本，便于用户阅读理解。

附图说明

[11]为了更清楚地说明本说明书实施例或相关技术中的技术方案，下面将对实施例或相关技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本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。

[12]图 1 是本说明书实施例提供了一种文本合并判断模型判断文本是否可以合并的流程示意图；

- [13]图 2 是本说明书实施例提供的一种文本合并判断模型的训练方法；
- [14]图 3 是本说明书实施例提供的一种获取负样本组的流程示意图；
- [15]图 4 是本说明书实施例提供的一种获取负样本组的流程示意图；
- [16]图 5 是本说明书实施例提供的一种文本合并判断模型的结构示意图；
- [17]图 6 是本说明书实施例提供的一种文本合并判断模型判断文本是否可以合并的流程示意图；
- [18]图 7 是本说明书实施例提供的一种文本合并判断方法的场景示意图；
- [19]图 8 是本说明书实施例提供的一种文本合并判断方法的流程示意图；
- [20]图 9 是本说明书实施例提供的一种文本合并判断模型的训练装置的结构示意图；
- [21]图 10 是本说明书实施例提供的一种文本合并判断装置的结构示意图；
- [22]图 11 是本说明书实施例提供的一种电子设备的结构示意图。

具体实施方式

[23]下面将结合本说明书实施例中的附图，对本说明书实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本说明书一部分实施例，而不是全部的实施例。基于本说明书中的实施例，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，都属于本说明书保护的范围。

[24]在本说明书的描述中，需要理解的是，术语“第一”、“第二”等仅用于描述目的，而不能理解为指示或暗示相对重要性。在本说明书的描述中，需要说明的是，除非另有明确的规定和限定，“包括”和“具有”以及它们任何变形，意图在于覆盖不排他的包含。例如包含了一系列步骤或单元的过程、方法、系统、产品或设备没有限定于已列出的步骤或单元，而是可选地还包括没有列出的步骤或单元，或可选地还包括对于这些过程、方法、产品或设备固有的其他步骤或单元。对于本领域的普通技术人员而言，可以具体情况理解上述术语在本说明书中的具体含义。此外，在本说明书的描述中，除非另有说明，“多个”是指两个或两个以上。“和/或”，描述关联对象的关联关系，表示可以存在三种关系，例如，A 和/或 B，可以表示：单独存在 A，同时存在 A 和 B，单独存在 B 这三种情况。字符“/”一般表示前后关联对象是一种“或”的关系。

[25]下面结合具体的实施例对本说明书进行详细说明。

[26]自然语言处理（Nature Language processing, NLP）是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此，这一领域的研究将涉及自然语言，即人们日常使用的语言，所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。

[27]随着网络技术的不断发展，人工智能技术已应用到各个领域，比如判断两个文本是否可以合并的技术。通常情况下，将一个长文本分割成多个句子，可以通过“。”、“！”、“？”乃至“，”来分割。但由于文本的生成环境非常复杂，录入的文本可能存在错误使用分割情况。例如，用户通过移动终端的触摸屏幕输入文本，但错误地使用了分割符号、大量使用空格和错误使用分行的情况，又例如，用户通过语音输入文本，但语音录入环境的条件恶劣或用户录入时不正常地停顿，都会导致语音输入的文本存在分割错误的情况。

[28]例如，用户输入的文本为“针对这个问题。我有别的看法，希望大家听一下”，文本1“针对这个问题”和文本2“我有别的看法，希望大家听一下”之间错误地使用了句号作为分割，实际上文本1和文本2应该合并，合并后的正确文本为“针对这个问题，我有别的看法，希望大家听一下”。

[29]因此，用于判断两个待检测文本是否可以合并的文本合并判断模型应运而生。如图1所示，图1为本说明书实施例提供的一种文本合并判断模型判断文本是否可以合并的流程示意图，将文本1011和文本1012输入至文本合并判断模型102中，以使文本合并判断模型102判断文本1011和文本1012是否可以合并，并输出判断结果103，判断结果103至少包括两个结果，其一是“可以合并”，其二是“不可以合并”。

[30]例如，输入的文本1011为“针对这个问题”，文本1012为“我有别的看法，希望大家听一下”，文本合并判断模型102判断文本1011和文本1012是否可以合并，输出判断结果103为“可以合并”，并相应执行后续的文本处理任务。

[31]相关技术中的文本合并判断模型主要分为两类，一类是基于人工智能中的机器学习方法构建的模型，另一类是基于人工智能中的深度学习的方法构建的模型。具体而言，基于机器学习方法构建的模型的判断过程为：将文本合并判断问题拆分成了特征工程和分类器两部分；特征工程包括两个文本预处理、特征提取、文本表示等部分；首先分别对两个文本进行清洗，利用分词工具分别对两个文本分词，再利用词袋法、TF-IDF等

方法将每个文本表示成向量形式再将其分别输入到分类器如 SVM、决策树等以得到最终结果。基于深度学习方法构建的模型的判断过程为：利用神经网络获取两个文本分别对应的有效特征，如卷积神经网络和循环神经网络；首先分别对两个文本进行清洗与分词，然后通过 word2vec 等基于神经网络思想的方法分别将两个文本转化为稠密的分布式词向量，再通过神经网络如 CNN 或 LSTM 对上述词向量对应的数据进行训练以得到最终结果。

[32]在一个实施例中，如图 2 所示，图 2 为本说明书实施例提出的一种文本合并判断模型的训练方法。该方法可依赖于计算机程序实现，可运行于基于冯诺依曼体系的文本合并判断训练装置上。该计算机程序可集成在应用中，也可作为独立的工具类应用运行。

[33]具体而言，文本合并判断模型的训练方法包括步骤 S102 和步骤 S104。

[34]S102、获取至少一个正样本组，以及获取至少一个负样本组。

[35]每个正样本组包括两个不可以合并的文本，两个文本具有单独且完整的语义。例如，正样本组中两个文本来自语文课本、报纸、新闻网站等发布的文本段落中，且两个文本中通过“。”、“！”或“？”进行连接，则该正样本组中的两个文本为正确分割的两个文本，不可以合并。当向文本合并判断模型输入正样本组时，训练至收敛的文本合并判断模型的判断结果应该是“不可以合并”。

[36]每个负样本组包括两个可以合并的文本，也即两个文本在语义上具有关联关系，且只有当两个文本合并后才具有完整的语义。例如，一个长文本中包括符号“，”、“、”、“：”以及“——”，在通过上述符号中任意一个符号处对该文本分割，得到两个文本，每个文本不能单独表达完整的意思。当向文本合并判断模型输入负样本组时，训练至收敛的文本合并判断模型的判断结果应该是“可以合并”。换言之，在一个实施例中，获取两个可以合并的文本组成负样本组的方法为：获取至少一个待分割的样本文本，根据至少一个待分割的样本文本中的预设符号，分别将待分割的样本文本进行分割，得到至少一个负样本组。预设符号可以是“，”、“、”、“：”以及“——”中任意一个，或其他本领域相关技术人员按需设置的符号。

[37]在一个实施例中，如图 3 所示，为本说明书实施例提供的一种获取负样本组的流程示意图，包括下述步骤 S1022 至步骤 S1028。

[38]S1022、获取待分割的样本文本。

[39]样本文本包括多个字符，获取待分割的样本文本的方法可以是任意一种已知且可以

实现的获取方法，样本文本的具体内容可以是任意一种。以医疗场景进行举例：样本文本包含有患者病例，样本文本为针对病例的至少一组问答对，问答对中包含患者针对该病例提出的问题和医生针对患者的问题给出的答案，或样本文本为医生针对该病例列出的诊断结果以及治疗方案，如“患者表现出严重的贫血症状，应注意饮食，以及注意用餐时间”。

[40]S1024、分别确定位于每个待分割的样本文本的中间位置的字符为目标字符。

[41]每个待分割的样本文本包括多个字符，字符分为符号字符和非符号字符。根据每个待分割的样本文本包括的字符的数量，以阅读顺序为判断顺序，将位于中间位置的字符为目标字符。

[42]例如，待分割的样本文本为“患者表现出严重的贫血症状，应注意饮食，以及注意用餐时间”，该待分割的文本包括 26 个字符，26 个字符中包括 2 个类型为符号的字符，因此确定位于第 14 个字符的目标字符“应”为目标字符。

[43]S1026、检测位于目标字符左边的 N 个字符中是否存在预设符号，以及检测位于目标字符右边的 N 个字符中是否存在预设符号。

[44]N 为大于 1 的正整数，由相关技术人员按需设置。例如，N 为 3 或 4 或 5。以 N 为窗口，检测在目标字符的左窗口和右窗口内是否存在预设符号。预设符号可以是“，”、“、”、“：”以及“——”中任意一个，或其他本领域相关技术人员按需设置的符号。

[45]检测位于目标字符左边的 N 个字符中是否存在预设符号，以及检测位于目标字符右边的 N 个字符中是否存在预设符号的顺序可以是按照阅读顺序，例如，阅读顺序为从左到右时，先检测目标字符左边的 N 个字符中是否存在预设符号，若为是，执行 S1028，若为否，也即目标字符左边的 N 个字符中不存在预设符号，则继续检测目标字符右边的 N 个字符中是否存在预设符号，若为是，则执行 S1028，若为否，也即位于目标字符左边的 N 个字符中不存在预设符号，且位于目标字符右边的 N 个字符中不存在预设符号，则执行 S1022、获取待分割的样本文本。

[46]S1028、以预设符号为界对每个待分割的样本文本进行分割，得到至少一个负样本组。

[47]若检测位于目标字符左边的 N 个字符中存在预设符号，或检测位于目标字符右边的 N 个字符中是否存在预设符号时，以预设符号为界，将待分割的样本文本进行分割，得到两个文本，将上述两个文本组成一个负样本组。

[48]在另一个实施例中，当检测到位于目标字符左边的 N 个字符中存在预设符号，且检

测位于目标字符右边的 N 个字符中存在预设符号时，以预设符号为界，将待分割的样本文本进行分割，得到两个负样本组。

[49]举例来说，如图 4 所示，是本说明书实施例提供的一种获取负样本组的流程示意图。获取待分割的样本文本 200，以位于待分割的样本文本 200 的中间位置的目标字符为界，将待分割的样本文本 200 切分为样本文本 201 和样本文本 202。进一步的，判断目标字符的左窗口 2011 和右窗口 2021 中是否存在预设符号。例如，在图 4 中，目标字符的左窗口 2011 中存在预设符号，根据预设符号，将待分割的样本文本 200 切分为样本文本 203 和样本文本 204，将样本文本 203 和样本文本 204 作为一个负样本组输入至文本合并判断模型中。

[50]本实施例提供了一种更合理且零成本的样本构造方法，不仅可以降低构造正样本组和负样本组的人工标注成本，且避免通过简单地按照预设符号来进行文本分割时，如果样本文本存在符号乱用的情况，会导致对待分割的样本文本少切、多切和误切的问题。

[51]在另一个实施例中，根据每个待分割的样本文本中每个预设字符对应的位置，以每个预设符号为界对待分割的样本文本进行分割，得到每个预设符号对应的负样本组。例如，待分割的样本文本为“患者表现出严重的贫血症状，应注意饮食，以及注意用餐时间”，该待分割的文本包括 26 个字符，26 个字符中包括 2 个类型为符号的字符，将待分割的样本文本分割为两组负样本组，第一个负样本组包括“患者表现出严重的贫血症状”和“应注意饮食”，另一个负样本组包括“应注意饮食”和“以及注意用餐时间”。本实施例提供的文本切割的方法逻辑简单，创造负样本组的效率较高。

[52]S104、通过至少一个正样本组和至少一个负样本组训练文本合并判断模型，直至文本合并判断模型收敛。

[53]获取至少一个正样本组和至少一个负样本组，将每个正样本组和负样本组输入至训练过程中的文本合并判断模型中，通过理想结果调整文本合并判断模型，直至文本合并判断模型收敛。在本说明书中，直至文本合并判断模型收敛的条件可以是预先设置的训练轮次，或者是根据训练过程中的停止条件确定的，停止条件可以是文本合并判断模型的损失函数收敛至期望值，或损失函数达到稳定在某一值后出现差异。

[54]训练过程可以包括迁移学习、多任务学习和对抗训练，对至少一个正样本组和负样本组包括数据增强处理。其中，迁移学习是利用在相似任务上训练的模型作为模型初始点在原本任务上进行再训练的方法，通过共享模型学到的知识，迁移学习可以加快模型

的学习效率并提高模型的泛化性。多任务学习是利用在相似任务上训练的模型作为模型初始点在原本任务上进行再训练的方法，通过共享模型学到的知识，迁移学习可以加快模型的学习效率并提高模型的泛化性。数据增强包含一系列用来生成新训练样本的技术，这些技术是通过原始数据采用随机抖动和扰动而类标签未变化来实现。应用数据增强的目标是增加模型的泛化性。对抗训练是一种增强模型鲁棒性的重要表示。在对抗训练的过程中，至少一个正样本组和至少一个负样本组会增加一些微小的扰动，使文本合并判断模型犯错，从而文本合并判断模型在训练的过程中能够适应扰动，以增强文本合并判断模型的鲁棒性。

[55]在一个实施例中，如图 5 所示，图 5 为本说明书实施例提供的一种文本合并判断模型的结构示意图，文本合并判断模型 40 包括：多个编码器、至少一个全连接层 402 和判断器 403，其中，多个编码器包括编码器 4011、编码器 4012、编码器 4013、...、编码器 401M，M 为大于或等于 2 的正整数。

[56]多个编码器，用于对输入的待检测文本进行编码，以得到每个待检测文本对应的多个特征向量。多个编码器为下述的一个或多个：双向编码器表示 BERT 模型的编码器、循环神经网络的编码器、卷积神经网络的编码器。其中，变压器的双向编码器表示 (Bidirectional Encoder Representation from Transformers, BERT) 是一个基于 Transformer 的在大规模语料库上进行掩码语言模型 (Mask Language Model, MLM) 和下一句预测 (Next Sentence Prediction, NSP) 多任务训练得到的预训练语言模型，循环神经网络 (Recurrent Neural Network, RNN) 为一类以序列 (sequence) 数据为输入，在序列的演进方向进行递归 (recursion) 且所有节点 (循环单元) 按链式连接的递归神经网络。可以理解的是，本说明书实施例还包括其他类型的编码器，对此不作限制。

[57]全连接层 402，用于对两个文本分别对应的多个特征向量进行全连接处理，得到至少一个连接结果。在一个实施例中，全连接层 402 的数量为一个或多个，至少一个全连接层 402 包括下述一个或多个全连接层：将所有特征向量依次连接的全连接层、将每个文本的头部字符对应的特征向量连接的全连接层、将一个文本的头部字符对应的特征向量与另一个文本的尾部字符对应的特征向量连接的全连接层。

[58]判断器 403，用于根据至少一个连接结果，判断至少两个文本是否可以合并。具体而言，判断器 403 对至少一个连接结果进行约束处理，得到至少两个文本可以合并的概率；根据至少两个文本可以合并的概率，判断至少两个文本是否可以合并。例如，将两个待检测文本输入至文本合并判断模型 40 中，通过多个编码器得到每个文本对应的多

个特征向量，通过至少一个全连接层 402 对每个文本对应的多个特征向量进行连接，得到至少一个连接结果，最后通过判断器 403 对至少一个连接结果进行约束处理，得到判断结果，判断该两个待检测文本是否可以合并。

[59]具体而言，如图 6 所示，图 6 是本说明书实施例提供的一种文本合并判断模型判断文本是否可以合并的流程示意图。

[60]首先，获取两个待检测文本，分别是待检测文本 501 和待检测文本 502。进一步的，根据分词规则将待检测文本 501 和待检测文本 502 进行最低粒度的分割，得到待检测文本 501 对应的多个分词 token，以及待检测文本 502 对应的多个分词 token，在待检测文本 501 对应的多个分词 token 的开头设置一个[CLS]分类，并将待检测文本 501 对应的多个分词 token 和待检测文本 502 对应的多个分词 token 通过[SEP]连接，并在待检测文本 502 对应的多个分词 token 后设置[SEP]作为结尾。

[61]进一步的，通过文本分类模型的编码层 401 的多个编码器分别对待检测文本 501 对应的多个分词 token 和待检测文本 502 对应的多个分词 token 进行编码，得到每个分词 token 对应的向量 embedding。例如，编码层 401 首先对每个分词 token 都会输出一个 1×1024 的向量作为这个分词 token 的第一特征向量，再通过多个 transformer layers 层对多个第一特征向量编码出第二特征向量，如图 6 所示，得到包括 T_1 到 T_N 以及 T'_1 到 T'_M 的多个第二特征向量，transformer layers 层包括 12 个。根据第一特征向量得到第二特征向量的方法可以是：识别出待检测文本 501 和待检测文本 502 中关键词的词性，关键词倾向于包含更多有效的信息，词性标签包含名词、动词、形容词、副词、数字或外文词。将第一特征向量输入编码层 401 中，通过编码层 401 中引入的关键词突出操作，根据第一特征向量中用于表征关键词的特征向量，对第一特征向量中用于表征文本信息的特征向量进行关键词突出，以获得待检测文本 501 和待检测文本 502 对应的多个第二特征向量。可以理解的是，图 6 中所示 transformer layers 层和全连接层 402 的数量仅为示意，本实施例对此不作限制。

[62]最后，在待检测文本 501 对应的多个第二特征向量的开头设置字符[CLS]，并通过字符[SEP]连接待检测文本 502 对应的多个第二特征向量，并将字符[CLS]设置为结尾，将设置完毕的向量语句作为全连接层 402 的输入，再经过 calss label 判断器 403 针对文本合并判断任务得到最终的输出，也即输出待检测文本 501 和待检测文本 502 是否可以合并的判断结果。

[63]本说明书实施例合理构建至少一个正样本组和负样本组，正样本组包括不可以合并

的文本，负样本组包括可以合并的文本，通过至少一个正负样本组使文本合并判断模型可以自监督式地学习两个文本中是否存在可合并的关系，直至文本合并判断模型收敛，从而提高文本合并判断模型的训练效率，以及通过至少一个正负样本对使文本合并判断模型进行多轮训练，以使训练完成的文本合并判断模型具有较好的抗干扰性和鲁棒性，执行判断两个文本是否合并的任务的准确性较高，从而得到具有完整的语义的合并文本，便于用户阅读理解。

[64]在介绍完本说明书对文本合并判断模型的设计思想之后，下面对本申请设置的应用场景进行简要说明。

[65]如图 7 所示，为本申请实施例提供的一种文本合并判断模型应用的场景示意图。该应用场景中包括终端设备 602 和服务器 601。终端设备 602 与服务器 601 之间可以通过通信网络进行通信。在一个实施例中，通信网络是有线网络或无线网络。终端设备 602 和服务器 601 可以通过有线或无线通信方式进行直接或间接的连接，本说明书实施例在此不做限制。

[66]在本申请实施例中，终端设备 602 为用户使用的电子设备，该电子设备可以是个人计算机、手机、平板电脑、笔记本、电子书阅读器等具有一定计算能力并且运行有即时通信类软件及网站或者社交类软件及网站的计算机设备。终端设备 602 可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等，但并不局限于此。

[67]服务器 601 可以是独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统，还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN(Content Delivery Network，内容分发网络)、以及大数据和人工智能平台等基础云计算服务的云服务器。

[68]文本分类模型可部署于服务器 601 上进行训练，服务器 601 中可存储有大量训练样本，包含至少一个正样本组和负样本组，用于训练文本合并判断模型。可选的，在基于本说明书实施例中的训练方法训练得到文本合并判断模型之后，可直接将训练好的文本合并判断模型部署于服务器 601 或终端设备 602 上。一般情况下都是直接将文本合并判断模型部署于服务器 601 上，在本申请实施例中，文本合并判断模型常用于对用户输入的问题和对应的两个待检测文本进行分析，以基于确定两个待检测文本是否可以合并。

[69]在一种可能的应用场景中，为了便于降低通信时延，可以在各个地区部署服务器 601，或为了负载均衡，可以由不同的服务器 601 分别去服务各个终端设备 602 对应的地区。

多个服务器 601 以通过区块链实现数据的共享，多个服务器 601 相当于多个服务器 601 组成的数据共享系统。例如终端设备 602 位于地点 a，与服务器 601 之间进行通信连接，终端设备 602 位于地点 b，与其他服务器 601 之间通信连接。

[70]在一个实施例中，如图 8 所示，图 8 为本说明书实施例提出的一种文本合并判断的方法。该方法可依赖于计算机程序实现，可运行于基于冯诺依曼体系的文本合并判断装置上。该计算机程序可集成在应用中，也可作为独立的工具类应用运行。

[71]具体而言，文本合并判断的方法包括步骤 S202 至步骤 S204。

[72]S202、获取两个待检测文本。

[73]获取两个待检测文本的方法可以获取用户在终端设备 602 上通过语音、触摸输入等方式输入的文本，或接收来自终端设备 602 发送的待检测文本。

[74]S204、将两个待检测文本输入至文本合并判断模型中，得到两个待检测文本是否可以合并的判断结果。

[75]本说明书实施例通过至少一个正样本组和负样本组训练文本合并判断模型，直至文本合并判断模型收敛，以使文本合并判断模型被用于判断两个文本是否合并，提高文本合并判断模型判断的准确性。进一步的，本实施例提供的文本合并判断模型结合目前流行的自然语言处理模型，在多个编码层后自定义地接入一层或多层的全连接层，以对多个编码层得到的多个特征向量进行特征压缩处理，提升文本合并判断模型的算法效果。

[76]需要说明的是，本申请实施例提供的文本合并判断模型可以应用于各种包含有文本合并判断的应用场景下。例如医疗领域、金融领域或教育领域中的各种自然语言处理任务中的文本合并判断这样的基础任务，但这样的基础任务往往对后续的任务至关重要。

[77]下述为本说明书装置实施例，可以用于执行本说明书方法实施例。对于本说明书装置实施例中未披露的细节，请参照本说明书方法实施例。

[78]请参见图 9，其示出了本说明书一个示例性实施例提供的文本合并判断模型的训练装置的结构示意图。该文本合并判断装置可以通过软件、硬件或者两者的结合实现成为装置的全部或一部分。该装置包括样本获取模块 901 和模型训练模块 902。

[79]样本获取模块 901，用于获取至少一个正样本组，以及获取至少一个负样本组，所述正样本组包括两个不可以合并的文本，所述负样本组包括两个可以合并的文本；模型训练模块 902，用于通过所述至少一个正样本组和所述至少一个负样本组训练所述文本

合并判断模型，直至所述文本合并判断模型收敛。

[80]在一个实施例中，样本获取模块 901，包括：样本获取单元，用于获取至少一个待分割的样本文本；样本分割单元，用于根据所述至少一个待分割的样本文本中的预设符号，分别将所述待分割的样本文本进行分割，得到至少一个负样本组。

[81]在一个实施例中，样本分割单元，包括：目标确定子单元，用于分别确定位于每个所述待分割的样本文本的中间位置的字符为目标字符；符号检测子单元，用于检测位于所述目标字符左边的 N 个字符中是否存在所述预设符号，以及检测位于所述目标字符右边的 N 个字符中是否存在预设所述预设符号，N 为大于 1 的整数；目标分割子单元，用于若位于所述目标字符左边的 N 个字符中存在预设符号，或位于所述目标字符右边的 N 个字符中存在所述预设符号，则以所述预设符号为界对每个所述待分割的样本文本进行分割，得到至少一个所述负样本组。

[82]在一个实施例中，样本分割单元，包括：符号分割子单元，用于根据每个所述待分割的样本文本中每个预设字符对应的位置，以每个所述预设符号为界对所述待分割的样本文本进行分割，得到每个所述预设符号对应的负样本组。

[83]在一个实施例中，所述文本合并判断模型包括：多个编码器、至少一个全连接层和判断器；其中，所述多个编码器，用于对所述文本进行编码，以得到所述文本对应的多个特征向量；所述至少一个全连接层，用于对两个所述文本分别对应的多个特征向量进行全连接处理，得到至少一个连接结果；所述判断器，用于根据所述至少一个连接结果，判断所述至少两个文本是否可以合并。

[84]在一个实施例中，所述至少一个全连接层包括下述一个或多个全连接层：将所有特征向量依次连接的全连接层、将每个所述文本的头部字符对应的特征向量连接的全连接层、将一个所述文本的头部字符对应的特征向量与另一个所述文本的尾部字符对应的特征向量连接的全连接层。

[85]在一个实施例中，所述判断器具体用于：对所述至少一个连接结果进行约束处理，得到所述至少两个文本可以合并的概率；根据所述至少两个文本可以合并的概率，判断所述至少两个文本是否可以合并。

[86]在一个实施例中，所述多个编码器为下述的一个或多个：双向编码器表示 BERT 模型的编码器、循环神经网络的编码器、卷积神经网络的编码器。

[87]本说明书实施例合理构建至少一个正样本组和负样本组，正样本组包括不可以合并

的文本，负样本组包括可以合并的文本，通过至少一个正负样本组使文本合并判断模型可以自监督式地学习两个文本中是否存在可合并的关系，直至文本合并判断模型收敛，从而提高文本合并判断模型的训练效率，以及通过至少一个正负样本对使文本合并判断模型进行多轮训练，以使训练完成的文本合并判断模型具有较好的抗干扰性和鲁棒性，执行判断两个文本是否合并的任务的准确性较高，从而得到具有完整的语义的合并文本，便于用户阅读理解。需要说明的是，上述实施例提供的文本合并判断模型的训练装置在执行文本合并判断模型的训练方法时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将设备的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的文本合并判断模型的训练装置与文本合并判断模型的训练方法实施例属于同一构思，其体现实现过程详见方法实施例，这里不再赘述。

[88]请参见图 10，其示出了本说明书一个示例性实施例提供的文本合并判断装置的结构示意图。该文本合并判断装置可以通过软件、硬件或者两者的结合实现成为装置的全部或一部分。该装置包括文本获取模块 1001 和结果获取模块 1002。

[89]文本获取模块 1001，用于获取两个待检测文本；结果获取模块 1002，用于将所述两个待检测文本输入至文本合并判断模型中，得到所述两个待检测文本是否可以合并的判断结果；其中，所述文本合并判断模型为采用上述实施例所述的文本合并判断模型的训练方法训练得到的模型。

[90]本说明书实施例通过至少一个正样本组和负样本组训练文本合并判断模型，直至文本合并判断模型收敛，以使文本合并判断模型被用于判断两个文本是否合并，提高文本合并判断模型判断的准确性。进一步的，本实施例提供的文本合并判断模型结合目前流行的自然语言处理模型，在多个编码层后自定义地接入一层或多层的全连接层，以对多个编码层得到的多个特征向量进行特征压缩处理，提升文本合并判断模型的算法效果。

[91]需要说明的是，上述实施例提供的文本合并判断装置在执行文本合并判断方法时，仅以上述各功能模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能模块完成，即将设备的内部结构划分成不同的功能模块，以完成以上描述的全部或者部分功能。另外，上述实施例提供的文本合并判断装置与文本合并判断方法实施例属于同一构思，其体现实现过程详见方法实施例，这里不再赘述。

[92]上述本说明书实施例序号仅仅为了描述，不代表实施例的优劣。

[93]本说明书实施例还提供了一种计算机存储介质，所述计算机存储介质可以存储有多条指令，所述指令适于由处理器加载并执行如上述图 1-图 8 所示实施例的所述文本合并判断方法，具体执行过程可以参见图 1-图 8 所示实施例的具体说明，在此不进行赘述。

[94]本说明书还提供了一种计算机程序产品，该计算机程序产品存储有至少一条指令，所述至少一条指令由所述处理器加载并执行如上述图 1-图 8 所示实施例的所述文本合并判断方法，具体执行过程可以参见图 1-图 8 所示实施例的具体说明，在此不进行赘述。

[95]请参见图 11，为本说明书实施例提供了一种电子设备的结构示意图。如图 11 所示，所述电子设备 1100 可以包括：至少一个处理器 1101，至少一个网络接口 1104，用户接口 1103，存储器 1105，至少一个通信总线 1102。

[96]其中，通信总线 1102 用于实现这些组件之间的连接通信。

[97]其中，用户接口 1103 可以包括显示屏（Display）、摄像头（Camera），可选用户接口 1103 还可以包括标准的有线接口、无线接口。

[98]其中，网络接口 1104 可选的可以包括标准的有线接口、无线接口（如 WI-FI 接口）。

[99]其中，处理器 1101 可以包括一个或者多个处理核心。处理器 1101 利用各种接口和线路连接整个电子设备 1100 内的各个部分，通过运行或执行存储在存储器 1105 内的指令、程序、代码集或指令集，以及调用存储在存储器 1105 内的数据，执行电子设备 1100 的各种功能和处理数据。可选的，处理器 1101 可以采用数字信号处理（Digital Signal Processing, DSP）、现场可编程门阵列（Field-Programmable Gate Array, FPGA）、可编程逻辑阵列（Programmable Logic Array, PLA）中的至少一种硬件形式来实现。处理器 1101 可集成处理器（Central Processing Unit, CPU）、图像处理器（Graphics Processing Unit, GPU）和调制解调器等中的一种或几种的组合。其中，CPU 主要处理操作系统、用户界面和应用程序等；GPU 用于负责显示屏所需要显示的内容的渲染和绘制；调制解调器用于处理无线通信。可以理解的是，上述调制解调器也可以不集成到处理器 1101 中，单独通过一块芯片进行实现。

[100]其中，存储器 1105 可以包括随机存储器（Random Access Memory, RAM），也可以包括只读存储器（Read-Only Memory）。可选的，该存储器 1105 包括非瞬时性计算机可读介质（non-transitory computer-readable storage medium）。存储器 1105 可用于存储指令、程序、代码、代码集或指令集。存储器 1105 可包括存储程序区和存储数据区，其中，存储程序区可存储用于实现操作系统的指令、用于至少一个功能的指令（比如触

控功能、声音播放功能、图像播放功能等)、用于实现上述各个方法实施例的指令等;存储数据区可存储上面各个方法实施例中涉及到的数据等。存储器 1105 可选的还可以是至少一个位于远离前述处理器 1101 的存储装置。如图 11 所示,作为一种计算机存储介质的存储器 1105 中可以包括操作系统、网络通信模块、用户接口模块以及应用程序,应用程序为文本合并判断模型的训练方法的应用程序和/或文本合并判断方法的应用程序。

[101]在图 11 所示的电子设备 1100 中,用户接口 1103 主要用于为用户提供输入的接口,获取用户输入的数据;而处理器 1101 可以用于调用存储器 1105 中存储的文本合并判断模型的训练应用程序,并具体执行以下操作:获取至少一个正样本组,以及获取至少一个负样本组,所述正样本组包括两个不可以合并的文本,所述负样本组包括两个可以合并的文本;通过所述至少一个正样本组和所述至少一个负样本组训练所述文本合并判断模型,直至所述文本合并判断模型收敛。

[102]在一个实施例中,处理器 1101 执行所述获取至少一个负样本组,具体执行:获取至少一个待分割的样本文本;根据所述至少一个待分割的样本文本中的预设符号,分别将所述待分割的样本文本进行分割,得到至少一个负样本组。

[103]在一个实施例中,处理器 1101 执行所述根据所述至少一个待分割的样本文本中的预设字符,分别将所述待分割的样本文本进行分割,得到至少一个负样本组,具体执行:分别确定位于每个所述待分割的样本文本的中间位置的字符为目标字符;检测位于所述目标字符左边的 N 个字符中是否存在所述预设符号,以及检测位于所述目标字符右边的 N 个字符中是否存在预设所述预设符号, N 为大于 1 的整数;若位于所述目标字符左边的 N 个字符中存在预设符号,或位于所述目标字符右边的 N 个字符中存在所述预设符号,则以所述预设符号为界对每个所述待分割的样本文本进行分割,得到至少一个所述负样本组。

[104]在一个实施例中,处理器 1101 执行所述根据所述至少一个待分割的样本文本中的预设字符,分别将所述待分割的样本文本进行分割,得到至少一个负样本组,具体执行:根据每个所述待分割的样本文本中每个预设字符对应的位置,以每个所述预设符号为界对所述待分割的样本文本进行分割,得到每个所述预设符号对应的负样本组。

[105]在一个实施例中,所述文本合并判断模型包括:多个编码器、至少一个全连接层和判断器;其中,所述多个编码器,用于对所述文本进行编码,以得到所述文本对应的多个特征向量;所述至少一个全连接层,用于对两个所述文本分别对应的多个特征向量进

行全连接处理,得到至少一个连接结果;所述判断器,用于根据所述至少一个连接结果,判断所述至少两个文本是否可以合并。

[106]在一个实施例中,至少一个全连接层包括下述一个或多个全连接层:将所有特征向量依次连接的全连接层、将每个所述文本的头部字符对应的特征向量连接的全连接层、将一个所述文本的头部字符对应的特征向量与另一个所述文本的尾部字符对应的特征向量连接的全连接层。

[107]在一个实施例中,所述判断器具体用于:对所述至少一个连接结果进行约束处理,得到所述至少两个文本可以合并的概率;根据所述至少两个文本可以合并的概率,判断所述至少两个文本是否可以合并。

[108]在一个实施例中,所述多个编码器为下述的一个或多个:双向编码器表示 BERT 模型的编码器、循环神经网络的编码器、卷积神经网络的编码器。

[109]在一个实施例中,处理器 1101 可以用于调用存储器 1105 中存储的文本合并判断应用程序,并具体执行以下操作:获取两个待检测文本;将所述两个待检测文本输入至文本合并判断模型中,得到所述两个待检测文本是否可以合并的判断结果;其中,所述文本合并判断模型为采用上述实施例所述的文本合并判断模型的训练方法训练得到的模型。

[110]本说明书实施例合理构建至少一个正样本组和负样本组,正样本组包括不可以合并的文本,负样本组包括可以合并的文本,通过至少一个正负样本组使文本合并判断模型可以自监督式地学习两个文本中是否存在可合并的关系,直至文本合并判断模型收敛,从而提高文本合并判断模型的训练效率,以及通过至少一个正负样本对使文本合并判断模型进行多轮训练,以使训练完成的文本合并判断模型具有较好的抗干扰性和鲁棒性,执行判断两个文本是否合并的任务的准确性较高,从而得到具有完整的语义的合并文本,便于用户阅读理解。本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,所述的程序可存储于一计算机可读取存储介质中,该程序在执行时,可包括如上述各方法的实施例的流程。其中,所述的存储介质可为磁碟、光盘、只读存储记忆体或随机存储记忆体等。

[111]以上所揭露的仅为本说明书较佳实施例而已,当然不能以此来限定本说明书之权利范围,因此依本说明书权利要求所做的等同变化,仍属本说明书所涵盖的范围。

权利要求书

1、一种文本合并判断模型的训练方法，所述方法包括：

获取至少一个正样本组，以及获取至少一个负样本组，所述正样本组包括两个不能够合并的文本，所述负样本组包括两个能够合并的文本；

通过所述至少一个正样本组和所述至少一个负样本组训练所述文本合并判断模型，直至所述文本合并判断模型收敛。

2、根据权利要求1所述的方法，所述获取至少一个负样本组，包括：

获取至少一个待分割的样本文本；

根据所述至少一个待分割的样本文本中的预设符号，分别将所述待分割的样本文本进行分割，得到至少一个负样本组。

3、根据权利要求2所述的方法，所述根据所述至少一个待分割的样本文本中的预设符号，分别将所述待分割的样本文本进行分割，得到至少一个负样本组，包括：

分别确定位于每个所述待分割的样本文本的中间位置的字符为目标字符；

检测位于所述目标字符左边的N个字符中是否存在所述预设符号，以及检测位于所述目标字符右边的N个字符中是否存在所述预设符号，N为大于1的整数；

若位于所述目标字符左边的N个字符中存在预设符号，或位于所述目标字符右边的N个字符中存在所述预设符号，则以所述预设符号为界对每个所述待分割的样本文本进行分割，得到至少一个所述负样本组。

4、根据权利要求2所述的方法，所述根据所述至少一个待分割的样本文本中的预设符号，分别将所述待分割的样本文本进行分割，得到至少一个负样本组，包括：

根据每个所述待分割的样本文本中每个预设符号对应的位置，以每个所述预设符号为界对所述待分割的样本文本进行分割，得到每个所述预设符号对应的负样本组。

5、根据权利要求1所述的方法，所述文本合并判断模型包括：多个编码器、至少一个全连接层和判断器；

其中，所述多个编码器，用于对所述文本进行编码，以得到所述文本对应的多个特征向量；

所述至少一个全连接层，用于对两个所述文本分别对应的多个特征向量进行全连接处理，得到至少一个连接结果；

所述判断器，用于根据所述至少一个连接结果，判断所述至少两个文本是否能够合并。

6、根据权利要求5所述的方法，所述至少一个全连接层包括下述一个或多个全连

接层：将所有特征向量依次连接的全连接层、将每个所述文本的头部字符对应的特征向量连接的全连接层、将一个所述文本的头部字符对应的特征向量与另一个所述文本的尾部字符对应的特征向量连接的全连接层。

7、根据权利要求5所述的方法，所述判断器具体用于：

对所述至少一个连接结果进行约束处理，得到所述至少两个文本能够合并的概率；根据所述至少两个文本能够合并的概率，判断所述至少两个文本是否能够合并。

8、根据权利要求5所述的方法，所述多个编码器为下述的一个或多个：双向编码器表示BERT模型的编码器、循环神经网络的编码器、卷积神经网络的编码器。

9、一种文本合并判断的方法，所述方法包括：

获取两个待检测文本；

将所述两个待检测文本输入至文本合并判断模型中，得到所述两个待检测文本是否能够合并的判断结果；其中，所述文本合并判断模型为采用权利要求1至8任意一项所述的文本合并判断模型的训练方法训练得到的模型。

10、一种文本合并判断模型的训练装置，所述装置包括：

样本获取模块，用于获取至少一个正样本组，以及获取至少一个负样本组，所述正样本组包括两个不能够合并的文本，所述负样本组包括两个能够合并的文本；

模型训练模块，用于通过所述至少一个正样本组和所述至少一个负样本组训练所述文本合并判断模型，直至所述文本合并判断模型收敛。

11、一种文本合并判断的装置，所述装置包括：

文本获取模块，用于获取两个待检测文本；

结果获取模块，用于将所述两个待检测文本输入至文本合并判断模型中，得到所述两个待检测文本是否能够合并的判断结果；其中，所述文本合并判断模型为采用权利要求1至8任意一项所述的文本合并判断模型的训练方法训练得到的模型。

12、一种计算机存储介质，所述计算机存储介质存储有多条指令，所述指令适于由处理器加载并执行如权利要求1~9任意一项所述的方法。

13、一种计算机程序产品，所述计算机程序产品存储有多条指令，所述指令适于由处理器加载并执行如权利要求1~9任意一项所述的方法。

14、一种电子设备，包括：处理器和存储器；其中，所述存储器存储有计算机程序，所述计算机程序适于由所述处理器加载并执行如权利要求1~9任意一项所述的方法。