

## 一种基于跨语言数据增强的分词方法及装置

本申请要求在2021年01月11日提交中国专利局,申请号为202110034450.2的中国专利申请的优先权,该申请的全部内容通过引用结合在本申请中。

### 技术领域

本申请实施例涉及分词技术领域,尤其涉及一种基于跨语言数据增强的分词方法、一种基于跨语言数据增强的分词装置、基于跨语言数据增强的分词设备及存储介质。

### 背景技术

由于信息化进程的发展,对用户搜索和推荐服务是目前信息时代的普遍需求。在为用户提供搜索和推荐服务的过程中,通常需要对句子进行合理的分词。传统的分词任务中,对于英文这类天然带有空格的句子,只需要按照空格分词即可。然而与传统的分词任务的假设不同的是,现实中的用户并不会严格按照语法进行分词,而是常常把若干个词连在一起输入。错误的分词会影响到下游的任务的执行,比如:实体识别、语义识别等。因此,需要根据业务场景训练特殊的分词模型。而分词模型的训练需要大量的语料,但是对于一些数据相对稀缺的国家和地区,由于业务处于较为早期的阶段,缺少足够的用户数据,也缺乏相应的标注资源,语料资源的获取变得尤为艰难。

相关的分词技术方案主要分为两类,一类是基于概率统计的词典法,另一类是基于神经网络的模型法。词典法的主要逻辑是收集足够多的词和这些词的词频,通过计算不同的分词组合的概率得到最终的分词结果。模型法的主要逻辑则是采用序列标注的方法,通过特征的转移概率计算出全局最优的序列组合,并将这个序列转换为分词结果。这两种方法都需要足够多的训练语料,但是对低资源地区的训练语料的获取较为困难。一种涉及低资源地区的较为流行的解决方案是借助谷歌提供的多语言BERT模型进行分词处理。其首先下载在大规模数据集上预训练好的模型,再在每个LRL上用小数据集进行精调。BERT模型会将文字转换成向量,再将这些向量作为特征输入后续的模型中进行预测。

但是谷歌提供的 BERT 预训练模型是已经分好词的，向量也是基于词的向量，因此不能被运用于分词任务中。而且 BERT 预训练模型是在正式语料（如新闻、博客文章等）中进行训练的，缺少在一个产品的社交生态中产生的独特的语境。

## 发明内容

本申请实施例提供了一种基于跨语言数据增强的分词方法、装置、设备及存储介质，以实现利用高资源语言地区的语料作为低资源语言地区的扩充和验证以对低资源语言进行分词，打破语言资源不平衡的问题。

第一方面，本申请实施例提供了一种基于跨语言数据增强的分词方法，包括：

采集若干组高资源语言数据，根据每一组所述高资源语言数据处理得到一组第一分词语料；若干组第一分词语料构成第一分词语料库；

采集若干低资源语言数据，根据低资源语言数据处理得到若干候选分词，基于每一个候选分词与第一分词语料库之间的匹配度从若干候选分词中选取第二分词语料；

基于第二分词语料训练得到分词模型，将若干待分词数据输入至分词模型中以输出多个分词候选结果；

基于每一个分词候选结果与第一分词语料库之间的匹配度，选取匹配度最高的分词候选结果作为分词结果。

第二方面，本申请实施例提供了一种基于跨语言数据增强的分词装置，包括：

第一采集模块，配置为采集若干组高资源语言数据，根据每一组所述高资源语言数据处理得到一组第一分词语料；若干组第一分词语料构成第一分词语料库；

第二采集模块，配置为采集若干低资源语言数据，根据低资源语言数据处理得到若干候选分词，基于每一个候选分词与第一分词语料库之间的匹配度从若干候选分词中选取第二分词语料；

分词候选模块，配置为基于第二分词语料训练得到分词模型，将若干待分词数据输入至分词模型中以输出多个分词候选结果；

分词确定模块，配置为基于每一个分词候选结果与第一分词语料库之间的匹配度，选取匹配度最高的分词候选结果作为分词结果。

第三方面，本申请实施例提供了一种基于跨语言数据增强的分词设备，包括：存储器以及一个或多个处理器；

所述存储器，配置为存储一个或多个程序；

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如第一方面所述的一种基于跨语言数据增强的分词方法。

第四方面，本申请实施例提供了一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行时配置为执行如第一方面所述的一种基于跨语言数据增强的分词方法。

第五方面，本申请实施例还提供一种基于跨语言数据增强的分词处理的程序，该程序被执行时，可以实现如第一方面所述的基于跨语言数据增强的分词方法有关的操作。

本申请实施例首先采集高资源语言数据进行处理得到分词语料，之后采集低资源语言数据获取候选分词，并根据从高资源语言数据处获得的分词语料对候选分词进行甄选，选择出与分词语料匹配度高的作为低资源语言数据的分词语料，并根据该低资源分词语料进行分词模型的训练，从而可以实现对低资源语言数据基于模型自动输出分词候选结果，并结合分词候选结果与高资源语言数据的分词语料的匹配度进行选取分词结果，通过使用高资源语言的语料对低资源语言的模型训练数据进行自动扩充和验证，解决了数据资源和标注资源在不同语言之间不平衡的问题，为跨国产品快速在低资源语言中迭代分词模型提供了一种更轻量而高效的解决方式，并且由于低资源语言的训练语料通过自动化方式生成，因此后续更新时只需要在高资源语言中进行重点维护，节省了维护成本。

## 附图说明

图1是本申请实施例提供的一种基于跨语言数据增强的分词方法的流程图；

图2是本申请实施例提供的另一种基于跨语言数据增强的分词方法的流程图；

图3是本申请实施例提供的另一种基于跨语言数据增强的分词方法的流程图；

图4是本申请实施例提供的另一种基于跨语言数据增强的分词方法的流程图；

图5是本申请实施例提供的另一种基于跨语言数据增强的分词方法的流程图；

图6是本申请实施例提供的一种基于跨语言数据增强的分词装置的结构示意图；

图7是本申请实施例提供的一种基于跨语言数据增强的分词设备的结构示意图。

### 具体实施方式

为了使本申请的目的、技术方案和优点更加清楚，下面结合附图对本申请具体实施例作进一步的详细描述。可以理解的是，此处所描述的具体实施例仅用于解释本申请，而非对本申请的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与本申请相关的部分而非全部内容。在更加详细地讨论示例性实施例之前应当提到的是，一些示例性实施例被描述成作为流程图描绘的处理或方法。虽然流程图将各项操作（或步骤）描述成顺序的处理，但是其中的许多操作可以被并行地、并发地或者同时实施。此外，各项操作的顺序可以被重新安排。当其操作完成时所述处理可以被终止，但是还可以具有未包括在附图中的附加步骤。所述处理可以对应于方法、函数、规程、子例程、子程序等等。

本申请实施例提供了一种基于跨语言数据增强的分词方法、一种基于跨语言数据增强的分词装置、基于跨语言数据增强的分词设备及存储介质。本申请实施例首先采集高资源语言数据进行处理得到分词语料，之后采集低资源语言数据并获取候选分词，根据从高资源语言数据处获得的分词语料对候选分词进行甄选，选择出与分词语料匹配度高的作为低资源语言数据的分词语料，并根据该低资源分词语料进行分词模型的训练，从而可以实现对低资源语言数据基于模型自动输出分词候选结果，并结合分词候选结果与高资源语言数据的分词语料的匹配度进行选取分词结果。

本申请实施例使用机器学习的技术自动地生成低资源地区（主要为小语种或者方言使用地区）的分词训练数据和分词模型。具体使用方式是：首先在高

资源语言（如英语）中利用用户的自然输入数据生成一批该语言的分词语料，并进行半人工的验证。接着在低资源地区利用本方案自动化地生成一批分词语料，利用高资源语言的语料进行自动化验证。最后，结合全域的高频词库进行低资源语言中的分词模型训练。实施例通过使用高资源语言的语料对低资源语言的模型训练数据进行自动扩充和验证，解决了数据资源和标注资源在不同语言之间不平衡的问题，为跨国产品快速在低资源语言中迭代分词模型提供了一种更轻量而高效的解决方式，并且由于低资源语言的训练语料通过自动化方式生成，因此后续更新时只需要在高资源语言中进行重点维护，节省了维护成本。

下面分别进行详细说明。

图1给出了本申请实施例提供的一种基于跨语言数据增强的分词方法的流程图，本申请实施例提供的一种基于跨语言数据增强的分词方法可以由一种基于跨语言数据增强的分词装置来执行，该基于跨语言数据增强的分词装置可以通过硬件和/或软件的方式实现，并集成在计算机设备中。

下述以基于跨语言数据增强的分词装置执行基于跨语言数据增强的分词方法为例进行描述。参考图1，该基于跨语言数据增强的分词方法包括：

101：采集若干组高资源语言数据，根据每一组所述高资源语言数据处理得到一组第一分词语料；若干组第一分词语料构成第一分词语料库。

在本申请实施例中，高资源语言是指有较多语料数据和标注数据的语言，通俗的说是更为广泛使用的语言，例如英语、中文等。而低资源语言则与高资源语言相对而言，表示语料和标注数据都较为缺乏的语言，常见于小语种，例如西班牙语、德语、法语、越南语、泰语等；或者各类方言，例如粤语、客家话、潮汕语等。

本申请的应用基于对同一产品的不同的语言的数据具有相似的数据分布这个假设。在现实应用场景中，这个假设在许多跨国产品的使用中是成立的，因为不同国家的用户使用同一套的底层服务，例如输入法、推文话题等，并且有相似的语言生态，例如一些热门词汇、习惯表达等。

采集高资源语言数据，也即是收集高资源语言地区的天然的用户输入数据，例如用户的搜索数据、用户的评论数据。本申请实施例中对高资源语言数据的采集的场景可以是，在某个高资源语言国家应用的一款视频客户端，采集该视频客户端的注册用户对用户通过该视频客户端进行视频播放时所进行的视频本身的评论数据，或者用户使用该视频客户端时对视频进行搜索时的搜索数据。

容易理解的是，评论数据通常为字段，在一实施例中，字段指包含众多由字或者单词组成的句子、或者就是本身由字或者单词组成的文本。当然评论数据还可能是表情符等其他语言单位的体现。而搜索数据往往包含的语言单位量少，通常以关键词进行体现，包括人名、时间等数据表达，一般是短语或者短句。在其他的一些示例性中，高资源语言数据的采集场景可以是短视频本身所涉及的语音数据，将语音转变成的整段文本作为高资源语言数据，还可以是论文网站，获取公开的论文中的语言数据作为高资源语言数据。更多的示例性中，高资源语言数据可以是各种能够容易获得的语言数据。

本申请实施例对这批高资源语言数据进行清洗和转换，生成多组短句。其中每一组短句包含有若干个相同语言单位数的短语。例如将高资源语言数据“happy birthday to you”生成2个语言单位的短句，则为 happy birthday, birthday to, to you；在该短句中，happy birthday, birthday to, to you 是在该短句集合中的三个短语。若是将“happy birthday to you”生成3个语言单位的短句，则为 happy birthday to, birthday to you。若是将“happy birthday to you”生成1个语言单位的短句，则为 happy, birthday, to, you, happy, birthday, to, you 分别作为短句“happy, birthday, to, you”的短语。本申请实施例中，语言单位表示不同语种、不同语言场景等情况下，对应的一个词或者一个字。例如对于中文而言，一句话包含若干个字，而每一个字都作为一个语言单位独立存在。或者如英文而言，一句话包含若干个单词，每一个单词都作为一个语言单位独立存在。

所采集的高资源语言数据，通常包括字段或句子，其中一个段落或者一个句子均可作为一组高资源语言数据，对每一组高资源语言数据进行处理分别得到第一分词语料。例如以将英文的国家为例，由于英文被众多国家作为母语使用，并且英语是国际化通用语言，所采集的英语国家的语言数据则为高资源语言数据。采集的高资源语言数据例如是“happy birthday to you”，为一个句子，还可以是段落，如“Today is a good day. I prepared a delicious dinner”。其中，“happy birthday to you”和“Today is a good day. I prepared a delicious dinner”作为不同组的高资源语言数据，可以分别处理得到第一分词语料。本申请中，没有固定的规则对高资源语言数据分组，本实施例描述的多组高资源语言数据是表示每一个句子、或者多个句子组成的段落均可作为独立的一组高资源语言数据进行处理为第一分词语料的操作。

在另外一个实施例中，根据每一组所述高资源语言数据处理得到一组第一分词语料的方式如图2中的1011至1013所示。其中，包括：1011、采集若干组高资源语言数据，分别对每一组高资源语言数据进行预处理，将预处理后的每一组高资源语言数据进行数据转换以形成多组短句；其中，每一组短句包含若干个具有相同语言单位数的短语；1012、将所包含的语言单位完全相同的短语划分为同一个短语集合；其中，所述短语集合中包含若干种短语；1013、分别获取每一种短语在为同一个短语集合中的出现频率，根据出现频率从高至低的排序选取排名前第一阈值的短语为第一分词语料。

在本实施例中，第一阈值为自然整数，例如1,2,3。作为本实施例中的一种方案，第一阈值可以选择为3。

在一实施例中，预处理包括按照第一预设规则对高资源语言数据中的标点符号进行转换、按照第二预设规则对高资源语言数据中的表情符号转换为语言单位。预处理采用统一的转换规则对高资源语言数据中的一些特殊语言单位进行转换，包括标点符合、表情符合等。例如从评价数据中获得高资源语言数据，而评价数据携带有表情符合，则可以通过识别该表情符号的含义，将该表情符合转换为与含义对应的语言单位。例如“开心”对应的表情符为一个笑脸表情，则当评价数据中出现该笑脸表情，将笑脸表情转换为语言单位“开心”。则获取表情符号的含义，根据将表情符合转换为与含义对应的语言单位为本申请实施例所述的第二预设规则。其他示例性中，第二预设规则还可以是提前设置表情符与语言单位的映射关系，通过该表情符可以更为直接的获取到对应的语言单位。在更多的示例性中，甚至预处理方式对表情符号的处理不一定是转换，可能直接将表情符号删除。同样，对于第一预设规则中对标点符合进行转换，例如是一个完整的句子结束后用句号表达，而中文和英文对句号的表达方式不同，第一预设规则将使标点符合有统一的表达。

在一实施例中，假设预处理后的一组高资源语言数据为“今天给朋友过生日，祝生日快乐。”基于该高资源语言数据，可以转换为多组短句，每一组短句分别所对应的连续语言单位数不相等。假设分为连续语言单位数为2的短句，则形成“今天、天给、给朋、朋友、友过、过生、生日、日祝、祝生、生日、日快、快乐”的一组短句。同样还可以形成连续语言单位数为3的短句，为“今天给、天给朋、给朋友、朋友过、友过生、过生日”的一组短句，每一组短句中，包括若干个短语，例如“今天给”、“给朋友”则是本实施例所谓的短语。

可以理解到每一组短句中的短语的连续语言单位数是相同的。在上述短语中，将语言单位也完全相同的短句划分为一组短语集合，例如上述出现了“生日”是完全相同的，可以划分为一组。需要说明的是，本申请分组的规则定义语言单位完全一致，包括语言单位的顺序也必须一致，但是不考究语言单位与语言单位之间的空格。例如短句为我（空格）喜欢你、我喜欢（空格）你，我（空格）喜欢（空格）你，以上三个短句，去除掉空格，实际都是“我喜欢你”，那么以语言单位串“我喜欢你”为关键词进行分组为同一个短语集合。短语聚合中包含若干种短语，例如“我喜欢你”作为关键词，所有去掉空格后都能形成该关键词的短语共同组成的短语集合，由于短语的空格设置位置不同，因此导致每一种短语虽然关键词相同，但是实际不相同，也即是对应不同种的短语，例如我（空格）喜欢你和我喜欢（空格）你则是两种不同的短语。假设有一组短语集合，包括的短语为：我（空格）喜欢你、我（空格）喜欢你、我喜欢（空格）你、我（空格）喜欢（空格）你，可见我（空格）喜欢你出现频率为2次，相比其他两种短语出现频率更高。本实施例，则出现频率从高至低的排序选取排名前第一阈值的短语为第一分词语料。第一阈值在一个实施方式中，可以选用3，但是在其他示例性中，也可以是其他数值，如5，如4，等等，本申请并不作限定。又如，有一组短语集合，包括的短语为：我（空格）喜欢你、我（空格）喜欢你、我喜欢（空格）你、我（空格）喜欢（空格）你，我（空格）喜欢（空格）你。假设第一阈值=2，则挑选我（空格）喜欢你和我（空格）喜欢（空格）你作为一组第一分词语料。由于采集的高资源语言数据足够庞大，因此可以形成很多的第一分词语料，足够数量的第一分词语料相当于构成了语料库，也即是本申请所述的第一分词语料库。

在另外一个实施方式中，根据出现频率从高至低的排序选取排名前第一阈值的短语为第一分词语料，包括：根据出现频率从高至低的排序选取排名前第一阈值的短语为候选第一分词语料；从同一个短语集合中任意挑选一个短语，将短语中的全部空格去除后形成关键字；输入关键字至预设分词模型以输出预测分词语料；比对候选第一分词语料和预测分词语料，当第一分词语料与预测分词语料一致时，将候选第一分词语料作为第一分词语料；当第一分词语料与预测分词语料不一致时，对候选第一分词语料进行调整，将调整后的候选第一分词语料作为第一分词语料。

第一阈值作为一个预先设定的数值，为自然数。例如第一阈值为2,3,5，等

等，本申请不作限定。上述操作中提供给第一分词语料的半自动验证方式。也即是，根据出现频率从高至低的排序选出排名前第一阈值的短语并不直接作为第一分词语料归入第一分词语料库中，而是作为候选第一分词语料，并对候选第一分词语料进行验证。验证方式包括从同一个短语集合中任意挑选一个短语，将短语中的全部空格去除后形成关键字，例如 ABCD 这连续四个语言单位对应的短语集合中有短语 A（空格）BC（空格）D、A（空格）BCD，随意挑选 A（空格）BC（空格）D 的短语，去除空格后形成关键字 ABCD。将关键字 ABCD 输入预设分词模型中。该预设分词模型是指目前相关技术中已经公开的、被使用到的分词模型。预设分词模型对关键字 ABCD 输出若干预测分词语料，例如输出 A（空格）BC（空格）D、A（空格）BCD、AB（空格）CD，比对候选第一分词语料与预测分词语料，也即是将“A（空格）BC（空格）D、A（空格）BCD”与“A（空格）BC（空格）D、A（空格）BCD、AB（空格）CD”进行比对，则 A（空格）BC（空格）D、A（空格）BCD 均有一只对应的预测分词语料，将该 A（空格）BC（空格）D、A（空格）BCD 作为第一分词语料。在另外一种情况中，假设候选第一分词语料为“A（空格）BC（空格）D、A（空格）BCD、AB（空格）CD”，而预测分词语料为“A（空格）BC（空格）D、A（空格）BCD”，可以知道候选第一分词语料中的 AB（空格）CD 在预测分词语料中没有对应，则该 AB（空格）CD 不作为第一分词语料，另外的 A（空格）BC（空格）D、A（空格）BCD 则作为第一分词语料。

102：采集若干低资源语言数据，根据低资源语言数据处理得到若干候选分词，基于每一个候选分词与第一分词语料库之间的匹配度从若干候选分词中选取第二分词语料。

低资源语言数据是相对于高资源语言数据而言和存在的。如前面所介绍的，高资源语言是指有较多语料数据和标注数据的语言，通俗的说是更为广泛使用的语言，而低资源语言则是较少被应用、受众群体较少、语料较为缺乏的语言。

采集低资源语言数据同样可以是收集低资源语言地区的用户的搜索数据、用户的评论数据。对低资源语言数据的采集的应用场景可以与高资源语言数据的采集场景完全相同。例如，在某个低资源语言国家应用的一款视频客户端，采集该视频客户端的注册用户对用户通过该视频客户端进行视频播放时所进行的视频本身的评论数据，以及或者用户使用该视频客户端时对视频进行搜索时的搜索数据。容易理解的是，评论数据通常为字段，在实施例的字段指包含众

多由字或者单词组成的句子、或者就是本身由字或者单词组成的文本。当然评论数据还可能是表情符等其他语言单位的体现。而搜索数据往往包含的语言单位量少，通常以关键词进行体现，包括人名、时间等数据表达，一般是短语或者短句。在其他的一些示例性中，低资源语言数据的采集场景可以是短视频本身所涉及的语音数据，将语音转变成的整段文本作为低资源语言数据。更多的示例性中，低资源语言数据可以是各种能够容易获得的语言数据，例如当地的新闻联播的字幕。

本申请实施例对采集的低资源语言数据同样可进行清洗和转换，实质上清洗和转换则是对数据的预处理。对预处理后的低资源语言数据处理得到候选分词。候选分词在本申请实施例中旨在用于后续基于从候选分词中选择第二分词语料，即低资源语言数据的分词语料。

在一个实施例中，作为对步骤102的进一步细化，如图3中步骤1021至1023所示，采集若干低资源语言数据，根据低资源语言数据处理得到若干候选分词，可包括：1021、采集若干低资源语言数据，分别对每一组低资源语言数据进行预处理，将预处理后的每一组低资源语言数据进行数据转换以形成多组短句，其中，每一组短句包含若干个具有相同语言单位数的短语；1022、将所包含的语言单位完全相同的短语划分为同一个短语集合，其中，所述短语集合中包含若干种短语；1023、分别获取每一种短语在同一个短语集合中的出现频率，根据出现频率从高至低的排序选取排名前第二阈值的短语为候选分词。

本申请实施例中，对低资源语言数据进行预处理的方式可以与低高资源语言数据的方式相同，即按照第一预设规则对低资源语言数据中的标点符号进行转换、按照第二预设规则对低资源语言数据中的表情符号转换为语言单位。预处理将收集的低资源语言数据进行特殊语言单位的转换，与高资源语言数据相同的，例如从评价数据中获得低资源语言数据，而评价数据携带有表情符合，则可以通过识别该表情符号的含义，将该表情符合转换为与含义对应的语言单位。例如“开心”对应的表情符为一个笑脸表情，则当评价数据中出现该笑脸表情，将笑脸表情转换为语言单位“开心”。则获取表情符号的含义，根据将表情符合转换为与含义对应的语言单位为本申请实施例所述的第二预设规则。其他示例性中，第二预设规则还可以是提前设置表情符与语言单位的映射关系，通过该表情符可以更为直接的获取到对应的语言单位。在更多的示例性中，甚至预处理方式对表情符号的处理不一定是转换，可能直接将表情符号删除。同

样，对于第一预设规则中对标点符合进行转换，例如是一个完整的句子结束后用句号表达，而中文和英文对句号的表达方式不同，第一预设规则将使标点符合有统一的表达。

本实施例对与预处理后的低资源语言数据进行数据转换形成短句，形成短句的方式与高资源语言数据形成短句的方式相同。每一组短句包含若干个具有相同语言单位数的短语。

示例性的，低资源语言数据为“ABCDEFADDEGCABCD”，其中每一个英文字母代表一个语言单位。语言单位在本示例性中的含义为有实质意义的字，不包含符号、空格等。将“ABCDEFADDEGCABCD”转换为多组短句，意思是可以将“ABCDEFADDEGCABCD”拆解为连续语言单位数为4的短句，也可以是连续语言单位数为3的短句，还可能是其他连续语言单位数的短句。在此以将“ABCDEFADDEGCABCD”拆解为连续语言单位数为4的短句，形成“ABCD、BCDE、CDEF、DEFA、EFAD、FADE、ADEG、DEGC、EGCA、GCAB、CABC、ABCD”的这样一组短句，可见该组短句，包含有12组短语，每一组短语的连续语言单位数都是4。为了方便理解，上面的低资源语言数据的表达实际上已经去除了空格，仅仅将语言单位进行体现。但是实际上，任何两个语言单位之间都可能包含有空格。基于上述一组短语，可能实际存在的情况是，AB（空格）CDE（空格）FAD（空格）EGCABC（空格）D。那么所形成的连续语言单位数为4的一组短句为“AB（空格）CD、B（空格）CDE、CDE（空格）F、DE（空格）FA、EFAD、FAD（空格）E、AD（空格）EG、DEGC、EGCA、GCAB、CABC、ABC（空格）D”。

本实施例的做法，实际上将所包含的语言单位完全相同的短语划分为同一个短语集合，所述短语集合中包含若干种短语。以上述举例为例，也技术“AB（空格）CD、B（空格）CDE、CDE（空格）F、DE（空格）FA、EFAD、FAD（空格）E、AD（空格）EG、DEGC、EGCA、GCAB、CABC、ABC（空格）D”中的短语进行分组。例如AB（空格）CD去掉空格后的实质语言单位组成的短句为ABCD，与B（空格）CDE去掉空格后实质语言单位组成的短句BCDE相比，很显然语言单位不完全相同。容易发现，上述短句的组合中，AB（空格）CD与ABC（空格）D的语言单位实质上是完全相同的，因此对着两个短句进行分组，划分在同一个短语集合。在该短语集合中，包含AB（空格）CD与ABC（空格）D这两种短语。

分别获取每一种短语在为同一个短语集合中的出现频率，根据出现频率从高至低的排序选取排名前第二阈值的短语为候选分词。例如在上述 AB（空格）CD 所在的短语集合中，由于只有两种短语，且两种短语的数量均为 1 个，也即是，AB（空格）CD 与 ABC（空格）D 在该短语集合中出现的频率相同。在其他的示例性中，例如一组短语集合包括有短语为：AB（空格）CD、ABC（空格）D、ABC（空格）D、ABC（空格）D、AB（空格）CD、A（空格）BCD、A（空格）BC（空格）D。在上述短语集合中，总共有 7 个短语，并形成 4 种短语。其中 AB（空格）CD 相同的短语有 2 个，ABC（空格）D 相同的短语有 3 个，A（空格）BCD 相同的短语有 1 个，A（空格）BC（空格）D 相同的短语有 1 个。因此 ABC（空格）D 在短语集合中的出现频率最高，为七分之三，出现频率第二的是 AB（空格）CD，出现频率为七分之二。由于按照出现频率从高至低的顺序选取排名前第二阈值的短语为候选分词，在此第二阈值可以与第一阈值的数值相同，也可以不相同，同样第二阈值为自然数，假设第二阈值为 2，则在上述举例中，选择 ABC（空格）D 和 AB（空格）CD 作为候选分词。

上述根据低资源语言数据处理得到若干候选分词与根据高资源语言数据处理得到第一分词语料的原理实质相同。在另外一个实施中，同样是对步骤 102 的进一步细化，如图 4 所示，将图 3 中的步骤 1022 和 1023 替换为 1024 和 1025 即：1024、将任意一个短语去除短语中的空格后形成关键字；1025、采用动态规划编程方法在关键字中加入空格，以形成所有可能的分词作为候选分词。

本实施方式与上一种实施方式相同的是，均需要先将预处理后的每一组低资源语言数据进行数据转换以形成多组短句，并且将任意一个短语去除短语中的空格后形成关键字。上一个实施例方式中，没有写明将任意一个短语去除短语中的空格后形成关键字，但实质执行了该操作，体现在对同一组短句中的短语分组上，实质上就是以短语去掉空格后形成关键字，关键字相同的为一组。而本实施方式中，虽然形成关键字，但是并不基于关键字对短语进行分组。而是形成关键字之后，采用动态规划编程方法在关键字中加入空格，以形成所有可能的分词作为候选分词。动态规划编程方法是在编程中较常运用到的编程基础思想，例如关键字为 ABCD，动态加入空格为：A（空格）BCD、AB（空格）CD、ABC（空格）D、ABCD、A（空格）B（空格）CD、A（空格）BC（空格）D、AB（空格）C（空格）D、A（空格）B（空格）C（空格）D。则，以上 A（空格）BCD、AB（空格）CD、ABC（空格）D、ABCD、A（空格）B（空格）

CD、A（空格）BC（空格）D、AB（空格）C（空格）D、A（空格）B（空格）C（空格）D 全部都作为分词候选。

在一实施例中，基于每一个候选分词与第一分词语料库之间的匹配度从若干候选分词中选取第二分词语料，包括：

使用翻译模型将每一个候选分词分别翻译成多个均与高资源语言数据对应高资源候选分词；获取与高资源候选分词一致的第一分词语料在第一分词语料库中的出现频率；根据高资源候选分词的概率以及与该高资源候选分词一致的第一分词语料在第一分词语料库中的出现频率计算频率指数；将频率指数最高的第一分词语料所对应的高资源候选分词作为第二分词语料。根据频率指数的方式有多种，本申请不作限定，例如可以是对出现频率和翻译概率进行相加得到。

本实施例中应用到翻译模型，翻译模型是市面上现有的、常规使用的，例如 Transfomer，RNN-encoder-decoder、Attention 等。本实施例采用翻译模型将低资源语言转换为高资源语言。翻译模型通常体现为翻译软件，被承载在网页端或者智能终端作为一种客户端，翻译模型往往提供给用户交互操作界面，以进行可视化的翻译操作过程。例如候选分词为 abcd，需要翻译成高资源语言，为 ABCD，abcd 与 ABCD 往往表达意思实质相同，仅仅是两种不同语言体系的不同表达。例如“happy birthday day”可以翻译成祝你生日快乐，也可能翻译成愿你生日开心，翻译出来的结果实质意思是相同的，但是用词、表达等会存在不同，而翻译模型翻译出来的不同结果同时对应有不同的翻译概率，翻译概率与翻译结果同时显示出来。

根据前述步骤，由于采集了足够的高资源语言，已经形成了丰富的第一分词语料库。本实施例中，获取与高资源候选分词一致的第一分词语料在第一分词语料库中的出现频率，例如第一分词语料库中有 15 个短语 ABCD，则根据 abcd 的低资源语言翻译成的高资源语言的高资源候选分词 ABCD 在第一分词语料库中匹配到 15 个，匹配频次为 15 次，假设第一分词语料库中总共有 M 个短语，M 为自然数，则 ABCD 的出现频率为  $15/M$ 。假设根据候选分词翻译的高资源候选分词有 A（空格）BCD、ABC（空格）D，两者分别在第一分词语料库中匹配到 12 个，23 个，显然 ABC（空格）D 的出现频率高于 A（空格）BCD 的出现频率，因此将 ABC（空格）D 作为第二分词语料。

103：基于第二分词语料训练得到分词模型，将若干待分词数据输入至分词

模型中以输出多个分词候选结果。

本申请实施例中，步骤 103 基于第二分词语料训练得到分词模型，将若干待分词数据输入至分词模型中以输出多个分词候选结果，具体参考图 5 所示，包括两个子步骤，为：1031、将第二分词语料作为训练数据输入至预设的 CRF 模型中训练得到分词模型；1032、将若干待分词数据输入至分词模型中以输出多个分词候选结果。

CRF 模型是一种条件随机场模型，本申请采用 CRF 模型作为模型的基础，在 CRF 模型上输入训练数据进行模型的训练。将上一个步骤得到的第二分词语料作为所述训练数据。训练好分词模型之后，当输入待分词数据，则可以自动输出多个分词候选结果。分词候选结果的数量可以根据实际情况进行设定，例如是 5 个。当然也可以是其他数值情况，本申请不作限定。需要说明的是，由于本申请的目的之一是希望通过一系列的方式之后可以提供给低资源语言地区方便，能够基于高资源语言关联低资源语言，从而便于对低资源语言进行分词。因此在本实施例中，待分词数据往往为低资源语言数据。

104：基于每一个分词候选结果与第一分词语料库之间的匹配度，选取匹配度最高的分词候选结果作为分词结果。

在一实施例中，结合图 5 所示，基于每一个分词候选结果与第一分词语料库之间的匹配度，选取匹配度最高的分词候选结果作为分词结果，包括：1041、获取与分词候选结果一致的第一分词语料在第一分词语料库中的出现频率；1042、将出现频率最高的第一分词语料所对应的分词候选结果作为分词结果。

对最终分词结果的选取统一基于第一分词语料库的出现频率作为匹配度的考察因素。由于待数据可能为低资源语言数据，那么对应的分词候选结果也是低资源语言数据，在此可以选择翻译模型将分词候选结果翻译为高资源语言模型，然后基于翻译后的分词候选结果从第一分词语料库中匹配一致的第一分词语料，并且统计出现频率，将出现频率最高的第一分词语料所对应的分词候选结果作为分词结果。

另外一方面，图 6 示出了本申请实施例提供的一组基于跨语言数据增强的分词装置。如图 6 所示，基于跨语言数据增强的分词装置包括第一采集模块 201、第二采集模块 202、分词候选模块 203、分词确定模块 204。其中，第一采集模块 201 配置为采集若干组高资源语言数据，根据每一组所述高资源语言数据处理得到一组第一分词语料；若干组第一分词语料构成第一分词语料库。第二采

集模块 202 配置为采集若干低资源语言数据，根据低资源语言数据处理得到若干候选分词，基于每一个候选分词与第一分词语料库之间的匹配度从若干候选份分词中选取第二分词语料。分词候选模块 203 配置为基于第二分词语料训练得到分词模型，将若干待分词数据输入至分词模型中以输出多个分词候选结果。分词确定模块 204 配置为基于每一个分词候选结果与第一分词语料库之间的匹配度，选取匹配度最高的分词候选结果作为分词结果。

在一实施例中，第一采集模块 201 中，根据所述高资源语言数据处理得到第一分词语料，包括：分别对每一组高资源语言数据进行预处理；将预处理后的每一组高资源语言数据进行数据转换以形成多组短句，每一组短句包含若干个具有相同语言单位数的短语；将所包含的语言单位完全相同的短语划分为同一个短语集合，所述短语集合中包含若干种短语；分别获取每一种短语在为同一个短语集合中的出现频率，根据出现频率从高至低的排序选取排名前第一阈值的短语为第一分词语料。

其中，预处理包括按照第一预设规则对高资源语言数据中的标点符号进行转换、按照第二预设规则对高资源语言数据中的表情符号转换为语言单位。

上述根据出现频率从高至低的排序选取排名前第一阈值的短语为第一分词语料，包括：根据出现频率从高至低的排序选取排名前第一阈值的短语为候选第一分词语料；从同一个短语集合中任意挑选一个短语，将短语中的全部空格去除后形成关键字；输入关键字至预设分词模型以输出预测分词语料；比对候选第一分词语料和预测分词语料，当第一分词语料与预测分词语料一致时，将候选第一分词语料作为第一分词语料；当第一分词语料与预测分词语料不一致时，根据选词预设规则从候选第一分词语料和预测分词语料选取出第一分词语料。

在第二采集模块 202 中，根据低资源语言数据处理得到若干候选分词，包括：分别对每一组低资源语言数据进行预处理；将预处理后的每一组低资源语言数据进行数据转换以形成多组短句，每一组短句包含若干个具有相同语言单位数的短语；将所包含的语言单位完全相同的短语划分为同一个短语集合，所述短语集合中包含若干种短语；分别获取每一种短语在为同一个短语集合中的出现频率，根据出现频率从高至低的排序选取排名前第二阈值的短语为候选分词。

其中，预处理包括按照第一预设规则对低资源语言数据中的标点符号进行

转换、按照第二预设规则对高资源语言数据中的表情符号转换为语言单位。

在第二采集模块 202 中，根据低资源语言数据处理得到若干候选分词还可以是：分别对每一组低资源语言数据进行预处理；将预处理后的每一组低资源语言数据进行数据转换以形成多组短句，每一组短句包含若干个具有相同语言单位数的短语；将任意一个短语去除短语中的空格后形成关键字；采用动态规划编程方法在关键字中加入空格，以形成所有可能的分词作为候选分词。

基于每一个候选分词与第一分词语料库之间的匹配度从若干候选分词中选取第二分词语料，包括：使用翻译模型将每一个候选分词分别翻译成与高资源语言数据对应高资源候选分词；获取与高资源候选分词一致的第一分词语料在第一分词语料库中的出现频率；将出现频率最高的第一分词语料所对应的高资源候选分词作为第二分词语料。

在一实施例中，基于第二分词语料训练得到分词模型，是将第二分词语料作为训练数据输入至预设的 CRF 模型中训练得到分词模型。基于每一个分词候选结果与第一分词语料库之间的匹配度，选取匹配度最高的分词候选结果作为分词结果，包括：获取与分词候选结果一致的第一分词语料在第一分词语料库中的出现频率；将出现频率最高的第一分词语料所对应的分词候选结果作为分词结果。

如图 7 所示，本申请实施例还提供一种基于跨语言数据增强的分词设备，包括：存储器 301 以及一个或多个处理器 302；所述存储器 301，配置为存储一个或多个程序；当所述一个或多个程序被所述一个或多个处理器 302 执行，使得所述一个或多个处理器实现如本申请所述的基于跨语言数据增强的分词方法。

本申请实施例还提供一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行时配置为执行如上述实施例提供的基于跨语言数据增强的分词方法。

当然，本申请实施例所提供的一种包含计算机可执行指令的存储介质，其计算机可执行指令不限于如上所述的基于跨语言数据增强的分词方法，还可以执行本申请任意实施例所提供的基于跨语言数据增强的分词方法中的相关操作。

本申请实施例还提供一种基于跨语言数据增强的分词处理的程序，该程序被执行时，可以实现如上述实施例提及的基于跨语言数据增强的分词方法有关

的操作。

上述仅为本申请的较佳实施例及所运用的技术原理。本申请不限于这里所述的特定实施例，对本领域技术人员来说能够进行的各种明显变化、重新调整及替代均不会脱离本申请的保护范围。因此，虽然通过以上实施例对本申请进行了较为详细的说明，但是本申请不仅仅限于以上实施例，在不脱离本申请构思的情况下，还可以包括更多其他等效实施例，而本申请的范围由权利要求的范围决。

## 权 利 要 求 书

1.一种基于跨语言数据增强的分词方法，应用于服务器中，其中，包括：

采集若干组高资源语言数据，根据每一组所述高资源语言数据处理得到一组第一分词语料；若干组第一分词语料构成第一分词语料库；

采集若干低资源语言数据，根据低资源语言数据处理得到若干候选分词，基于每一个候选分词与第一分词语料库之间的匹配度从若干候选分词中选取第二分词语料；

基于第二分词语料训练得到分词模型，将若干待分词数据输入至分词模型中以输出多个分词候选结果；

基于每一个分词候选结果与第一分词语料库之间的匹配度，选取匹配度最高的分词候选结果作为分词结果。

2.根据权利要求1所述的分词方法，其中，根据每一组所述高资源语言数据处理得到第一分词语料，包括：

分别对每一组高资源语言数据进行预处理；

将预处理后的每一组高资源语言数据进行数据转换以形成多组短句，每一组短句包含若干个具有相同语言单位数的短语；

将所包含的语言单位完全相同的短语划分为同一个短语集合，所述短语集合中包含若干种短语；

分别获取每一种短语在为同一个短语集合中的出现频率，根据出现频率从高至低的排序选取排名前第一阈值的短语为第一分词语料。

3.根据权利要求2所述的分词方法，其中，所述预处理包括按照第一预设规则对高资源语言数据中的标点符号进行转换和/或按照第二预设规则对高资源语言数据中的表情符号转换为语言单位。

4.根据权利要求2所述的分词方法，其中，根据出现频率从高至低的排序选取排名前第一阈值的短语为第一分词语料，包括：

根据出现频率从高至低的排序选取排名前第一阈值的短语为候选第一分词语料；

从同一个短语集合中任意挑选一个短语，将短语中的全部空格去除后形成关键字；

输入关键字至预设分词模型以输出预测分词语料；

比对候选第一分词语料和预测分词语料，当第一分词语料与预测分词语料一致时，将候选第一分词语料作为第一分词语料；当第一分词语料与预测分词

语料不一致时，对候选第一分词语料进行调整，将调整后的候选第一分词语料作为第一分词语料。

5. 根据权利要求 1-4 中任一项所述的分词方法，其中，根据低资源语言数据处理得到若干候选分词，包括：

分别对每一组低资源语言数据进行预处理；

将预处理后的每一组低资源语言数据进行数据转换以形成多组短句，每一组短句包含若干个具有相同语言单位数的短语；

将所包含的语言单位完全相同的短语划分为同一个短语集合，所述短语集合中包含若干种短语；

分别获取每一种短语在为同一个短语集合中的出现频率，根据出现频率从高至低的排序选取排名前第二阈值的短语为候选分词。

6. 根据权利要求 1-4 中任一项所述的分词方法，其中，根据低资源语言数据处理得到若干候选分词，包括：

分别对每一组低资源语言数据进行预处理；

将预处理后的每一组低资源语言数据进行数据转换以形成多组短句，每一组短句包含若干个具有相同语言单位数的短语；

将任意一个短语去除短语中的空格后形成关键字；

采用动态规划编程方法在关键字中加入空格，以形成所有可能的分词作为候选分词。

7. 根据权利要求 5 或 6 所述的分词方法，其中，基于每一个候选分词与第一分词语料库之间的匹配度从若干候选分词中选取第二分词语料，包括：

使用翻译模型将每一个候选分词分别翻译成多个均与高资源语言数据对应高资源候选分词；每一个所述高资源候选分词对应有翻译模型分配的翻译概率；

获取与高资源候选分词一致的第一分词语料在第一分词语料库中的出现频率；

根据高资源候选分词的概率以及与该高资源候选分词一致的第一分词语料在第一分词语料库中的出现频率计算频率指数；

将频率指数最高的第一分词语料所对应的高资源候选分词作为第二分词语料。

8. 根据权利要求 1-7 中任一项所述的分词方法，其中，基于第二分词语料训练得到分词模型，包括：

将第二分词语料作为训练数据输入至预设的 CRF 模型中训练得到分词模型。

9.根据权利要求 1-8 中任一项所述的分词方法，其中，基于每一个分词候选结果与第一分词语料库之间的匹配度，选取匹配度最高的分词候选结果作为分词结果，包括：

获取与分词候选结果一致的第一分词语料在第一分词语料库中的出现频率；

将出现频率最高的第一分词语料所对应的分词候选结果作为分词结果。

10.一种基于跨语言数据增强的分词装置，其中，包括：

第一采集模块，配置为采集若干组高资源语言数据，根据每一组所述高资源语言数据处理得到一组第一分词语料；若干组第一分词语料构成第一分词语料库；

第二采集模块，配置为采集若干低资源语言数据，根据低资源语言数据处理得到若干候选分词，基于每一个候选分词与第一分词语料库之间的匹配度从若干候选分词中选取第二分词语料；

分词候选模块，配置为基于第二分词语料训练得到分词模型，将若干待分词数据输入至分词模型中以输出多个分词候选结果；

分词确定模块，配置为基于每一个分词候选结果与第一分词语料库之间的匹配度，选取匹配度最高的分词候选结果作为分词结果。

11.一种基于跨语言数据增强的分词设备，其中，包括：存储器以及一个或多个处理器；

所述存储器，配置为存储一个或多个程序；

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求 1-9 任一所述的一种基于跨语言数据增强的分词方法。

12.一种包含计算机可执行指令的存储介质，其中，所述计算机可执行指令在由计算机处理器执行时配置为执行如权利要求 1-9 任一所述的一种基于跨语言数据增强的分词方法。