

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2022 年 11 月 3 日 (03.11.2022)



(10) 国际公布号
WO 2022/228041 A1

(51) 国际专利分类号:
G06F 40/126 (2020.01) **G06F 40/247** (2020.01)

(21) 国际申请号: PCT/CN2022/084963

(22) 国际申请日: 2022 年 4 月 2 日 (02.04.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110454958.8 2021 年 4 月 26 日 (26.04.2021) CN

(71) 申请人: 北京有竹居网络技术有限公司(**BEIJING YOUZHUJU NETWORK TECHNOLOGY CO. LTD.**)
[CN/CN]; 中国北京市平谷区林荫北街 13 号信息大厦 802 室, Beijing 101299 (CN)。

(72) 发明人: 潘晓(**PAN, Xiao**); 中国北京市海淀区知春路 63 号中国卫星通信大厦今日头条小邮局, Beijing 100086 (CN)。王明轩(**WANG, Mingxuan**); 中国北京市海淀区知春路 63 号中国卫星通信大厦今日头条小邮局, Beijing 100086 (CN)。吴礼蔚(**WU, Liwei**); 中国北京市海淀区

知春路 63 号中国卫星通信大厦今日头条小邮局, Beijing 100086 (CN)。李磊(**LI, Lei**); 中国北京市海淀区知春路 63 号中国卫星通信大厦今日头条小邮局, Beijing 100086 (CN)。

(74) 代理人: 北京品源专利代理有限公司(**BEYOND ATTORNEYS AT LAW**); 中国北京市海淀区莲花池东路 39 号西金大厦 6 层, Beijing 100036 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,

(54) **Title:** TRANSLATION MODEL TRAINING METHOD, APPARATUS, AND DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 翻译模型的训练方法、装置、设备和存储介质

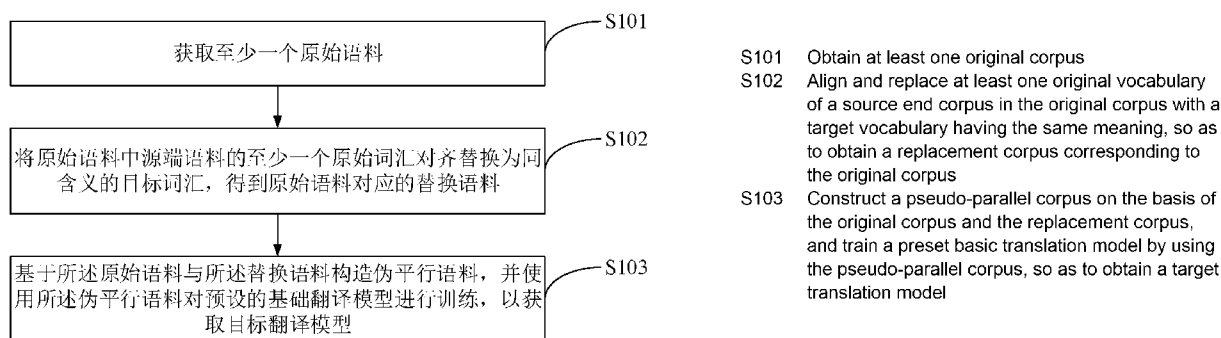


图 1

(57) **Abstract:** A translation model training method, apparatus, and device, and a storage medium. The translation model training method comprises: obtaining at least one original corpus (S101); aligning and replacing at least one original vocabulary of a source end corpus in the original corpus with a target vocabulary having the same meaning, so as to obtain a replacement corpus corresponding to the original corpus (S102), wherein the original vocabulary and the target vocabulary are different in language; and constructing a pseudo-parallel corpus on the basis of the original corpus and the replacement corpus, and training a preset basic translation model by using the pseudo-parallel corpus, so as to obtain a target translation model (S103).

(57) **摘要:** 一种翻译模型的训练方法、装置、设备和存储介质。该翻译模型的训练方法包括: 获取至少一个原始语料 (S101); 将所述原始语料中源端语料的至少一个原始词汇对齐替换为同含义的目标词汇, 得到所述原始语料对应的替换语料 (S102); 其中, 所述原始词汇与所述目标词汇的语种不同; 基于所述原始语料与所述替换语料构造伪平行语料, 并使用所述伪平行语料对预设的基础翻译模型进行训练, 以获取目标翻译模型 (S103)。

NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布：

— 包括国际检索报告(条约第21条(3))。

翻译模型的训练方法、装置、设备和存储介质

本申请要求在2021年04月26日提交中国专利局、申请号为202110454958.8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

本公开实施例涉及计算机技术领域，例如涉及一种翻译模型的训练方法、装置、设备和存储介质。

背景技术

随着计算机技术的不断发展，各种各样的翻译软件应运而生，成为了人们获取外部信息的重要渠道。

翻译软件的语言翻译模型的建立，通常是基于一通用语种为中心的平行语料进行训练得到的，用于实现通用语种与其它语种之间的翻译（以通用语种为英语为例，如实现英译法等）。但是，这样的翻译软件在其它非通用语种对（例如，德译法）上的翻译准确性较低。

发明内容

本公开提供一种翻译模型的训练方法、装置、设备和存储介质，以提高翻译模型在各场景下的翻译准确性。

本公开实施例提供了一种翻译模型的训练方法，包括：

获取至少一个原始语料；

将所述原始语料中源端语料的至少一个原始词汇对齐替换为同含义的目标词汇，得到所述原始语料对应的替换语料；其中，所述原始词汇与所述目标词汇的语种不同；

基于所述原始语料与所述替换语料构造伪平行语料，并使用所述伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型。

本公开实施例提供了一种翻译模型的训练装置，包括：

获取模块，设置为获取至少一个原始语料；

替换模块，设置为将所述原始语料中源端语料的至少一个原始词汇对齐替换为同含义的目标词汇，得到所述原始语料对应的替换语料；其中，所述原始词汇与所述目标词汇的语种不同；

构造模块，设置为基于所述原始语料与所述替换语料构造伪平行语料；

训练模块，设置为使用所述伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型。

本公开实施例提供了一种翻译模型的训练设备，包括存储器和处理器，所述存储器存储有计算机程序，所述处理器执行所述计算机程序时实现上述的翻译模型的训练方法。

本公开实施例提供了一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现上述的翻译模型的训练方法。

附图说明

图1为本公开实施例提供的一种翻译模型的训练方法的流程示意图；

图2为本公开实施例提供的一种伪平行语料的构造过程的原理示意图；

图3为本公开实施例提供的另一种伪平行语料的构造过程的原理示意图；

图4为本公开实施例提供的另一种翻译模型的训练方法的流程示意图；

图5为本公开实施例提供的一种翻译模型的训练过程的原理示意图；

图6为本公开实施例提供的一种翻译模型的训练装置的结构示意图；

图7为本公开实施例提供的一种翻译模型的训练设备的结构示意图。

具体实施方式

下面将参照附图描述本公开的实施例。虽然附图中显示了本公开的一些实施例，然而，本公开可以通过各种形式来实现，提供这些实施例是为了理解本公开。本公开的附图及实施例仅用于示例性作用。

本公开的方法实施方式中记载的各个步骤可以按照不同的顺序执行，和/或并行执行。此外，方法实施方式可以包括附加的步骤和/或省略执行示出的步骤。本公开的范围在此方面不受限制。

本文使用的术语“包括”及其变形是开放性包括，即“包括但不限于”。术语“基于”是“至少部分地基于”。术语“一个实施例”表示“至少一个实施例”；术语“另一实施例”表示“至少一个另外的实施例”；术语“一些实施例”表示“至少一些实施例”。其他术语的相关定义将在下文描述中给出。

本公开中提及的“第一”、“第二”等概念仅用于对不同的装置、模块或单元进行区分，并非用于限定这些装置、模块或单元所执行的功能的顺序或者相互依存关系。

本公开中提及的“一个”、“多个”的修饰是示意性而非限制性的，本领域技术人员应当理解，除非在上下文另有指出，否则应该理解为“一个或多个”。

本公开实施方式中的多个装置之间所交互的消息或者信息的名称仅用于说明性的目的，而并不是用于对这些消息或信息的范围进行限制。

下文中将结合附图对本公开的实施例进行说明。需要说明的是，在不冲突的情况下，本公开中的实施例及实施例中的特征可以相互任意组合。

下述方法实施例的执行主体可以是翻译模型的训练装置，该装置可以通过软件、硬件或者软硬件结合的方式实现成为电子设备的部分或者全部。可选的，该电子设备可以为客户端，包括但不限于智能手机、平板电脑、电子书阅读器以及车载终端等。该电子设备也可以为独立的服务器或者服务器集群，本公开实施例对电子设备的形式不做限定。下述方法实施例以执行主体是电子设备为例进行说明。

图1为本公开实施例提供的一种翻译模型的训练方法的流程示意图。本实施例涉及的是电子设备如何训练多语种翻译模型的过程。如图1所示，该方法可以包括：

S101、获取至少一个原始语料。

原始语料作为训练数据，用于训练基础翻译模型。该原始语料的模态可以为图像、文本、视频或者音频中的至少一种。原始语料中所包含的语种可以为单一语种，也可以为多个语种，本实施例对此不做限定。

可选地，原始语料可以包括单语语料和/或平行语料。其中，所述平行语料包括成对的源端语料和目标端语料。源端语料可以理解为翻译之前的语料，目标端语料可以理解为源端语料经过翻译后的语料。以平行语料为文本语料为例，中英平行语料包括一个中文文档和一个对应的英文文档，如果通过翻译模型进行中译英操作，那么中文文档即为源端语料，英文文档即为目标端语料。

单语语料可以为源端语料或者目标端语料，缺少对应的平行语料。例如，在中医领域，可以获取到大量的中文语料，也可以获取到英文语料，但是中英文相互对应的平行语料就很难获取。

通常，语料数据库中存储有大量的原始语料，因此，电子设备可以直接从语料数据库中获取至少一个原始语料。

S102、将原始语料中源端语料的至少一个原始词汇对齐替换为同含义的目标词汇，得到原始语料对应的替换语料。

目标词汇是原始词汇的同义词，且原始词汇与目标词汇的语种不同。由于

多语种的同义词词典相对比较容易获取，因此，电子设备可以基于多语种的同义词词典，将原始语料中源端语料的至少一个原始词汇替换成其它任意语种的同义词，从而得到原始语料对应的替换语料。

为了能够实现多语种之间的翻译，可选地，所述替换语料中的各目标词汇的语种至少部分不同。

以下基于平行语料和单语语料，介绍上述替换语料的获取过程：

当原始语料为平行语料时，可以将平行语料中源端语料的部分词汇对齐替换为其它任意语种的同义词汇，目标端语料保持不变。示例性的，如图 2 所示，假设平行语料的源端语料为“I like sing and dance”，其语种为英语，目标端语料为“J'adore chanter et danser”，其语种为法语。对此，电子设备可以将源端语料中的词汇“sing”对齐替换为同含义的中文词汇“唱歌”，以及将词汇“dance”对齐替换为同含义的中文词汇“跳舞”，从而形成源端语料对应的替换语料“I like 唱歌 and 跳舞”，目标端语料不替换，保持不变。对齐替换后的各目标词汇的语种可以相同，也可以不同，即可以将上述词汇“sing”对齐替换为同含义的德语词汇，将上述词汇“dance”对齐替换为同含义的中文词汇。

当原始语料为单语语料时，单语语料即为上述源端语料，其不存在对应的翻译语料，此时，可以将单语语料中的部分词汇对齐替换为其它任意语种的同义词汇。替换后的各同义词汇的语种至少部分不同，即替换后的各同义词汇可以为同一语种，也可以为不同语种。示例性的，如图 3 所示，假设单语语料为中文语料“你喜欢哪种类型的音乐呢”，为了能够使用大量的单语语料进行翻译模型的训练，电子设备可以将上述单语语料中的词汇“喜欢”对齐替换为同含义的英语词汇“like”、词汇“哪种”对齐替换为同含义的法语词汇“quel”，词汇“音乐”对齐替换为同含义的德语词汇“Musik”，从而形成单语语料对应的替换语料“你 like quel 类型的 Musik 呢”。

S103、基于所述原始语料与所述替换语料构造伪平行语料，并使用所述伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型。

在得到原始语料对应的替换语料之后，可以基于原始语料与对应的替换语料来构造平行语料。由于该平行语料并不是标准的语料，而是通过同义词对齐替换后获取得到的，因此，该平行语料被标注为伪平行语料。

以下针对不同的原始语料，介绍伪平行语料的构造过程：

当原始语料为平行语料时，在将平行语料中源端语料的部分词汇对齐替换为其它任意语种的同义词汇，得到源端语料对应的替换语料后，可以将平行语料中源端语料对应的替换语料作为伪源端语料，将平行语料中目标端语料作为

伪目标端语料，组成伪平行语料。目标端语料保持不变，能够保证翻译模型学习到正确的翻译结果。继续参见图 2，将替换语料“I like 唱歌 and 跳舞”作为伪源端语料，平行语料中目标端语料“J adore chanter et danser”继续作为伪目标端语料，从而形成伪平行语料。

当原始语料为单语语料时，在将单语语料中的部分词汇对齐替换为其它任意语种的同义词汇，得到单语语料对应的替换语料后，可以将单语语料对应的替换语料作为伪源端语料以及将单语语料作为伪目标端语料，组成伪平行语料。继续参见图 3，将替换语料“你 like quel 类型的 Musik 呢”作为伪源端语料，将单语语料本身“你喜欢哪种类型的音乐呢”作为伪目标端语料，从而形成伪平行语料。

在得到大量的伪平行语料之后，电子设备可以使用该伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型。对于单语语料来说，经过简单的同义词汇对齐替换后，就可以将单语语料直接应用于翻译模型的训练过程中，不需要采用“回译”技术，大大缩短了使用单语语料训练翻译模型的流程，从而提高了翻译模型的训练效率。

可选地，上述基础翻译模型可以包括序列到序列（sequence to sequence, seq2seq）模型，是一种编码（Encoder）-解码（Decoder）结构的神经网络，输入是一个序列（Sequence），输出也是一个序列；在 Encoder 中，将可变长度的序列转变为固定长度的向量表示，Decoder 将这个固定长度的向量表示转换为可变长度的目标信号序列，进而实现不定长的输入到不定长的输出。序列到序列模型可以包括多种类型，例如，基于循环神经网络（Recurrent Neural Network, RNN）的 seq2seq 模型和基于卷积运算（Convolution, CONV）的 seq2seq 模型等，本实施例中对基础翻译模型的类型不做限定。

通过将原始语料中源端语料的部分词汇替换为其它任意语种的同义词汇，使得所构造的伪平行语料中包含了原始语料中没有出现的其它语种，进而通过所构造的伪平行语料对基础翻译模型进行训练，使得基础翻译模型能够学习到其它语种之间的语法结构以及词汇关联，增强了翻译模型在其它非通用语种对上的翻译准确性。例如，翻译模型通常是以英语为中心的平行语料进行训练得到的，那么在其它非英语语对上的翻译效果达不到期望要求，为此，继续参见图 2，本公开实施例将原始语料中的源端语料的英文词汇“sing”对齐替换为中文词汇“唱歌”，目标端语料保持不变，使用替换语料和平行语料中目标端语料所构造的伪平行语料训练基础翻译模型，可以使得基础翻译模型学习到中文词汇“唱歌”与法语词汇“chanter”是同一含义，并能学习到句子“I like 唱歌 and 跳舞”与句子“J adore chanter et danser”之间的语法结构和词汇关联，从而实现了中文到法语对之间的翻译，且保证了中文到法语之间的翻译准确性。

本公开实施例提供的翻译模型的训练方法，获取至少一个原始语料，将原始语料中源端语料的至少一个原始词汇对齐替换为同含义的目标词汇，得到原始语料对应的替换语料，且原始词汇与目标词汇的语种不同；基于原始语料与替换语料构造伪平行语料，并使用伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型。也就是说，通过将原始语料中源端语料的至少一个原始词汇对齐替换为其它任意语种的同义词汇后，能够构造出大量包含其它任意语种的伪平行语料，使用该伪平行语料训练翻译模型，使得翻译模型能够学习到其它任意语种之间的语法结构以及词汇关联，从而提高了翻译模型在其它非通用语种对上的翻译准确性。

在一个实施例中，上述基础翻译模型包括编码器和解码器。编码器可以对输入序列进行特征提取，得到特征向量。为了提高编码器对输入序列特征提取的准确性，在上述实施例的基础上，可选地，上述 S103 中使用伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型的过程可以为：使用所述伪平行语料，通过第一损失函数对所述基础翻译模型的编码器进行训练。

第一损失函数为对比学习损失函数，用于更新编码器的参数。由于伪平行语料是基于原始语料对应的替换语料与原始语料所构造的，即替换语料可以认为是原始语料的同义句，因此，为了提高翻译准确性，在使用伪平行语料训练基础翻译模型时，可以使用对比损失函数对基础翻译模型的编码器进行训练，从而拉近同义句经过编码器编码后的高维表达，同时拉远不相关句子经过编码器编码后的高维表达。也就是说，通过对比学习损失函数对编码器进行训练后，原本相似的两个输入语料，在经过编码器编码后，在特征空间中，两个输入语料的特征仍旧相似；而原本不相似的两个输入语料，在经过编码器编码后，在特征空间中，两个输入语料的特征也仍旧不相似。这样，在使用由平行语料和/或单语语料所构造的伪平行语料训练基础翻译模型时，能够保证训练得到的目标翻译模型的翻译效果。

可选地，如图 4 所示，上述使用所述伪平行语料，通过第一损失函数对所述基础翻译模型的编码器进行训练的过程可以为：

S401、构建所述伪平行语料中的伪源端语料的正例语料和负例语料。

伪平行语料可以包括成对的伪源端语料和伪目标端语料，伪源端语料可以理解为翻译之前的语料，伪目标端语料可以理解为伪源端语料经过翻译后的语料。上述正例语料是指与伪源端语料匹配度大于第一预设值的语料，即两者为相似语料，当两个语料完全相似时，两者的匹配度为 1。上述负例语料是指与伪源端语料匹配度小于第二预设值的语料，即两者为不相关语料，当两个语料完全不相关时，两者的匹配度为 0。上述第一预设值大于第二预设值。

可选地，上述伪源端语料通常为原始语料中源端语料对应的替换语料，因此，上述正例语料可以为当前伪平行语料中的伪目标端语料，上述负例语料为其它伪平行语料中的伪目标端语料。

当原始语料为单语语料时，上述伪平行语料中伪源端语料可以为单语语料对应的替换语料，伪目标端语料可以为单语语料本身，这样，电子设备可以将单语语料本身作为伪源端语料的正例语料，将其它伪平行语料中的伪目标端作为伪源端语料的负例语料。

当原始语料为平行语料时，上述伪平行语料中伪源端语料可以为平行语料中源端语料对应的替换语料，伪目标端语料可以为平行语料中目标端语料，这样，电子设备可以将当前平行语料中目标端语料作为伪源端语料的正例语料，将其它伪平行语料中的伪目标端语料作为伪源端语料的负例语料。

示例性的，参见图 5，假设伪平行语料中的伪源端语料为“I love you”，伪目标端语料为“Je t'aime”，该伪源端语料是原始语料中源端语料经过其它任意语种同义词替换后的替换语料，此时，可以将伪目标端语料“Je t'aime”作为替换语料“I love you”的正例语料，选择其它伪平行语料中的伪目标端语料作为替换语料“I love you”的负例语料（如图 5 中的英语语料“It's sunny”，法语语料“C'est la vie”以及中文语料“你是谁”）。另外，可以选取多个负例语料对编码器进行训练，从而提高翻译模型的训练效率。

S402、使用所述伪源端语料、所述正例语料以及所述负例语料，通过第一损失函数对所述编码器进行训练。

在得到伪源端语料的正例语料以及负例语料后，电子设备可以将伪源端语料、正例语料以及负例语料作为编码器的训练数据，采用对比学习损失函数对编码器进行反复训练，以不断更新编码器的参数，直至达到训练目标。训练目标为最大化伪源端语料和正例语料的向量表示之间的相似度，最小化伪源端语料和负例语料的向量表示之间的相似度。

继续参见图 5，电子设备将伪目标端语料“Je t'aime”作为锚点“I love you”的正例语料，将从其它伪平行语料的伪目标端语料中采样的英语语料“It's sunny”，法语语料“C'est la vie”以及中文语料“你是谁”作为锚点“I love you”的负例语料，采用对比损失函数 L^{cl} 对基础翻译模型的编码器进行训练，使得编码器能够拉近同义句编码后的高维表达，同时拉远不相关句子编码后的高维表达。

作为一种可选地实施方式，上述 S402 的过程可以包括以下步骤：

S4021、将所述伪源端语料、所述正例语料以及所述负例语料输入至所述编码器中，得到所述伪源端语料对应的第一向量表示、所述正例语料对应的第二

向量表示以及所述负例语料对应的第三向量表示。

编码器用于提取输入语料的特征向量，因此，将伪源端语料、伪源端语料的正例语料以及伪源端语料的负例语料分别输入至编码器中，经过编码器对伪源端语料、正例语料以及负例语料进行编码，从而得到伪源端语料对应的第一向量表示、正例语料对应的第二向量表示以及负例语料对应的第三向量表示。

S4022、根据所述第一向量表示和所述第二向量表示，确定第一损失函数的第一损失值，并基于所述第一损失值更新所述编码器的参数，直至所述第一损失函数的第一损失值满足收敛条件。

S4023、根据所述第一向量表示和所述第三向量表示，确定所述第一损失函数的第二损失值，并基于所述第二损失值更新所述编码器的参数，直至所述第一损失函数的第二损失值满足收敛条件。

第一损失函数为对比学习损失函数，其优化目标是当输入语料相似时，希望经过编码器编码后两个输入语料对应的向量表示也相似，当输入语料不相似时，希望经过编码器编码后两个输入语料对应的向量表示也不相似。因此，电子设备可以基于第一向量表示、以及第二向量表示伪源端语料与正例语料之间的匹配度，确定对比学习损失函数的第一损失值。当该第一损失值不满足收敛条件时，对编码器的参数进行更新，并将更新后的编码器作为 S4021 中的编码器，继续执行上述 S4021，直至对比损失函数的第一损失值满足收敛条件。

同理，电子设备可以基于第一向量表示、以及第三向量表示伪源端语料与负例语料之间的匹配度，确定对比学习损失函数的第二损失值。当该第二损失值不满足收敛条件时，对编码器的参数进行更新，并将更新后的编码器作为 S4021 中的编码器，继续执行上述 S4021，直至对比损失函数的第二损失值满足收敛条件。

在实际应用中，对编码器的训练可以叠加在整个基础翻译模型的训练过程中，使得编码器能够共享多任务训练的参数。为此，在上述实施例的基础上，可选地，上述 S103 中使用所述伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型的过程可以包括：使用伪平行语料，通过第一损失函数和第二损失函数对所述基础翻译模型进行多任务训练，以获取目标翻译模型。

所述第一损失函数为对比学习损失函数，用于更新编码器的参数，所述第二损失函数用于更新编码器和解码器的参数，即第二损失函数用于对整个基础翻译模型进行训练。可选地，第二损失函数可以为交叉熵损失函数。上述多任务至少可以包括对编码器的训练任务以及对编码器和解码器的训练任务，编码器可以共享多任务训练所更新的参数。其中一个任务是使用锚点语料、正例语

料以及负例语料，通过对比损失函数训练编码器，使得训练后的编码器能够拉近同义句编码后的高维表达，同时拉远不相关句子编码后的高维表达；另一个任务是使用伪平行语料，通过第二损失函数训练基础翻译模型，使得基础翻译模型学习到伪平行语料中所包含的其它任意语种之间的语法结构以及词汇关联，从而实现在零资源场景下以及无监督场景下的多语种之间的相互翻译。即在使用伪平行语料，通过第二损失函数训练基础翻译模型的基础上，加入对比学习损失函数进行多任务训练，使得训练后的目标翻译模型能够支持任意方向的多语种之间的翻译，且确保了翻译结果的准确性。

继续参见图 5，电子设备将伪平行语料中伪源端语料（即锚点）“I love you”作为基础翻译模型的输入，将伪平行语料中伪目标端语料“Je t'aime”作为期望输出，采用第二损失函数 L^m 对基础翻译模型中的编码器和解码器进行训练。同时，电子设备将锚点如“I love you”、锚点的正例语料如“Je t'aime”以及锚点的负例语料如“It's sunny”、“C'est la vie”以及“你是谁”等，输入至编码器中，经过编码器编码后得到对应的第一向量表示、第二向量表示以及第三向量表示，基于第一向量表示、第二向量表示以及第三向量表示采用对比损失函数对编码器进行训练，直至达到对比损失函数 L^{cl} 和第二损失函数 L^m 的收敛条件。

在本实施例中，在基础翻译模型的训练过程中，加入对比损失函数对基础翻译模型的编码器进行训练，使得训练后的编码器能够拉近同义句编码后的高维表达，同时拉远不相关句子编码后的高维表达。这样，使用其它任意语种同义词对齐替换后的替换语料所构造的伪平行语料训练基础翻译模型，使得训练后的目标翻译模型能够支持任意方向的多语种之间的翻译，且确保了翻译结果的准确性。

图 6 为本公开实施例提供的一种翻译模型的训练装置的结构示意图。如图 6 所示，该装置可以包括：获取模块 601、替换模块 602、构造模块 603 和训练模块 604。

获取模块 601 设置为获取至少一个原始语料；替换模块 602 设置为将所述原始语料中源端语料的至少一个原始词汇对齐替换为同含义的目标词汇，得到所述原始语料对应的替换语料；其中，所述原始词汇与所述目标词汇的语种不同；构造模块 603 设置为基于所述原始语料与所述替换语料构造伪平行语料；训练模块 604 设置为使用所述伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型。

本公开实施例提供的翻译模型的训练装置，获取至少一个原始语料，将原始语料中源端语料的至少一个原始词汇对齐替换为同含义的目标词汇，得到原始语料对应的替换语料，且原始词汇与目标词汇的语种不同；基于原始语料与

替换语料构造伪平行语料，并使用伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型。也就是说，通过将原始语料中源端语料的至少一个原始词汇对齐替换为其它任意语种的同义词汇后，能够构造出大量包含其它任意语种的伪平行语料，使用该伪平行语料训练翻译模型，使得翻译模型能够学习到其它任意语种之间的语法结构以及词汇关联，从而提高了翻译模型在其它非通用语种对上的翻译准确性。

可选地，所述原始语料包括单语语料和/或平行语料；其中，所述单语语料为所述源端语料，所述平行语料包括成对的源端语料和目标端语料。

在上述实施例的基础上，可选地，所述基础翻译模型包括编码器和解码器；上述训练模块 604 包括：第一训练单元；第一训练单元设置为使用所述伪平行语料，通过第一损失函数对所述基础翻译模型的编码器进行训练；其中，所述第一损失函数为对比学习损失函数，用于更新所述编码器的参数。

在上述实施例的基础上，可选地，训练模块 604 还包括：第二训练单元；第二训练单元设置为使用所述伪平行语料，通过第一损失函数和第二损失函数对所述基础翻译模型进行多任务训练，以获取目标翻译模型；其中，所述第一损失函数为对比学习损失函数，用于更新所述编码器的参数，所述第二损失函数用于更新所述编码器和所述解码器的参数。

在上述实施例的基础上，可选地，第一训练单元设置为构建所述伪平行语料中的伪源端语料的正例语料和负例语料；使用所述伪源端语料、所述正例语料以及所述负例语料，通过第一损失函数对所述编码器进行训练；其中，训练目标为最大化所述伪源端语料和所述正例语料的向量表示之间的相似度，最小化所述伪源端语料和所述负例语料的向量表示之间的相似度。

在上述实施例的基础上，可选地，第一训练单元设置为将所述伪源端语料、所述正例语料以及所述负例语料输入至所述编码器中，得到所述伪源端语料对应的第一向量表示、所述正例语料对应的第二向量表示以及所述负例语料对应的第三向量表示；根据所述第一向量表示和所述第二向量表示，确定第一损失函数的第一损失值，并基于所述第一损失值更新所述编码器的参数，直至所述第一损失函数的第一损失值满足收敛条件；根据所述第一向量表示和所述第三向量表示，确定所述第一损失函数的第二损失值，并基于所述第二损失值更新所述编码器的参数，直至所述第一损失函数的第二损失值满足收敛条件。

在上述实施例的基础上，可选地，当所述原始语料为单语语料时，构造模块 603 设置为将所述单语语料对应的替换语料作为伪源端语料以及将所述单语语料作为伪目标端语料，组成伪平行语料。

在上述实施例的基础上，可选地，当所述原始语料为平行语料时，构造模块 603 设置为将所述平行语料中源端语料对应的替换语料作为伪源端语料，以及将所述平行语料中目标端语料作为伪目标端语料，组成伪平行语料。

可选地，所述正例语料为所述伪目标端语料。

可选地，所述负例语料为其它伪平行语料中的伪目标端语料。

可选地，所述替换语料中的各目标词汇的语种至少部分不同。

下面参考图 7，其示出了适于用来实现本公开实施例的电子设备 700（即翻译模型的训练设备）的结构示意图。本公开实施例中的电子设备可以包括但不限于诸如移动电话、笔记本电脑、数字广播接收器、个人数字助理（Personal Digital Assistant, PDA）、平板电脑（PAD）、便携式多媒体播放器（Portable Media Player, PMP）、车载终端（例如车载导航终端）等等的移动终端以及诸如数字电视（Television, TV）、台式计算机等等的固定终端。图 7 示出的电子设备仅仅是一个示例，不应对本公开实施例的功能和使用范围带来任何限制。

如图 7 所示，电子设备 700 可以包括处理装置（例如中央处理器、图形处理器等）701，其可以根据存储在只读存储器（Read-Only Memory, ROM）702 中的程序或者从存储装置 708 加载到随机访问存储器（Random Access Memory, RAM）703 中的程序而执行各种适当的动作和处理。在 RAM 703 中，还存储有电子设备 700 操作所需的各种程序和数据。处理装置 701、ROM 702 以及 RAM 703 通过总线 704 彼此相连。输入/输出（Input/Output, I/O）接口 705 也连接至总线 704。

通常，以下装置可以连接至 I/O 接口 705：包括例如触摸屏、触摸板、键盘、鼠标、摄像头、麦克风、加速度计、陀螺仪等的输入装置 706；包括例如液晶显示器（Liquid Crystal Display, LCD）、扬声器、振动器等等的输出装置 707；包括例如磁带、硬盘等的存储装置 708；以及通信装置 709。通信装置 709 可以允许电子设备 700 与其他设备进行无线或有线通信以交换数据。虽然图 7 示出了具有各种装置的电子设备 700，但并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。

根据本公开的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本公开的实施例包括一种计算机程序产品，其包括承载在非暂态计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信装置 709 从网络上被下载和安装，或者从存储装置 708 被安装，或者从 ROM 702 被安装。在该计算机程序被处理装置 701 执行时，执行本公开实施例的方法中限定的上

述功能。

本公开上述的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、RAM、ROM、可擦式可编程只读存储器（Erasable Programmable Read-Only Memory, EPROM 或闪存）、光纤、便携式紧凑磁盘只读存储器（Compact Disc Read-Only Memory, CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开中，计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读信号介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：电线、光缆、射频（Radio Frequency, RF）等等，或者上述的任意合适的组合。

在一些实施方式中，客户端、服务器可以利用诸如超文本传输协议（HyperText Transfer Protocol, HTTP）之类的任何当前已知或未来研发的网络协议进行通信，并且可以与任意形式或介质的数字数据通信（例如，通信网络）互连。通信网络的示例包括局域网（Local Area Network, LAN），广域网（Wide Area Network, WAN），网际网（例如，互联网）以及端对端网络（例如，ad hoc 端对端网络），以及任何当前已知或未来研发的网络。

上述计算机可读介质可以是上述电子设备中所包含的；也可以是单独存在，而未装配入该电子设备中。

上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被该电子设备执行时，使得该电子设备：获取至少一个原始语料；将所述原始语料中源端语料的至少一个原始词汇对齐替换为同含义的目标词汇，得到所述原始语料对应的替换语料；其中，所述原始词汇与所述目标词汇的语种不同；基于所述原始语料与所述替换语料构造伪平行语料，并使用所述伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型。

可以以一种或多种程序设计语言或其组合来编写用于执行本公开的操作的

计算机程序代码，上述程序设计语言包括但不限于面向对象的程序设计语言——诸如 Java、Smalltalk、C++，还包括常规的过程式程序设计语言——诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括 LAN 或 WAN——连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。

附图中的流程图和框图，图示了按照本公开各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

描述于本公开实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现。其中，单元的名称在一种情况下并不构成对该单元本身的限定，例如，第一获取单元还可以被描述为“获取至少两个网际协议地址的单元”。

本文中以上描述的功能可以至少部分地由一个或多个硬件逻辑部件来执行。例如，非限制性地，可以使用的示范类型的硬件逻辑部件包括：现场可编程门阵列(Field Programmable Gate Array, FPGA)、专用集成电路(Application Specific Integrated Circuit, ASIC)、专用标准产品(Application Specific Standard Parts, ASSP)、片上系统(System on Chip, SOC)、复杂可编程逻辑设备(Complex Programmable Logic Device, CPLD)等等。

在本公开的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。机器可读存储介质的示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、RAM、

ROM、EPROM 或快闪存储器、光纤、CD-ROM、光学储存设备、磁储存设备、或上述内容的任何合适组合。

在一个实施例中，提供了一种翻译模型的训练设备，包括存储器和处理器，存储器存储有计算机程序，该处理器执行计算机程序时实现：

获取至少一个原始语料；

将所述原始语料中源端语料的至少一个原始词汇对齐替换为同含义的目标词汇，得到所述原始语料对应的替换语料；其中，所述原始词汇与所述目标词汇的语种不同；

基于所述原始语料与所述替换语料构造伪平行语料，并使用所述伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型。

可选地，所述原始语料包括单语语料和/或平行语料；其中，所述单语语料为所述源端语料，所述平行语料包括成对的源端语料和目标端语料。

可选地，所述基础翻译模型包括编码器和解码器；

在一个实施例中，处理器执行计算机程序时还实现：使用所述伪平行语料，通过第一损失函数对所述基础翻译模型的编码器进行训练；其中，所述第一损失函数为对比学习损失函数，用于更新所述编码器的参数。

在一个实施例中，处理器执行计算机程序时还实现：使用所述伪平行语料，通过第一损失函数和第二损失函数对所述基础翻译模型进行多任务训练，以获取目标翻译模型；其中，所述第一损失函数为对比学习损失函数，用于更新所述编码器的参数，所述第二损失函数用于更新所述编码器和所述解码器的参数。

在一个实施例中，处理器执行计算机程序时还实现：构建所述伪平行语料中的伪源端语料的正例语料和负例语料；使用所述伪源端语料、所述正例语料以及所述负例语料，通过第一损失函数对所述编码器进行训练；其中，训练目标为最大化所述伪源端语料和所述正例语料的向量表示之间的相似度，最小化所述伪源端语料和所述负例语料的向量表示之间的相似度。

在一个实施例中，处理器执行计算机程序时还实现：将所述伪源端语料、所述正例语料以及所述负例语料输入至所述编码器中，得到所述伪源端语料对应的第一向量表示、所述正例语料对应的第二向量表示以及所述负例语料对应的第三向量表示；根据所述第一向量表示和所述第二向量表示，确定第一损失函数的第一损失值，并基于所述第一损失值更新所述编码器的参数，直至所述第一损失函数的第一损失值满足收敛条件；根据所述第一向量表示和所述第三向量表示，确定所述第一损失函数的第二损失值，并基于所述第二损失值更新所述编码器的参数，直至所述第一损失函数的第二损失值满足收敛条件。

在一个实施例中，当所述原始语料为单语语料时，处理器执行计算机程序时还实现：将所述单语语料对应的替换语料作为伪源端语料以及将所述单语语料作为伪目标端语料，组成伪平行语料。

在一个实施例中，当所述原始语料为平行语料时，处理器执行计算机程序时还实现：将所述平行语料中源端语料对应的替换语料作为伪源端语料，以及将所述平行语料中目标端语料作为伪目标端语料，组成伪平行语料。

可选地，所述正例语料为所述伪目标端语料。

可选地，所述负例语料为其它伪平行语料中的伪目标端语料。

可选地，所述替换语料中的各目标词汇的语种至少部分不同。

在一个实施例中，还提供了一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现：

获取至少一个原始语料；

将所述原始语料中源端语料的至少一个原始词汇对齐替换为同含义的目标词汇，得到所述原始语料对应的替换语料；其中，所述原始词汇与所述目标词汇的语种不同；

基于所述原始语料与所述替换语料构造伪平行语料，并使用所述伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型。

上述实施例中提供的翻译模型的训练装置、设备以及存储介质可执行本公开任意实施例所提供的翻译模型的训练方法，具备执行该方法相应的功能模块和效果。未在上述实施例中详尽描述的技术细节，可参见本公开任意实施例所提供的翻译模型的训练方法。

根据本公开的一个或多个实施例，提供一种翻译模型的训练方法，包括：

获取至少一个原始语料；

将所述原始语料中源端语料的至少一个原始词汇对齐替换为同含义的目标词汇，得到所述原始语料对应的替换语料；其中，所述原始词汇与所述目标词汇的语种不同；

基于所述原始语料与所述替换语料构造伪平行语料，并使用所述伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型。

可选地，所述原始语料包括单语语料和/或平行语料；其中，所述单语语料为所述源端语料，所述平行语料包括成对的源端语料和目标端语料。

可选地，所述基础翻译模型包括编码器和解码器；

根据本公开的一个或多个实施例，提供了如上的翻译模型的训练方法，还包括：使用所述伪平行语料，通过第一损失函数对所述基础翻译模型的编码器进行训练；其中，所述第一损失函数为对比学习损失函数，用于更新所述编码器的参数。

根据本公开的一个或多个实施例，提供了如上的翻译模型的训练方法，还包括：使用所述伪平行语料，通过第一损失函数和第二损失函数对所述基础翻译模型进行多任务训练，以获取目标翻译模型；其中，所述第一损失函数为对比学习损失函数，用于更新所述编码器的参数，所述第二损失函数用于更新所述编码器和所述解码器的参数。

根据本公开的一个或多个实施例，提供了如上的翻译模型的训练方法，还包括：构建所述伪平行语料中的伪源端语料的正例语料和负例语料；使用所述伪源端语料、所述正例语料以及所述负例语料，通过第一损失函数对所述编码器进行训练；其中，训练目标为最大化所述伪源端语料和所述正例语料的向量表示之间的相似度，最小化所述伪源端语料和所述负例语料的向量表示之间的相似度。

根据本公开的一个或多个实施例，提供了如上的翻译模型的训练方法，还包括：将所述伪源端语料、所述正例语料以及所述负例语料输入至所述编码器中，得到所述伪源端语料对应的第一向量表示、所述正例语料对应的第二向量表示以及所述负例语料对应的第三向量表示；根据所述第一向量表示和所述第二向量表示，确定第一损失函数的第一损失值，并基于所述第一损失值更新所述编码器的参数，直至所述第一损失函数的第一损失值满足收敛条件；根据所述第一向量表示和所述第三向量表示，确定所述第一损失函数的第二损失值，并基于所述第二损失值更新所述编码器的参数，直至所述第一损失函数的第二损失值满足收敛条件。

根据本公开的一个或多个实施例，提供了如上的翻译模型的训练方法，还包括：当所述原始语料为单语语料时，将所述单语语料对应的替换语料作为伪源端语料以及将所述单语语料作为伪目标端语料，组成伪平行语料。

根据本公开的一个或多个实施例，提供了如上的翻译模型的训练方法，还包括：当所述原始语料为平行语料时，将所述平行语料中源端语料对应的替换语料作为伪源端语料，以及将所述平行语料中目标端语料作为伪目标端语料，组成伪平行语料。

可选地，所述正例语料为所述伪目标端语料。

可选地，所述负例语料为其它伪平行语料中的伪目标端语料。

可选地，所述替换语料中的各目标词汇的语种至少部分不同。

此外，虽然采用特定次序描绘了各操作，但是这不应理解为要求这些操作以所示出的特定次序或以顺序次序执行来执行。在一定环境下，多任务和并行处理可能是有利的。同样地，虽然在上面论述中包含了多个实现细节，但是这些不应被解释为对本公开的范围的限制。在单独的实施例的上下文中描述的一些特征还可以组合地实现在单个实施例中。相反地，在单个实施例的上下文中描述的各种特征也可以单独地或以任何合适的子组合的方式实现在多个实施例中。

权利要求书

1、一种翻译模型的训练方法，包括：

获取至少一个原始语料；

将所述原始语料中源端语料的至少一个原始词汇对齐替换为同含义的目标词汇，得到所述原始语料对应的替换语料；其中，所述原始词汇与所述目标词汇的语种不同；

基于所述原始语料与所述替换语料构造伪平行语料，并使用所述伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型。

2、根据权利要求1所述的方法，其中，所述原始语料包括单语语料和平行语料中的至少之一；其中，所述单语语料为所述源端语料，所述平行语料包括成对的源端语料和目标端语料。

3、根据权利要求1所述的方法，其中，所述基础翻译模型包括编码器和解码器；

所述使用所述伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型，包括：

使用所述伪平行语料，通过第一损失函数对所述基础翻译模型的编码器进行训练；其中，所述第一损失函数为对比学习损失函数，用于更新所述编码器的参数。

4、根据权利要求3所述的方法，其中，所述使用所述伪平行语料对预设的基础翻译模型进行训练，以获取目标翻译模型，包括：

使用所述伪平行语料，通过所述第一损失函数和第二损失函数对所述基础翻译模型进行多任务训练，以获取所述目标翻译模型；其中，所述第二损失函数用于更新所述编码器和所述解码器的参数。

5、根据权利要求3所述的方法，其中，所述使用所述伪平行语料，通过第一损失函数对所述基础翻译模型的编码器进行训练，包括：

构建所述伪平行语料中的伪源端语料的正例语料和负例语料；

使用所述伪源端语料、所述正例语料以及所述负例语料，通过所述第一损失函数对所述编码器进行训练；其中，训练目标为最大化所述伪源端语料和所述正例语料的向量表示之间的相似度，最小化所述伪源端语料和所述负例语料的向量表示之间的相似度。

6、根据权利要求5所述的方法，其中，所述使用所述伪源端语料、所述正例语料以及所述负例语料，通过所述第一损失函数对所述编码器进行训练，包

括：

将所述伪源端语料、所述正例语料以及所述负例语料输入至所述编码器中，得到所述伪源端语料对应的第一向量表示、所述正例语料对应的第二向量表示以及所述负例语料对应的第三向量表示；

根据所述第一向量表示和所述第二向量表示，确定所述第一损失函数的第一损失值，并基于所述第一损失值更新所述编码器的参数，直至所述第一损失函数的第一损失值满足收敛条件；

根据所述第一向量表示和所述第三向量表示，确定所述第一损失函数的第二损失值，并基于所述第二损失值更新所述编码器的参数，直至所述第一损失函数的第二损失值满足收敛条件。

7、根据权利要求5所述的方法，其中，在所述原始语料为单语语料的情况下，所述基于所述原始语料与所述替换语料构造伪平行语料，包括：

将所述单语语料对应的替换语料作为所述伪源端语料以及将所述单语语料作为伪目标端语料，组成所述伪平行语料。

8、根据权利要求5所述的方法，其中，在所述原始语料为平行语料的情况下，所述基于所述原始语料与所述替换语料构造伪平行语料，包括：

将所述平行语料中源端语料对应的替换语料作为所述伪源端语料，以及将所述平行语料中目标端语料作为伪目标端语料，组成所述伪平行语料。

9、根据权利要求7或8所述的方法，其中，所述正例语料为所述伪目标端语料。

10、根据权利要求7或8所述的方法，其中，所述负例语料为其它伪平行语料中的伪目标端语料。

11、根据权利要求1至8中任一项所述的方法，其中，所述替换语料中的各目标词汇的语种至少部分不同。

12、一种翻译模型的训练装置，包括：

获取模块，设置为获取至少一个原始语料；

替换模块，设置为将所述原始语料中源端语料的至少一个原始词汇对齐替换为同含义的目标词汇，得到所述原始语料对应的替换语料；其中，所述原始词汇与所述目标词汇的语种不同；

构造模块，设置为基于所述原始语料与所述替换语料构造伪平行语料；

训练模块，用于使用所述伪平行语料对预设的基础翻译模型进行训练，以

获取目标翻译模型。

13、一种翻译模型的训练设备，包括存储器和处理器，所述存储器存储有计算机程序，其中，所述处理器执行所述计算机程序时实现权利要求 1 至 11 中任一项所述的翻译模型的训练方法。

14、一种计算机可读存储介质，存储有计算机程序，其中，所述计算机程序被处理器执行时实现权利要求 1 至 11 中任一项所述的翻译模型的训练方法。

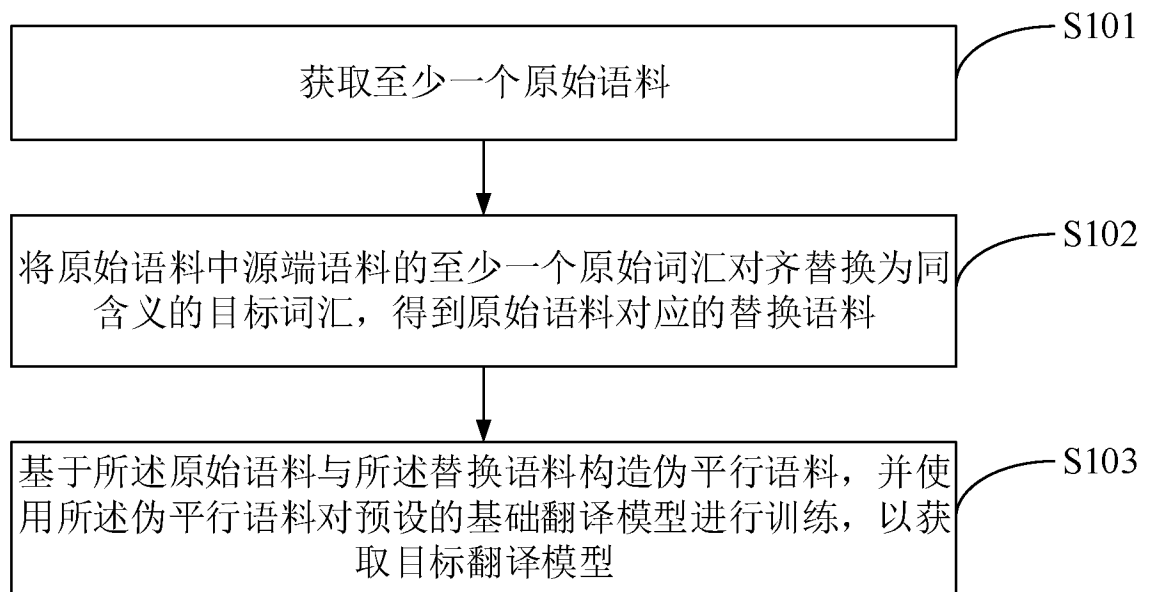


图 1

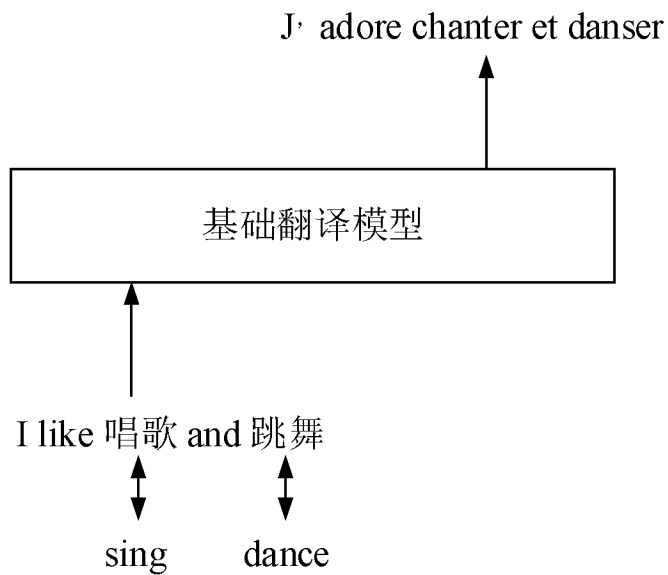


图 2

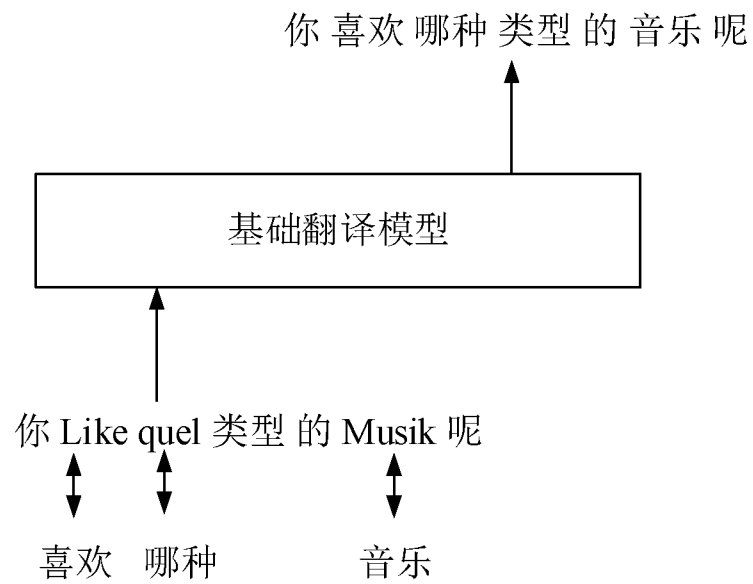


图 3

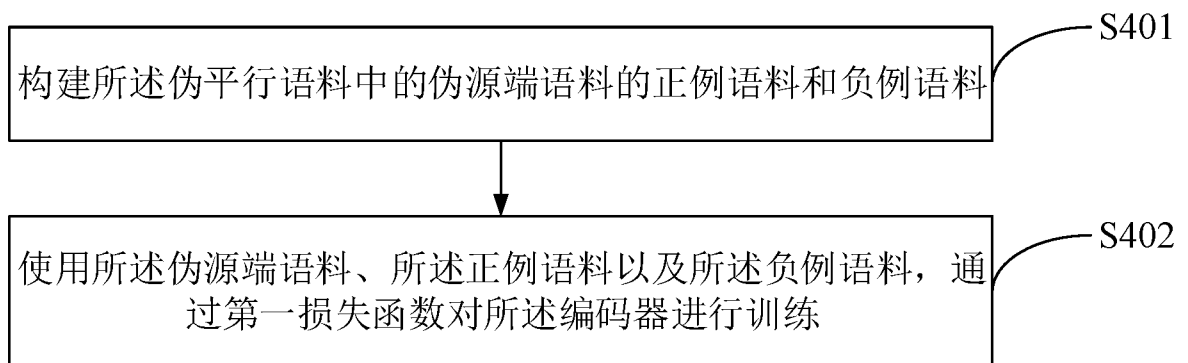


图 4

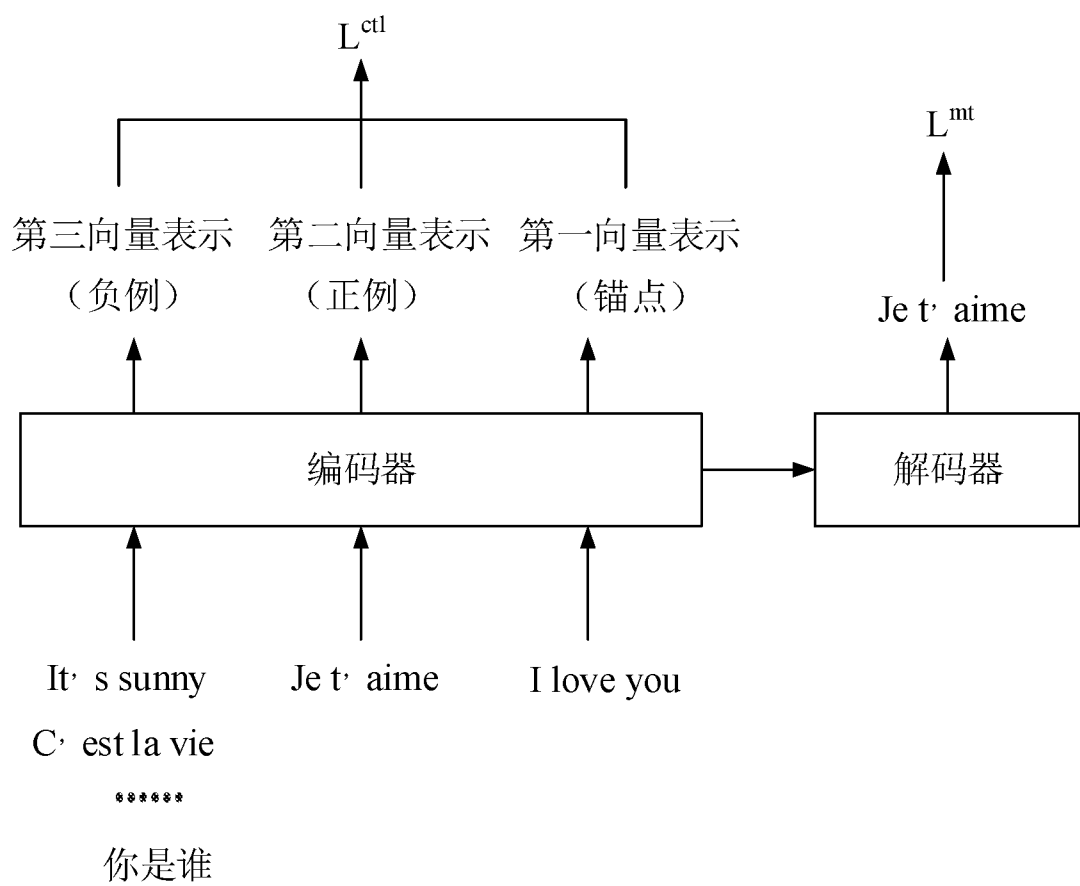


图 5

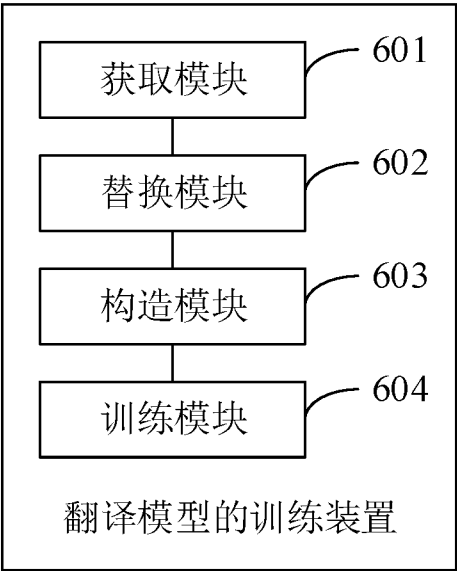


图 6

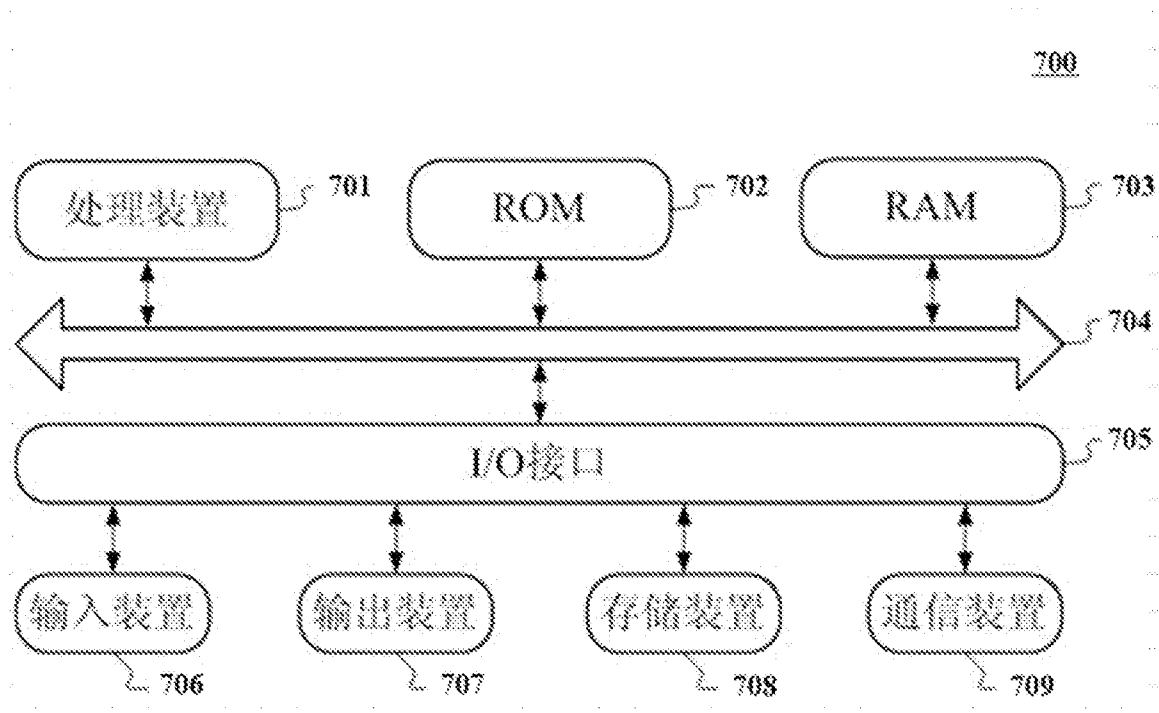


图 7

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/084963

A. CLASSIFICATION OF SUBJECT MATTER

G06F 40/126(2020.01)i; G06F 40/247(2020.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, CNKI, WPI, EPODOC: 损失函数, 翻译, 训练, 正 s 语料, 负 s 语料, 语种, 语言, 平行, 代替, 替代, loss function, translation, training, positive s corpus, negative s corpus, language, parallel, substitute

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 113139391 A (BEIJING YOUZHUJU NETWORK TECHNOLOGY CO., LTD.) 20 July 2021 (2021-07-20) claims 1-14	1-14
X	CN 111046677 A (BEIJING BYTEDANCE NETWORK TECHNOLOGY CO., LTD.) 21 April 2020 (2020-04-21) description paragraphs 0028-0034, and claims 9-10	1-3, 11-14
Y	CN 111046677 A (BEIJING BYTEDANCE NETWORK TECHNOLOGY CO., LTD.) 21 April 2020 (2020-04-21) description, paragraphs 0028-0034	4-10
Y	CN 112417902 A (BEIJING YOUZHUJU NETWORK TECHNOLOGY CO., LTD.) 26 February 2021 (2021-02-26) description, paragraph 0054	4
Y	CN 112668671 A (BEIJING BAIDU NETCOM SCIENCE AND TECHNOLOGY CO., LTD.) 16 April 2021 (2021-04-16) description, paragraph 0019	5-10
A	WO 2021037559 A1 (SONY CORP. et al.) 04 March 2021 (2021-03-04) entire document	1-14



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

06 June 2022

Date of mailing of the international search report

27 June 2022

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/084963

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 110941966 A (BEIJING XIAOMI MOBILE SOFTWARE CO., LTD.) 31 March 2020 (2020-03-31) entire document	1-14

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/084963

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	113139391	A	20 July 2021	None			
CN	111046677	A	21 April 2020	None			
CN	112417902	A	26 February 2021	None			
CN	112668671	A	16 April 2021	None			
WO	2021037559	A1	04 March 2021	None			
CN	110941966	A	31 March 2020	KR	20210073431	A	18 June 2021
				EP	3835998	A1	16 June 2021
				US	2021174019	A1	10 June 2021
				JP	2021093113	A	17 June 2021

国际检索报告

国际申请号

PCT/CN2022/084963

A. 主题的分类

G06F 40/126(2020.01) i; G06F 40/247(2020.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNPAT, CNKI, WPI, EP0D0C: 损失函数, 翻译, 训练, 正 s 语料, 负 s 语料, 语种, 语言, 平行, 代替, 替代, loss function, translation, training, positive s corpus, negative s corpus, language, parallel, substitute

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 113139391 A (北京有竹居网络技术有限公司) 2021年7月20日 (2021 - 07 - 20) 权利要求1-14	1-14
X	CN 111046677 A (北京字节跳动网络技术有限公司) 2020年4月21日 (2020 - 04 - 21) 说明书第0028-0034段, 权利要求9-10	1-3, 11-14
Y	CN 111046677 A (北京字节跳动网络技术有限公司) 2020年4月21日 (2020 - 04 - 21) 说明书第0028-0034段	4-10
Y	CN 112417902 A (北京有竹居网络技术有限公司) 2021年2月26日 (2021 - 02 - 26) 说明书第0054段	4
Y	CN 112668671 A (北京百度网讯科技有限公司) 2021年4月16日 (2021 - 04 - 16) 说明书第0019段	5-10
A	WO 2021037559 A1 (SONY CORP. 等) 2021年3月4日 (2021 - 03 - 04) 全文	1-14
A	CN 110941966 A (北京小米移动软件有限公司) 2020年3月31日 (2020 - 03 - 31) 全文	1-14

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2022年6月6日

国际检索报告邮寄日期

2022年6月27日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

范玉霞

电话号码 86-(10)-53961331

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/084963

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	113139391	A	2021年7月20日	无			
CN	111046677	A	2020年4月21日	无			
CN	112417902	A	2021年2月26日	无			
CN	112668671	A	2021年4月16日	无			
WO	2021037559	A1	2021年3月4日	无			
CN	110941966	A	2020年3月31日	KR	20210073431	A	2021年6月18日
				EP	3835998	A1	2021年6月16日
				US	2021174019	A1	2021年6月10日
				JP	2021093113	A	2021年6月17日