(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111797609 B
(45)授权公告日 2024.10.25

(21)申请号 202010639435.6

(22)申请日 2020.07.03

(65)同一申请的已公布的文献号
申请公布号 CN 111797609 A

(43)申请公布日 2020.10.20

(73)专利权人 阳光保险集团股份有限公司
地址 518000 广东省深圳市福田区红荔西路7002号第一世界广场A座17层

(72)发明人 蔡岩松 杜新凯 牛国扬 王彦昕 刘谦 高峰

(74)专利代理机构 北京超凡宏宇知识产权代理有限公司 11463
专利代理师 唐正瑜

(51)Int.Cl.
G06F 40/205(2020.01)
G06F 40/295(2020.01)
G06F 16/35(2019.01)

(56)对比文件
CN 110427627 A,2019.11.08
CN 111079447 A,2020.04.28

审查员 顾瑜尉

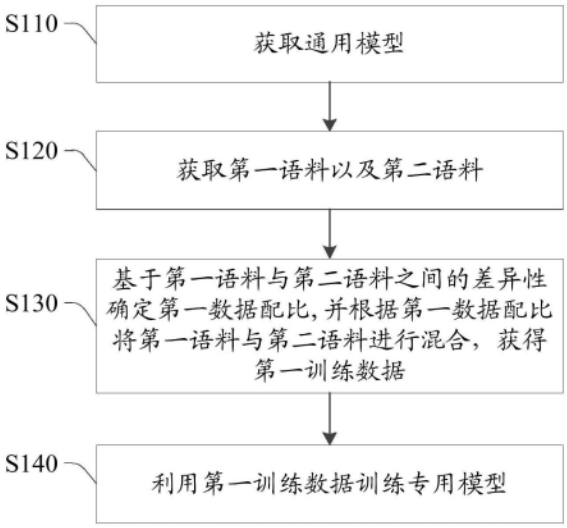
权利要求书2页 说明书13页 附图2页

(54)发明名称

模型训练方法及装置

(57)摘要

本申请涉及自然语言处理技术领域,提供一种模型训练方法及装置。其中,模型训练方法包括:获取通用模型,通用模型为预训练的、与任务无关的语言模型;获取第一语料以及第二语料,第一语料为通用领域内的语料,第二语料为目标领域内与目标任务相关的语料;基于第一语料与第二语料之间的差异性确定第一数据配比,并根据第一数据配比将两种语料混合,获得第一训练数据;利用第一训练数据训练用于执行目标任务的专用模型,专用模型中包括通用模型以及与目标任务相关的适配结构。该方法可视为一种对通用模型进行继续训练以实现领域偏移的解决方案,并且通过在第一训练数据中合理配比第一语料与第二语料,使得训练得到的专用模型的性能得以改善。



CN 111797609 B

1.一种模型训练方法,其特征在于,包括:

获取通用模型,所述通用模型为预训练的、与任务无关的语言模型;

获取第一语料以及第二语料;其中,所述第一语料为通用领域内的语料,所述第二语料为目标领域内与目标任务相关的语料,所述目标任务为自然语言处理任务,所述目标领域为所述目标任务所属的领域;

基于所述第一语料与所述第二语料之间的差异性确定第一数据配比,并根据所述第一数据配比将所述第一语料与所述第二语料进行混合,获得第一训练数据;其中,所述第一语料与所述第二语料之间的差异性和所述第一数据配比负相关;

利用所述第一训练数据训练用于执行所述目标任务的专用模型;其中,所述专用模型中包括所述通用模型以及与所述目标任务相关的适配结构;

所述获取通用模型,包括:

获取原始通用模型,所述原始通用模型为预训练的、所述通用领域内的语言模型;

获取第三语料以及第四语料;其中,所述第三语料为所述通用领域内的语料,所述第四语料为所述目标领域内的语料;

基于所述第三语料与所述第四语料之间的差异性确定第二数据配比,并根据所述第二数据配比将所述第三语料与所述第四语料进行混合,获得第二训练数据;其中,所述第三语料与所述第四语料之间的差异性和所述第二数据配比负相关;

利用所述第二训练数据训练所述原始通用模型,获得所述通用模型。

2.根据权利要求1所述的训练方法,其特征在于,所述基于所述第一语料与所述第二语料之间的差异性确定数据配比,包括:

获取第一差异系数,所述第一差异系数与所述目标领域内的关键词在所述目标领域内的测试语料中出现的频次正相关;

根据所述第一语料中的文本长度与所述第二语料中的文本长度之间的差异性计算第二差异系数,所述第二差异系数与文本长度之间的差异性正相关;

根据所述第一差异系数以及所述第二差异系数确定所述第一数据配比。

3.根据权利要求2所述的训练方法,其特征在于,所述目标任务为抽取式阅读理解任务,所述根据所述第一语料中的文本长度与所述第二语料中的文本长度之间的差异性计算第二差异系数,包括:

计算所述第一语料中阅读理解的文章平均长度L1、问题平均长度L2以及答案平均长度L3;

计算所述第二语料中阅读理解的文章平均长度P1、问题平均长度P2以及答案平均长度P3;

根据P1与L1的差异性、P2与L2的差异性以及P3与L3的差异性计算所述第二差异系数;其中,所述第二差异系数分别与所述P1与L1的差异性、所述P2与L2的差异性以及所述P3与L3的差异性正相关。

4.根据权利要求1所述的训练方法,其特征在于,所述利用所述第一训练数据训练用于执行所述目标任务的专用模型,包括:

在利用所述第一训练数据训练所述专用模型的过程中,定期利用验证集评估所述专用模型的收敛程度,并根据所述收敛程度设置训练过程中使用的学习率;其中,所述学习率被

设置为与所述收敛程度负相关。

5.根据权利要求4所述的训练方法,其特征在于,所述学习率为带Warm-up的衰减学习率,所述在利用所述第一训练数据训练所述专用模型的过程中,定期利用验证集评估所述专用模型的收敛程度,并根据所述收敛程度设置训练过程中使用的学习率,包括:

在利用所述第一训练数据训练所述专用模型的过程中的学习率衰减阶段,定期利用验证集评估所述专用模型的收敛程度,并根据所述收敛程度减小所述学习率的取值;其中,所述学习率取值的减小量与所述收敛程度负相关。

6.根据权利要求1所述的训练方法,其特征在于,所述目标任务为抽取式阅读理解任务,所述目标任务的答案满足第一统计规律,获取第二语料,包括:

根据所述目标领域搜索和/或构造用于抽取式阅读理解的语料,并从中筛选出答案满足所述第一统计规律的语料作为所述第二语料。

7.一种模型训练装置,其特征在于,包括:

第一模型获取模块,用于获取通用模型,所述通用模型为预训练的、与任务无关的语言模型;

第一数据获取模块,用于获取第一语料以及第二语料;其中,所述第一语料为通用领域内的语料,所述第二语料为目标领域内与目标任务相关的语料,所述目标任务为自然语言处理任务,所述目标领域为所述目标任务所属的领域;

第一数据混合模块,用于基于所述第一语料与所述第二语料之间的差异性确定第一数据配比,并根据所述第一数据配比将所述第一语料与所述第二语料进行混合,获得第一训练数据;其中,所述第一语料与所述第二语料之间的差异性和所述第一数据配比负相关;

第一训练模块,用于利用所述第一训练数据训练用于执行所述目标任务的专用模型;其中,所述专用模型中包括所述通用模型以及与所述目标任务相关的适配结构;

所述第一模型获取模块获取通用模块,包括:获取原始通用模型,所述原始通用模型为预训练的、所述通用领域内的语言模型;获取第三语料以及第四语料;其中,所述第三语料为所述通用领域内的语料,所述第四语料为所述目标领域内的语料;基于所述第三语料与所述第四语料之间的差异性确定第二数据配比,并根据所述第二数据配比将所述第三语料与所述第四语料进行混合,获得第二训练数据;其中,所述第三语料与所述第四语料之间的差异性和所述第二数据配比负相关;利用所述第二训练数据训练所述原始通用模型,获得所述通用模型。

模型训练方法及装置

技术领域

[0001] 本发明涉及自然语言处理技术领域,具体而言,涉及一种模型训练方法及装置。

背景技术

[0002] 近年来,机器阅读理解被广泛地应用于各种文章的动态信息抽取,以及各种问答机器人的辅助上。在自然语言处理领域,一般利用训练好的神经网络模型执行特定领域(如金融保险领域、政策法规领域、教育领域、通信及IT领域)内的阅读理解任务。然而,现有的预训练模型都是通用领域内的,用通用领域内的预训练模型去执行特定领域内的阅读理解任务会损失一定程度的准确度,因此,需要将通用领域内的预训练模型继续训练,来实现领域偏移,但对于该继续训练过程应该如何实现现有技术中并没有很好的解决方案。

发明内容

[0003] 本申请实施例的目的在于提供一种模型训练方法及装置,以改善上述技术问题。

[0004] 为实现上述目的,本申请提供如下技术方案:

[0005] 第一方面,本申请实施例提供一种模型训练方法,包括:获取通用模型,所述通用模型为预训练的、与任务无关的语言模型;获取第一语料以及第二语料;其中,所述第一语料为通用领域内的语料,所述第二语料为目标领域内与目标任务相关的语料,所述目标任务为自然语言处理任务,所述目标领域为所述目标任务所属的领域;基于所述第一语料与所述第二语料之间的差异性确定第一数据配比,并根据所述第一数据配比将所述第一语料与所述第二语料进行混合,获得第一训练数据;其中,所述第一语料与所述第二语料之间的差异性和所述第一数据配比负相关;利用所述第一训练数据训练用于执行所述目标任务的专用模型;其中,所述专用模型中包括所述通用模型以及与所述目标任务相关的适配结构。

[0006] 在上述方法中,通用模型是一个与任务无关的预训练模型,专用模型则是用于执行特定领域的目标任务的模型,并且该专用模型中包括通用模型,因此训练该专用模型的方案也可以视为一种对通用模型进行继续训练以实现领域偏移的解决方案。上述目标任务是自然语言处理任务,但并不限于阅读理解任务,也可以是文本分类任务、命名实体识别任务等。

[0007] 在该方案中,首先,利用第一语料(通用领域内的数据)与第二语料(特定领域内的数据)混合产生第一训练数据,既确保了第一训练数据具有良好的知识强度,又使其不会过分偏向于领域知识,而放弃了通用领域内的语言表达方式,甚至出现过拟合等问题,从而有利于改善训练得到的专用模型的性能。其次,该方案基于第一语料与第二语料之间的差异性确定两种数据在第一训练数据中的配比,有利于合理地平衡模型的领域性和通用性,进一步改善训练得到的专用模型的性能。

[0008] 在第一方面的一种实现方式中,所述基于所述第一语料与所述第二语料之间的差异性确定数据配比,包括:获取第一差异系数,所述第一差异系数与所述目标领域内的关键词在所述目标领域内的测试语料中出现的频次正相关;根据所述第一语料中的文本长度与

所述第二语料中的文本长度之间的差异性计算第二差异系数,所述第二差异系数与文本长度之间的差异性正相关;根据所述第一差异系数以及所述第二差异系数确定所述第一数据配比。

[0009] 在上述实现方式中,目标领域内的关键词可由领域专家指定,这些关键词基本只在目标领域内的语料中出现,在通用领域内则很少出现,从而,第一差异系数可以表征第一语料与第二语料在知识层面上的差异性,或者说主要描述的是两类语料因所属的领域不同而产生的差异。而第二差异系数由于是根据第一语料中与第二语料中的文本长度之间的差异性计算得到的,所以表征的是第一语料与第二语料在语言结构层面上的差异性,或者说主要描述的是两类语料本身的语言特征带来的差异。综合这两种差异性可以全面、有效地反映第一语料与第二语料之间的差异性。

[0010] 在第一方面的一种实现方式中,所述目标任务为抽取式阅读理解任务,所述根据所述第一语料中的文本长度与所述第二语料中的文本长度之间的差异性计算第二差异系数,包括:计算所述第一语料中阅读理解的文章平均长度L1、问题平均长度L2以及答案平均长度L3;计算所述第二语料中阅读理解的文章平均长度P1、问题平均长度P2以及答案平均长度P3;根据P1与L1的差异性、P2与L2的差异性以及P3与L3的差异性计算所述第二差异系数;其中,所述第二差异系数分别与所述P1与L1的差异性、所述P2与L2的差异性以及所述P3与L3的差异性正相关。

[0011] 抽取式阅读理解任务是指问题的答案为文章原文中的片段的阅读理解任务,抽取式阅读理解任务的三个要素是文章、问题以及答案。因此,在针对抽取式阅读理解任务时,语料中本文的长度包括文章、问题以及答案的长度。从而,在计算第一语料和第二语料的第二差异系数时,需要同时考虑两类语料中文章长度的差异性、问题长度的差异性以及答案长度的差异性。

[0012] 在第一方面的一种实现方式中,所述利用所述第一训练数据训练用于执行所述目标任务的专用模型,包括:在利用所述第一训练数据训练所述专用模型的过程中,定期利用验证集评估所述专用模型的收敛程度,并根据所述收敛程度设置训练过程中使用的学习率;其中,所述学习率被设置为与所述收敛程度负相关。

[0013] 在上述实现方式中,学习率的调整是一种根据模型的收敛程度而进行的动态调整,即根据模型训练的实际表现果来调整学习率,特别是在训练后期,使得学习率能够随着模型的收敛而衰减,调整方式灵活有效。

[0014] 在第一方面的一种实现方式中,所述学习率为带Warm-up的衰减学习率,所述在利用所述第一训练数据训练所述专用模型的过程中,定期利用验证集评估所述专用模型的收敛程度,并根据所述收敛程度设置训练过程中使用的学习率,包括:在利用所述第一训练数据训练所述专用模型的过程中的学习率衰减阶段,定期利用验证集评估所述专用模型的收敛程度,并根据所述收敛程度减小所述学习率的取值;其中,所述学习率取值的减小量与所述收敛程度负相关。

[0015] 带Warm-up的衰减学习率在训练初期使用一个较小的学习率,待模型稳定后再选择预设的学习率进行学习,即所谓的学习率预热(Warm-up),此举有利于加快模型的收敛速度,改善模型的训练效果。在使用预设的学习学习一段时间后,学习率开始衰减,避免在模型已经趋于收敛时因学习率过大产生过拟合等问题,在学习率的衰减过程中利用验证集进

行动态的学习率调整,调整方式灵活、客观,同样有利于改善训练效果。

[0016] 在第一方面的一种实现方式中,所述目标任务为抽取式阅读理解任务,所述目标任务的答案满足第一统计规律,获取第二语料,包括:根据所述目标领域搜索和/或构造用于抽取式阅读理解的语料,并从中筛选出答案满足所述第一统计规律的语料作为所述第二语料。

[0017] 在上述实现方式中,第二语料的获取过程与训练好的专用模型实际要执行的目标任务高度适配,这样获取到的语料更具针对性,训练效果也越理想(针对要执行的目标任务而言)。

[0018] 在第一方面的一种实现方式中,所述获取通用模型,包括:获取原始通用模型,所述原始通用模型为预训练的、所述通用领域内的语言模型;获取第三语料以及第四语料;其中,所述第三语料为所述通用领域内的语料,所述第四语料为所述目标领域内的语料;基于所述第三语料与所述第四语料之间的差异性确定第二数据配比,并根据所述第二数据配比将所述第三语料与所述第四语料进行混合,获得第二训练数据;其中,所述第三语料与所述第四语料之间的差异性和所述第二数据配比负相关;利用所述第二训练数据训练所述原始通用模型,获得所述通用模型。

[0019] 原始通用模型可以指现有的公开模型,例如ERNIE、BERT模型等,上述实现方式先通过训练对原始通用模型进行参数微调,获得通用模型,然后再基于通用模型构建专用模型并进一步训练。其中,对原始通用模型进行参数微调,其使用的第二训练数据类似于第一训练数据,也是由通用领域内的数据和目标领域内的数据构成的,并且在数据的配比上也与第一训练数据类似,从而有利于改善训练好的通用模型的质量。

[0020] 需要指出,第二训练数据和第一训练数据在形式上还是有所区别的,因为第一训练数据针对的是下游的目标任务,例如,阅读理解任务、文本分类任务等;而第二训练数据是与目标任务无关的,由上游的预训练任务所使用,例如,Masked LM、Next Sentence Prediction等任务。

[0021] 第二方面,本申请实施例提供一种模型训练方法,包括:获取原始通用模型,所述原始通用模型为预训练的、通用领域内的语言模型;获取第三语料以及第四语料;其中,所述第三语料为所述通用领域内的语料,所述第四语料为目标任务所属的目标领域内的语料,所述目标任务为自然语言处理任务;基于所述第三语料与所述第四语料之间的差异性确定第二数据配比,并根据所述第二数据配比将所述第三语料与所述第四语料进行混合,获得第二训练数据;其中,所述第三语料与所述第四语料之间的差异性和所述第二数据配比负相关;利用所述第二训练数据训练所述原始通用模型,获得通用模型;其中,所述通用模型用于与所述目标任务相关的适配结构组成专用模型,所述专用模型为用于执行所述目标任务的模型。

[0022] 第二方面提供的方法和第一方面的最后一种实现方式提供的方法获得通用模型的方式类似,但在第二方面提供的方法中,允许基于通用模型直接构建专用模型,并且允许该专用模型不经训练直接投入使用(当然该专用模型训练后再投入使用也是可以的)。由于在训练产生通用模型的过程中,合理地混合了通用领域内的数据和目标领域内的数据,使得获得的通用模型在领域性和通用性之间取得了良好的平衡,从而基于该通用模型构建的专用模型的性能也得到改善。

[0023] 第三方面,本申请实施例提供一种模型训练装置,包括:第一模型获取模块,用于获取通用模型,所述通用模型为预训练的、与任务无关的语言模型;第一数据获取模块,用于获取第一语料以及第二语料;其中,所述第一语料为通用领域内的语料,所述第二语料为目标领域内与目标任务相关的语料,所述目标任务为自然语言处理任务,所述目标领域为所述目标任务所属的领域;第一数据混合模块,用于基于所述第一语料与所述第二语料之间的差异性确定第一数据配比,并根据所述第一数据配比将所述第一语料与所述第二语料进行混合,获得第一训练数据;其中,所述第一语料与所述第二语料之间的差异性和所述第一数据配比负相关;第一训练模块,用于利用所述第一训练数据训练用于执行所述目标任务的专用模型;其中,所述专用模型中包括所述通用模型以及与所述目标任务相关的适配结构。

[0024] 第四方面,本申请实施例提供一种模型训练装置,包括:第二模型获取模块,获取原始通用模型,所述原始通用模型为预训练的、通用领域内的语言模型;第二数据获取模块,获取第三语料以及第四语料;其中,所述第三语料为所述通用领域内的语料,所述第四语料为目标任务所属的目标领域内的语料,所述目标任务为自然语言处理任务;第二数据混合模块,基于所述第三语料与所述第四语料之间的差异性确定第二数据配比,并根据所述第二数据配比将所述第三语料与所述第四语料进行混合,获得第二训练数据;其中,所述第三语料与所述第四语料之间的差异性和所述第二数据配比负相关;第二训练模块,利用所述第二训练数据训练所述原始通用模型,获得通用模型;其中,所述通用模型用于与所述目标任务相关的适配结构组成专用模型,所述专用模型为用于执行所述目标任务的模型。

[0025] 第五方面,本申请实施例提供一种计算机可读存储介质,所述计算机可读存储介质上存储有计算机程序指令,所述计算机程序指令被处理器读取并运行时,执行第一方面或第一方面的任意一种可能的实现方式提供的方法。

[0026] 第六方面,本申请实施例提供一种电子设备,包括:存储器以及处理器,所述存储器中存储有计算机程序指令,所述计算机程序指令被所述处理器读取并运行时,执行第一方面或第一方面的任意一种可能的实现方式提供的方法。

附图说明

[0027] 为了更清楚地说明本申请实施例的技术方案,下面将对本申请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申请的某些实施例,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。

[0028] 图1示出了本申请实施例提供的一种模型训练方法的流程图;

[0029] 图2示出了本申请实施例提供的一种模型训练装置的模块图;

[0030] 图3示出了本申请实施例提供的另一种模型训练装置的模块图;

[0031] 图4示出了本申请实施例提供的一种电子设备的示意图。

具体实施方式

[0032] 下面将结合本申请实施例中的附图,对本申请实施例中的技术方案进行描述。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。术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。术语“第一”、“第二”、“第三”等仅用于区分描述,而不能理解为指示或暗示相对重要性。

[0033] 图1示出了本申请实施例提供的一种模型训练方法的流程图。该方法可以但不限于由一电子设备执行,图4示出了该电子设备可能的结构,具体见后文关于图4的阐述。参照图1,该方法包括:

[0034] 步骤S110:获取通用模型。

[0035] 通用模型为与特定的自然语言处理任务无关的语言模型,就其设计目的而言,主要是为了学习语言本身的表达方式,并不用于(或者说多数情况下无法直接用于)执行具体任务;相对地,专用模型(在步骤S140中出现)为用于执行特定的自然语言处理任务的模型,为简单起见,后文将该特定的自然语言处理任务简称目标任务,目标任务可以是阅读理解任务、文本分类任务、命名实体识别任务等。

[0036] 目标任务往往涉及某个特定的领域,例如,金融保险领域、政策法规领域、教育领域、通信及IT领域等,不妨将其称为目标领域。相对地,一些文本并不属于某个特定的领域,例如来自日常对话,则可认为其属于通用领域。

[0037] 通用模型和专用模型在结构上均可采用神经网络实现,专用模型可由通用模型以及与目标任务相关的适配结构构成,例如,在通用模型之后加上全连接层、sofmax层等即可构成一个能够用于执行目标任务的专用模型,这里的全连接层、sofmax层即与目标任务相关的适配结构,可根据不同的目标任务进行不同的设计。多数情况下,通用模型是专用模型的主体结构,适配结构相对简单。

[0038] 通用模型是经过预训练的模型,训练通用模型的任务可称为预训练任务,预训练任务也可称为上游任务,而相对地,目标任务则可称为下游任务。预训练任务一般是与特定领域无关的,经常使用的预训练任务包括Masked LM、Next Sentence Prediction等。通用模型的获取方式主要有两种:

[0039] 方式一:直接从公开渠道获取,例如,Google公司发布的BERT中文模型、百度公司发布的ERNIE模型等。这些模型往往利用通用领域内的语料训练产生,模型输入可以是语句,输出则是对语句的向量化表示,或者可视为提取了语句的特征,后续则可基于这些特征做相应的下游任务。相对而言,ERNIE在中文处理上有一些优势,具体为:

[0040] (1)相比BERT中文模型,ERNIE的预训练任务中的Masked LM任务采用了更适合中文习惯的词遮盖,使得ERNIE在处理中文任务的时候有着更好的表现。

[0041] (2)相比BERT中文模型,ERNIE的预训练语料的数量质量都更优秀,使得ERNIE对于执行中文任务效果更好。

[0042] 因此,若目标任务为中文任务,则可优先选取ERNIE模型。不过应当理解,由于语言模型也处在不断的改进中,不排除BERT中文模型改进后质量超过ERNIE模型,因此模型选择策略并非一成不变的。

[0043] 方式二:先从公开渠道获取模型,不妨称此时获取到的为原始通用模型,然后通过

训练微调原始通用模型的参数,得到通用模型。例如,可以采用目标领域内的语料微调原始通用模型的参数,以使原始通用模型适当向目标领域偏移。关于方式二,后文还会进一步举例说明,这里暂不具体介绍。

[0044] 相对而言,通过方式一获得通用模型更加简单,通过方式二获得的通用模型由于进行了参数微调,质量可能更好,也更有利于后期对专用模型的训练。

[0045] 步骤S120:获取第一语料以及第二语料。

[0046] 第一语料是指通用领域内的语料,第二语料是指目标领域内与目标任务相关的语料。可以理解的,第二语料在选取上与目标任务的适配程度越高,训练获得的专用模型在执行目标任务时也会取得越好的效果。

[0047] 以目标任务为抽取式阅读理解任务为例,所谓抽取式阅读理解任务是指:阅读一篇文章后回答针对该文章的问题,该问题的答案必然是原文中的片段(如原文中的词语或句子)。若目标任务的答案满足第一统计规律,则可以根据目标领域搜索和/或构造用于抽取式阅读理解的语料,并从中筛选出答案满足第一统计规律的语料作为第二语料。下面举具体的例子解释其含义:

[0048] (1)若目标任务的答案较短(这里答案较短就是第一统计规律),例如答案是一些命名实体或时间(比如,在机票预订场景中,要从客户所说的一段话中分析出机票的目的地和出发时间),则可以首先根据目标领域搜索和/或构造用于抽取式阅读理解的语料,然后从中筛选出答案较短的和/或答案是命名实体或时间的语料作为第二语料。

[0049] (2)若目标任务的答案较长(这里答案较长就是第一统计规律),例如答案是对某些概念的解释或者对某些事实的阐述(比如,在历史课教学场景中,要根据课文内容回答推翻封建帝制给中国社会带来的影响),则可以首先根据目标领域搜索和/或构造用于抽取式阅读理解的语料,然后从中筛选出答案较长和/或答案是对某些概念的解释或者对某些事实的阐述的语料作为第二语料。

[0050] (3)若目标任务的答案可统计分析出某种规律,例如,设计某个专门用于回答理科问题的问答机器人,此时很难从答案的长度上去寻找统计规律,但可以确定答案都是理科方面的,从而答案属于理科就是第一统计规律。此时,可以首先根据目标领域搜索和/或构造用于抽取式阅读理解的语料,然后从中筛选出答案属于理科的语料作为第二语料。

[0051] 关于语料的获取既可以由程序自动执行,也可以由人工获取,或者也可以由程序进行初步筛选,然后由人工进行验证或进一步筛选。

[0052] 步骤S130:基于第一语料与第二语料之间的差异性确定第一数据配比,并根据第一数据配比将第一语料与第二语料进行混合,获得第一训练数据。

[0053] 之前提到过通用模型是与具体的任务无关的,因此要使得到的专用模型与目标任务相关,必须在专用模型的训练数据中引入与目标任务相关的第二语料以实现领域偏移。

[0054] 然而,在本申请的方案中,第一训练数据并非只使用第二语料,而是将第一语料与第二语料混合使用,其原因在于:特定领域内的数据(第二语料)虽然具有高强度的领域知识,但语言的表达能力不足,通用领域内的数据(第一语料)虽然不具备或者基本不具备领域知识,但语言表达能力较强,混合使用两种类型的语料既确保了第一训练数据具有良好的知识强度,又使其不会过分偏向于领域知识,而放弃了通用领域内的语言表达能力,甚至出现过拟合等问题,从而有利于改善训练得到的专用模型的性能。

[0055] 进一步的,在第一训练数据中第一语料与第二语料的配比(即第一数据配比,例如可以指第一训练数据中第一语料的样本数量与第二语料的样本数量的比值)十分重要,因为该配比直接决定了专用模型在领域性和通用性上的平衡,从而合理地设置第一数据配比,有利于改善训练得到的专用模型的性能。

[0056] 在本申请的方案中,第一数据配比基于第一语料与第二语料之间的差异性确定,且第一语料与所述第二语料之间的差异性和第一数据配比负相关。若第一语料与第二语料差异越大,说明目标领域的领域性越强,此时第一语料在第一训练数据中的占比将越小,因为需要更多的领域数据才能反映出强领域的这一特征;反之,若第一语料与第二语料差异越小,说明目标领域的领域性越弱,此时第一语料在第一训练数据中的占比将越大。本申请并不限定负相关关系的具体形式,例如,在最简单的情况下可以是反比例关系,但也可以是其他关系。

[0057] 在一些实现方式中,第一数据配比可以这样计算:

[0058] 步骤A:获取第一差异系数(记为λ1);其中,第一差异系数与目标领域内的关键词在目标领域内的测试语料中出现的频次正相关。本申请并不限定正相关关系的具体形式,例如,在最简单的情况下可以是正比例关系,但也可以是其他关系。

[0059] 目标领域内的关键词可由领域专家指定,这些关键词具有代表性,并且基本只在目标领域内的语料中出现,在通用领域内则很少出现,例如,通信及IT领域内的关键词“调制”、“解调”、“频分复用”等在通用领域内很少被使用。从而,第一差异系数可以表征第一语料与第二语料在知识层面上的差异性,或者说主要描述的是两类语料因所属的领域不同而产生的差异。目标领域内的测试语料选取方式不限,但该语料与第二语料无必然联系。下面列出了一些领域内通过上述方法算得的第一差异系数:

[0060]

领域	金融保险领域	政策法规领域	教育领域	通信及IT领域
第一差异系数	0.75	0.91	0.32	1

[0061] 其中,通信及IT领域的第一差异系数被人为设定成1(人为认定该领域的领域性很强),其余领域的第一差异系数都以通信及IT领域的第一差异系数为参考(不超过1)相应计算。

[0062] 第一差异系数的取值只和语料所属的领域有关,从而计算好后可以长期使用,不用或者至少不用经常性更新。例如,可以提前计算好第一差异系数并将结果存储到类似上面表格中,在需要计算第一数据配比时直接查询表格。

[0063] 步骤B:根据第一语料中的文本长度与第二语料中的文本长度之间的差异性计算第二差异系数(记为λ1);其中,第二差异系数与文本长度之间的差异性正相关。本申请并不限定正相关关系的具体形式,例如,在最简单的情况下可以是正比例关系,但也可以是其他关系。

[0064] 第二差异系数由于是根据第一语料中与第二语料中的文本长度之间的差异性计算得到的,所以表征的是第一语料与第二语料在语言结构层面上的差异性,或者说主要描述的是两类语料本身的语言特征带来的差异。

[0065] 例如,对于目标任务为抽取式阅读理解任务的情况,第一差异系数可以这样计算:

[0066] 首先,计算第一语料(例如,可以是SQUAD数据集)中阅读理解的文章平均长度L1、问题平均长度L2以及答案平均长度L3;

[0067] 然后,计算第二语料中阅读理解的文章平均长度P1、问题平均长度P2以及答案平均长度P3;

[0068] 最后,根据P1与L1的差异性、P2与L2的差异性以及P3与L3的差异性计算第二差异系数;其中,第二差异系数分别与P1与L1的差异性、P2与L2的差异性以及P3与L3的差异性正相关。

[0069] 比如,在一种可选的方案中,第二差异系数计算公式如下:

[0070] λ2=g(P1/L1)*g(P2/L2)*g(P3/L3)

[0071] 其中,g(.)是一个预定义的函数,若x>y,则g(x/y)=x/y,否则g(x/y)=y/x。

[0072] 可以理解的,第二差异系数的上述计算公式仅为示例,在不同的方案中也可能采用与之不同的公式,例如,在等式右侧再乘上某些系数,或者对等式右侧的三部分不是求乘积而是加权求和,或者,文本长度的差异性也未必要定义为比值,也可以定义为差值的绝对值,等等。

[0073] 还应当理解的是,以上是以抽取式阅读理解任务为例,若不是抽取式阅读理解任务,则可能语料中不存在文章、问题、答案等概念,此时计算文本长度的差异性也要采取另外的方式,例如计算语料中语句长度的差异性等。

[0074] 步骤C:根据第一差异系数以及第二差异系数确定第一数据配比。

[0075] 例如,第一数据配比可以定义为1:(λ1*λ2),当然也可以采取其他定义方式,例如1:(λ1+λ2)等,本申请并不限定。具体采用何种定义方式可以根据训练好的专用模型实际的性能确定。

[0076] 另外,也不排除在某些实现方式中只根据第一差异系数或第二差异系数确定第一数据配比。例如,对于阅读理解任务,存在被学界广泛认可的SQUAD数据集,但对于其他自然语言处理任务,则未必有这样定义良好的数据集,此时可以直接将第一差异系数作为第一数据配比。

[0077] 至于如何根据第一数据混合两类语料,本申请不限定,例如可以按照第一数据配比均匀地混合在一起,也可以按照第一数据配比随机混合,等等。

[0078] 在获得第一训练数据后,还可以对第一训练数据进行清洗以改善数据质量,当然,数据清洗步骤是可选的,例如,若第一训练数据的质量已经满足训练要求,则无须执行数据清洗。另外,数据清洗也可以在获得第一语料和第二语料后就进行,不一定要等到获得第一训练数据后才进行。数据清洗项目包括但不限于:

[0079] (1)将第一训练数据中的全角符号与半角符号转换为统一的全半角形式,如全部统一为半角符号,或者全部统一为全角符号;

[0080] (2)将第一训练数据中的标点符号转换为统一的标点形式。例如,中英文文本中同样的标点在符号上可能有区别(如中英文句号分别是圆圈和圆点),需要统一形式,避免计算机系统认为其属于不同含义的符号。

[0081] 在执行数据清洗时,可根据需求选择上述一项或多项执行。

[0082] 第一训练数据中样本的形式与目标任务有关,例如,若目标任务是抽取式阅读理解任务,则第一训练数据中的每个样本可以包括一篇文章、一个问题以及两个标签(用于标明答案在文章中的起止位置)。

[0083] 步骤S140:利用第一训练数据训练专用模型。

[0084] 在执行步骤S140之前,应当先基于步骤S110中获得的通用模型以及与目标任务相关的适配结构构建好专用模型。训了好的专用模型可用于执行目标任务,例如,若目标任务为抽取式阅读理解任务,则模型输入可以是文章和问题,输出可以是答案在文章中的起止位置。

[0085] 关于具体如何根据训练数据进行模型参数的更新,属于现有技术,此处不具体介绍。下面主要介绍在专用模型的训练过程中学习率参数的设置问题:

[0086] 本申请一些实现方式中,在利用第一训练数据训练专用模型的过程中,定期利用验证集评估专用模型的收敛程度,并根据收敛程度设置训练过程中使用的学习率。

[0087] 这里的定期可以是指指定的训练步数(采用一个批次的数据更新一次模型的参数定义为一步)或者指定的时长。模型的收敛程度可以根据模型的预测误差来定义:例如,可以直接定义为预测误差或者预测误差的映射值;又例如,可以预设一个基准误差作为模型在理想收敛状态下的误差,将模型当前计算得到的实际误差与该基准误差求差值,将该差值或者该差值的映射值作为模型的收敛程度,等等。

[0088] 其中,学习率被设置为与收敛程度负相关,也就是说在专用模型离收敛尚早时,学习率应当设置得大一些,加快模型的收敛速度,而在专用模型越接近于收敛,学习率应当设置得小一些,避免出现模型不稳定、过拟合等问题。

[0089] 需要指出,在上述实现方式中,学习率的调整是一种根据模型的收敛程度而进行的动态调整,即根据模型在验证集上的实际表现果来调整学习率,而不是每次调整固定的值,调整方式灵活、客观、有效,有利于改善训练得到的专用模型的质量。

[0090] 进一步的,上述学习率可以设置为Warm-up的衰减学习率,即在训练初期使用一个较小的学习率,待模型稳定后再选择预设的学习率进行学习,即所谓的学习率预热(Warm-up),此举有利于加快模型的收敛速度,改善模型的训练效果。在使用预设的学习学习一段时间后,开始进行学习率,避免在模型已经趋于收敛时因学习率过大产生不稳定、过拟合等问题。在学习率的衰减过程中,可以按照前述方式定期利用验证集评估专用模型的收敛程度,并根据收敛程度减小学习率的取值;其中,学习率取值的减小量与收敛程度负相关。

[0091] 前文提到,步骤S110中获取通用模型至少包括两种方式,下面再对其中的第二种方式进一步举例说明:

[0092] 在一些实现方式,通用模型可以按照如下步骤获得:

[0093] 步骤A:获取原始通用模型;其中,原始通用模型为预训练的、通用领域内的语言模型,例如,Google公司发布的BERT中文模型、百度公司发布的ERNIE模型等,在中文环境中可优先选择ERNIE模型。

[0094] 步骤B:获取第三语料以及第四语料;其中,第三语料为通用领域内的语料,第四语料为目标领域内的语料。关于第三语料和第四语料的获取可以参考步骤S120,但需要指出的是,由于通用模型是与具体任务无关的,所以第四语料直接选取目标领域内的数据就可以,不涉及与目标任务适配的问题。

[0095] 步骤C:基于第三语料与第四语料之间的差异性确定第二数据配比,并根据第二数据配比将第三语料与第四语料进行混合,获得第二训练数据。

[0096] 该步骤可以参照步骤S130实现,不再重复说明。不过需要指出,第二训练数据和第一训练数据在形式上还是有所区别的,因为第一训练数据针对的是下游的目标任务,而第

二训练数据是由上游的预训练任务所使用,例如,第二训练数据中的每个样本可以包括两个连续的语句。

[0097] 步骤D:利用第二训练数据训练原始通用模型,获得通用模型。

[0098] 该步骤可以参照步骤S140实现。或者,其学习率的调整可以采取简单一些的方式,比如每训练一轮(训练数据全都被使用过一次定义为一轮)后将衰减学习率为之前的70%,等等。

[0099] 上述实现方式通过对原始模型进行参数微调而获得通用模型,其所使用的第二训练数据类似于第一训练数据,也是由通用领域内的数据和目标领域内的数据构成的,并且在数据的配比上也与第一训练数据类似,从而通用模型可以兼顾领域性与通用性,其质量较高,有利于后续专用模型的训练。

[0100] 在上面提出的方案中,都会利用获得的通用模型构建专用模型,并进一步训练专用模型,最终获得可用于执行目标任务的模型。但也不排除在另一些方案中,利用获得的通用模型(例如,通过上面的步骤A至步骤D)构建专用模型后,直接将该专用模型投入使用,不再对专用模型进行调参(当然该专用模型调参后再投入使用也是可以的)。这些方案由于在训练产生通用模型的过程中,合理地混合了通用领域内的数据和目标领域内的数据,使得获得的通用模型在领域性和通用性之间取得了良好的平衡,从而基于该通用模型构建的专用模型的性能也得到改善。因此,在训练时间有限时,也不排除可以采用这样的做法。

[0101] 图2示出了本申请实施例提供的模型训练装置200的功能模块图。参照图2,模型训练装置200包括:

[0102] 第一模型获取模块210,用于获取通用模型,所述通用模型为预训练的、与任务无关的语言模型;

[0103] 第一数据获取模块220,用于获取第一语料以及第二语料;其中,所述第一语料为通用领域内的语料,所述第二语料为目标领域内与目标任务相关的语料,所述目标任务为自然语言处理任务,所述目标领域为所述目标任务所属的领域;

[0104] 第一数据混合模块230,用于基于所述第一语料与所述第二语料之间的差异性确定第一数据配比,并根据所述第一数据配比将所述第一语料与所述第二语料进行混合,获得第一训练数据;其中,所述第一语料与所述第二语料之间的差异性和所述第一数据配比负相关;

[0105] 第一训练模块240,用于利用所述第一训练数据训练用于执行所述目标任务的专用模型;其中,所述专用模型中包括所述通用模型以及与所述目标任务相关的适配结构。

[0106] 在模型训练装置200的一种实现方式中,第一数据混合模块230基于所述第一语料与所述第二语料之间的差异性确定数据配比,包括:获取第一差异系数,所述第一差异系数与所述目标领域内的关键词在所述目标领域内的测试语料中出现的频次正相关;根据所述第一语料中的文本长度与所述第二语料中的文本长度之间的差异性计算第二差异系数,所述第二差异系数与文本长度之间的差异性正相关;根据所述第一差异系数以及所述第二差异系数确定所述第一数据配比。

[0107] 在模型训练装置200的一种实现方式中,所述目标任务为抽取式阅读理解任务,第一数据混合模块230根据所述第一语料中的文本长度与所述第二语料中的文本长度之间的差异性计算第二差异系数,包括:计算所述第一语料中阅读理解的文章平均长度L1、问题平

均长度L2以及答案平均长度L3;计算所述第二语料中阅读理解的文章平均长度P1、问题平均长度P2以及答案平均长度P3;根据P1与L1的差异性、P2与L2的差异性以及P3与L3的差异性计算所述第二差异系数;其中,所述第二差异系数分别与所述P1与L1的差异性、所述P2与L2的差异性以及所述P3与L3的差异性正相关。

[0108] 在模型训练装置200的一种实现方式中,第一训练模块240利用所述第一训练数据训练用于执行所述目标任务的专用模型,包括:在利用所述第一训练数据训练所述专用模型的过程中,定期利用验证集评估所述专用模型的收敛程度,并根据所述收敛程度设置训练过程中使用的学习率;其中,所述学习率被设置为与所述收敛程度负相关。

[0109] 在模型训练装置200的一种实现方式中,所述学习率为带Warm-up的衰减学习率,第一训练模块240在利用所述第一训练数据训练所述专用模型的过程中,定期利用验证集评估所述专用模型的收敛程度,并根据所述收敛程度设置训练过程中使用的学习率,包括:在利用所述第一训练数据训练所述专用模型的过程中的学习率衰减阶段,定期利用验证集评估所述专用模型的收敛程度,并根据所述收敛程度减小所述学习率的取值;其中,所述学习率取值的减小量与所述收敛程度负相关。

[0110] 在模型训练装置200的一种实现方式中,所述目标任务为抽取式阅读理解任务,所述目标任务的答案满足第一统计规律,第一数据获取模块220获取第二语料,包括:根据所述目标领域搜索和/或构造用于抽取式阅读理解的语料,并从中筛选出答案满足所述第一统计规律的语料作为所述第二语料。

[0111] 在模型训练装置200的一种实现方式中,第一模型获取模块210获取通用模型,包括:获取原始通用模型,所述原始通用模型为预训练的、所述通用领域内的语言模型;获取第三语料以及第四语料;其中,所述第三语料为所述通用领域内的语料,所述第四语料为所述目标领域内的语料;基于所述第三语料与所述第四语料之间的差异性确定第二数据配比,并根据所述第二数据配比将所述第三语料与所述第四语料进行混合,获得第二训练数据;其中,所述第三语料与所述第四语料之间的差异性和所述第二数据配比负相关;利用所述第二训练数据训练所述原始通用模型,获得所述通用模型。

[0112] 本申请实施例提供的模型训练装置200,其实现原理及产生的技术效果在前述方法实施例中已经介绍,为简要描述,装置实施例部分未提及之处,可参考方法实施例中相应内容。

[0113] 图3示出了本申请实施例提供的模型训练装置300的功能模块图。参照图3,模型训练装置300包括:

[0114] 第二模型获取模块310,获取原始通用模型,所述原始通用模型为预训练的、通用领域内的语言模型;

[0115] 第二数据获取模块320,获取第三语料以及第四语料;其中,所述第三语料为所述通用领域内的语料,所述第四语料为目标任务所属的目标领域内的语料,所述目标任务为自然语言处理任务;

[0116] 第二数据混合模块330,基于所述第三语料与所述第四语料之间的差异性确定第二数据配比,并根据所述第二数据配比将所述第三语料与所述第四语料进行混合,获得第二训练数据;其中,所述第三语料与所述第四语料之间的差异性和所述第二数据配比负相关;

[0117] 第二训练模块340,利用所述第二训练数据训练所述原始通用模型,获得通用模型;其中,所述通用模型用于与所述目标任务相关的适配结构组成专用模型,所述专用模型为用于执行所述目标任务的模型。

[0118] 本申请实施例提供的模型训练装置300,其实现原理及产生的技术效果在前述方法实施例中已经介绍,为简要描述,装置实施例部分未提及之处,可参考方法施例中相应内容。

[0119] 图4示出了本申请实施例提供的电子设备400的一种可能的结构。参照图4,电子设备400包括:处理器410、存储器420以及通信接口430,这些组件通过通信总线440和/或其他形式的连接机构(未示出)互连并相互通讯。

[0120] 其中,存储器420包括一个或多个(图中仅示出一个),其可以是,但不限于,随机存取存储器(Random Access Memory,简称RAM),只读存储器(Read Only Memory,简称ROM),可编程只读存储器(Programmable Read-Only Memory,简称PROM),可擦除可编程只读存储器(Erasable Programmable Read-Only Memory,简称EPROM),电可擦除可编程只读存储器(Electric Erasable Programmable Read-Only Memory,简称EEPROM)等。处理器410以及其他可能的组件可对存储器420进行访问,读和/或写其中的数据。

[0121] 处理器410包括一个或多个(图中仅示出一个),其可以是一种集成电路芯片,具有信号的处理能力。上述的处理器410可以是通用处理器,包括中央处理器(Central Processing Unit,简称CPU)、微控制单元(Micro Controller Unit,简称MCU)、网络处理器(Network Processor,简称NP)或者其他常规处理器;还可以是专用处理器,包括图形处理器(Graphics Processing Unit,GPU)、数字信号处理器(Digital Signal Processor,简称DSP)、专用集成电路(Application Specific Integrated Circuits,简称ASIC)、现场可编程门阵列(Field Programmable Gate Array,简称FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。并且,在处理器410为多个时,其中的一部分可以是通用处理器,另一部分可以是专用处理器。

[0122] 通信接口430包括一个或多个(图中仅示出一个),可以用于和其他设备进行直接或间接地通信,以便进行数据的交互。通信接口430可以包括进行有线和/或无线通信的接口。

[0123] 在存储器420中可以存储一个或多个计算机程序指令,处理器410可以读取并运行这些计算机程序指令,以实现本申请实施例提供的模型训练方法以及其他期望的功能。

[0124] 可以理解,图4所示的结构仅为示意,电子设备400还可以包括比图4中所示更多或者更少的组件,或者具有与图4所示不同的配置。图4中所示的各组件可以采用硬件、软件或其组合实现。电子设备400可能是实体设备,例如PC机、笔记本电脑、平板电脑、手机、服务器、嵌入式设备等,也可能是虚拟设备,例如虚拟机、虚拟化容器等。并且,电子设备400也不限于单台设备,也可以是多台设备的组合或者大量设备构成的集群。

[0125] 本申请实施例还提供一种计算机可读存储介质,该计算机可读存储介质上存储有计算机程序指令,所述计算机程序指令被计算机的处理器读取并运行时,执行本申请实施例提供的模型训练方法。例如,计算机可读存储介质可以实现为图4中电子设备400中的存储器420。

[0126] 在本申请所提供的实施例中,应该理解到,所揭露装置和方法,可以通过其它的方

式实现。以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，又例如，多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些通信接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

[0127] 另外，作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

[0128] 再者，在本申请各个实施例中的各功能模块可以集成在一起形成一个独立的部分，也可以是各个模块单独存在，也可以两个或两个以上模块集成形成一个独立的部分。

[0129] 以上所述仅为本申请的实施例而已，并不用于限制本申请的保护范围，对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

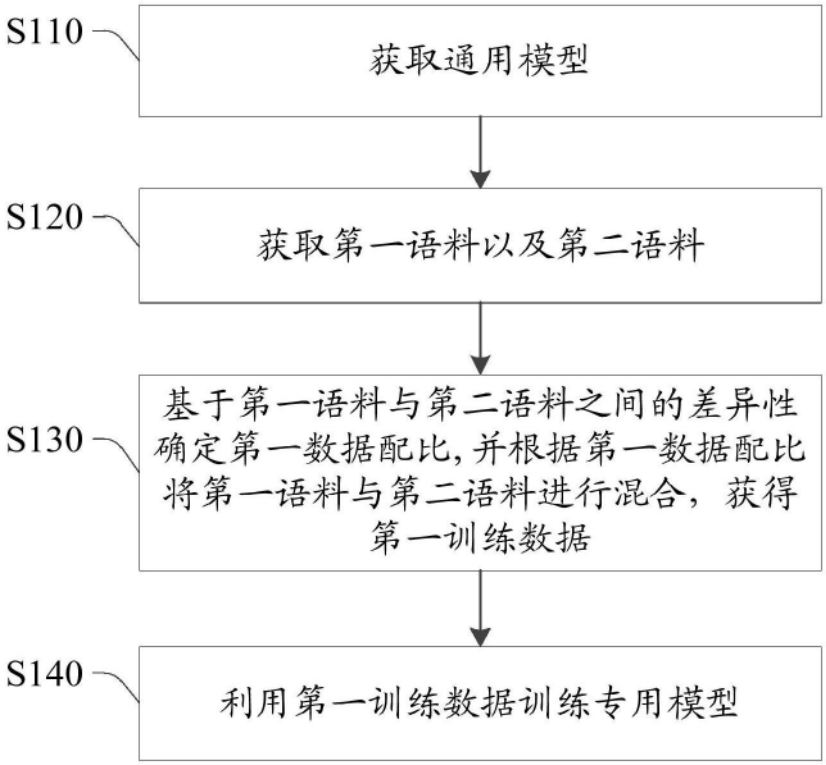

S110
获取通用模型
S120
获取第一语料以及第二语料
S130
基于第一语料与第二语料之间的差异性确定第一数据配比，并根据第一数据配比将第一语料与第二语料进行混合，获得第一训练数据
S140
利用第一训练数据训练专用模型

图1

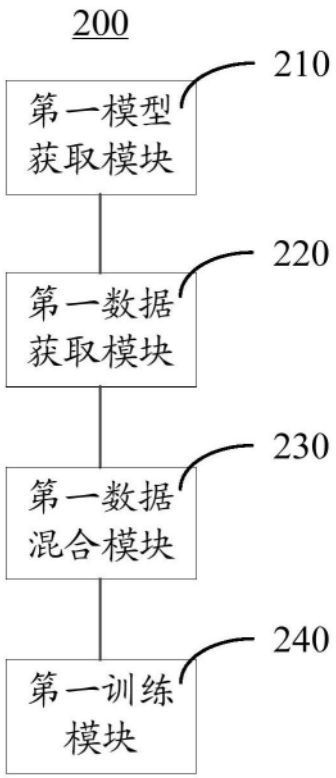

200
210
第一模型获取模块
220
第一数据获取模块
230
第一数据混合模块
240
第一训练模块

图2

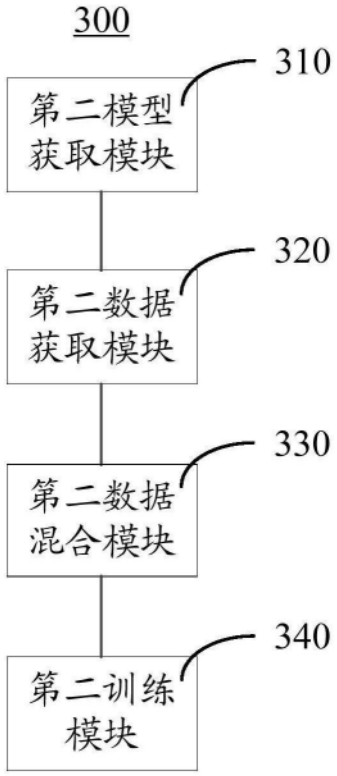

300
310
第二模型获取模块
320
第二数据获取模块
330
第二数据混合模块
340
第二训练模块

图3

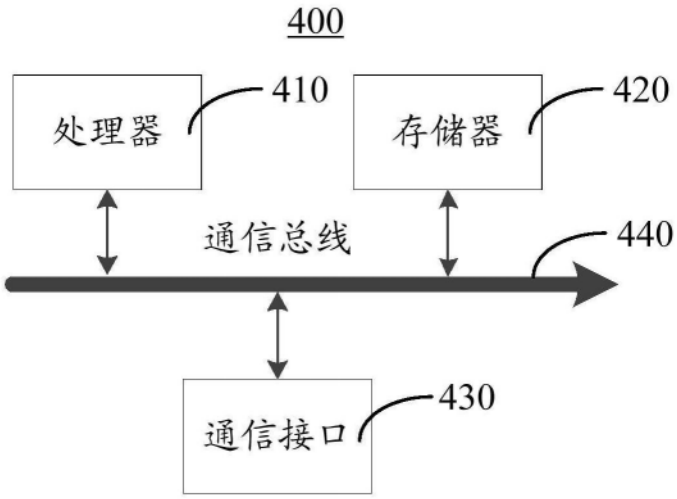

400
处理器
410
存储器
420
通信总线
440
通信接口
430

图4