(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 110991196 B
(45)授权公告日 2021.10.26

(21)申请号 201911309770.3
(22)申请日 2019.12.18
(65)同一申请的已公布的文献号
申请公布号 CN 110991196 A
(43)申请公布日 2020.04.10
(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层
(72)发明人 张睿卿 张传强 熊皓 何中军 吴华 李芝 王海峰
(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 石茵汀
(51)Int.Cl.
G06F 40/58(2020.01)
G06F 40/284(2020.01)
(56)对比文件
CN 104572633 A,2015.04.29
CN 109726385 A,2019.05.07
CN 105718443 A,2016.06.29
CN 108920467 A,2018.11.30
WO 2018110096 A1,2018.06.21
审查员 付琦

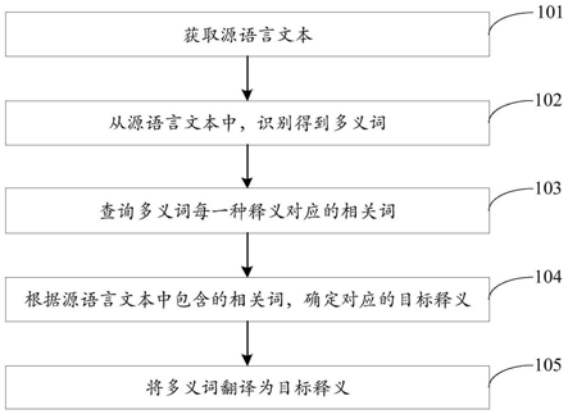
权利要求书2页 说明书11页 附图2页

(54)发明名称

多义词的翻译方法、装置、电子设备及介质

(57)摘要

本申请公开了一种多义词的翻译方法、装置、电子设备及介质，涉及自然语言处理技术领域中的翻译技术领域。具体实现方案为：通过获取源语言文本，从源语言文本中，识别得到多义词，查询多义词每一种释义对应的相关词，根据源语言文本中包含的相关词，确定对应的目标释义，将多义词翻译为目标释义。该方法根据源语言文本中包含的多义词的释义对应的相关词，对多义词进行翻译，实现了根据源语言文本的上下文对多义词进行翻译，确保多义词能够被正确的翻译出来，从而避免错误释义产生的情况。



CN 110991196 B

1.一种多义词的翻译方法,其特征在于,所述方法包括:

获取源语言文本;

从所述源语言文本中,识别得到多义词;

查询所述多义词每一种释义对应的相关词;其中,所述相关词的反文档频率高于其他词且所述相关词与相应释义的关联度高于其他词与所述相应释义的关联度;

根据所述源语言文本中包含的相关词,确定对应的目标释义;

将所述多义词翻译为所述目标释义;

其中,所述查询所述多义词每一种释义对应的相关词之前,还包括:

从语料库的各样本中,确定原文包含所述多义词的目标样本;

根据所述目标样本的原文中除所述多义词以外的词,确定多个候选词;

对每一个所述候选词,确定反文档频率,以及确定与各释义的关联度;其中,所述与各释义的关联度,是指原文中包含所述多义词和相应的候选词,且在对应译文中包含相应释义的概率;

对每一种释义,根据各候选词的反文档频率,以及根据各候选词与相应释义的关联度,从各候选词中,确定相应释义对应的相关词;

其中,所述对每一个所述候选词,确定反文档频率,以及确定与各释义的关联度,包括:

对一个候选词,统计所述语料库,确定原文中包含所述多义词和所述一个候选词,且在相应译文中包含释义Ti的样本个数yi;其中,i为所述多义词的释义序号,取值为1至n的自然数,n为所述多义词的释义总数;

确定原文中包含所述多义词,且在相应译文中包含释义Ti的样本个数Yi;

根据所述样本个数yi与样本个数Yi之间的比值,确定所述一个候选词与释义Ti的关联度。

2.根据权利要求1所述的翻译方法,其特征在于,所述对每一个所述候选词,确定反文档频率,以及确定与各释义的关联度,包括:

对一个候选词,统计所述语料库,确定原文中包含所述一个候选词的样本个数,以及所述语料库中包含的样本总数;

根据所述样本总数与原文中包含所述一个候选词的样本个数之间的比值,确定所述反文档频率。

3.根据权利要求1所述的翻译方法,其特征在于,所述对每一个所述候选词,确定反文档频率,以及确定与各释义的关联度之前,还包括:

对每一个所述候选词,确定各释义的词向量;

根据各释义的词向量之间的相似距离,对相应候选词的各释义进行合并。

4.根据权利要求1-3任一项所述的翻译方法,其特征在于,所述从所述源语言文本中,识别得到多义词,包括:

根据多义词库,从所述源语言文本中,识别得到所述多义词;

其中,所述多义词库是根据各单词的一词多义概率确定的;所述多义词的所述一词多义概率大于设定阈值;

所述一词多义概率包括相应单词e翻译为每一释义Ti的概率P(e|Ti),以及包括每一种释义Ti用于翻译为相应单词e的概率P(Ti|e);i为所述多义词的释义序号,取值为1至n的自

然数,n为所述多义词的释义总数。

5.一种多义词的翻译装置,其特征在于,所述装置包括:

获取模块,用于获取源语言文本;

识别模块,用于从所述源语言文本中,识别得到多义词;

查询模块,用于查询所述多义词每一种释义对应的相关词;其中,所述相关词的反文档频率高于其他词且所述相关词与相应释义的关联度高于其他词与所述相应释义的关联度;

确定模块,用于根据所述源语言文本中包含的相关词,确定对应的目标释义;

翻译模块,用于将所述多义词翻译为所述目标释义;

其中,所述装置,还包括:

样本处理模块,用于从语料库的各样本中,确定原文包含所述多义词的目标样本;

选择模块,用于根据所述目标样本的原文中除所述多义词以外的词,确定多个候选词;

计算模块,用于对每一个所述候选词,确定反文档频率,以及确定与各释义的关联度;其中,所述与各释义的关联度,是指原文中包含所述多义词和相应的候选词,且在对应译文中包含相应释义的概率;

关联模块,用于对每一种释义,根据各候选词的反文档频率,以及根据各候选词与相应释义的关联度,从各候选词中,确定相应释义对应的相关词;

其中,所述计算模块,还用于:

对一个候选词,统计所述语料库,确定原文中包含所述多义词和所述一个候选词,且在相应译文中包含释义Ti的样本个数yi;其中,i为所述多义词的释义序号,取值为1至n的自然数,n为所述多义词的释义总数;

确定原文中包含所述多义词,且在相应译文中包含释义Ti的样本个数Yi;

根据所述样本个数yi与样本个数Yi之间的比值,确定所述一个候选词与释义Ti的关联度。

6.一种电子设备,其特征在于,包括:

至少一个处理器;以及

与所述至少一个处理器通信连接的存储器;其中,

所述存储器存储有可被所述至少一个处理器执行的指令,所述指令被所述至少一个处理器执行,以使所述至少一个处理器能够执行权利要求1-4中任一项所述的多义词的翻译方法。

7.一种存储有计算机指令的非瞬时计算机可读存储介质,其特征在于,所述计算机指令用于使所述计算机执行权利要求1-4中任一项所述的多义词的翻译方法。

多义词的翻译方法、装置、电子设备及介质

技术领域

[0001] 本申请涉及自然语言处理技术领域中的翻译技术领域，尤其涉及一种多义词的翻译方法、装置、电子设备及介质。

背景技术

[0002] 随着人工智能的快速发展，在翻译领域，也出现了许多类型的翻译机器，解决了人工翻译效率低的问题。由于大量的词汇都具有多义性，对于此类词汇的准确翻译，即便是具有扎实的语言和专业基础的翻译人员也难免在翻译过程中出现疏漏。

[0003] 在实际使用翻译机器翻译文章时，依然会出现多义词翻译错误的情况，从而导致文章翻译的准确率较低。

发明内容

[0004] 本申请提出一种多义词的翻译方法，解决了相关技术中多义词翻译准确率低的技术问题。

[0005] 本申请第一方面实施例提出了一种多义词的翻译方法，包括：

[0006] 获取源语言文本；

[0007] 从所述源语言文本中，识别得到多义词；

[0008] 查询所述多义词每一种释义对应的相关词；

[0009] 根据所述源语言文本中包含的相关词，确定对应的目标释义；

[0010] 将所述多义词翻译为所述目标释义。

[0011] 作为本申请实施例的第一种可能的实现方式，所述查询所述多义词每一种释义对应的相关词之前，还包括：

[0012] 从语料库的各样本中，确定原文包含所述多义词的目标样本；

[0013] 根据所述目标样本的原文中除所述多义词以外的词，确定多个候选词；

[0014] 对每一个所述候选词，确定反文档频率，以及确定与各释义的关联度；其中，所述与各释义的关联度，是指原文中包含所述多义词和相应的候选词，且在对应译文中包含相应释义的概率；

[0015] 对每一种释义，根据各候选词的反文档频率，以及根据各候选词与相应释义的关联度，从各候选词中，确定相应释义对应的相关词。

[0016] 作为本申请实施例的第二种可能的实现方式，所述对每一个所述候选词，确定反文档频率，以及确定与各释义的关联度，包括：

[0017] 对一个候选词，统计所述语料库，确定原文中包含所述多义词和所述一个候选词，且在相应译文中包含释义Ti的样本个数yi；其中，i为所述多义词的释义序号，取值为1至n的自然数，n为所述多义词的释义总数；

[0018] 确定原文中包含所述多义词，且在相应译文中包含释义Ti的样本个数Yi；

[0019] 根据所述训练样本个数yi与训练样本个数Yi之间的比值，确定所述一个候选词与

释义Ti的关联度。

[0020] 作为本申请实施例的第三种可能的实现方式，所述对每一个所述候选词，确定反文档频率，以及确定与各释义的关联度，包括：

[0021] 对一个候选词，统计所述语料库，确定原文中包含所述一个候选词的样本个数，以及所述语料库中包含的样本总数；

[0022] 根据所述样本总数与原文中包含所述一个候选词的样本个数之间的比值，确定所述反文档频率。

[0023] 作为本申请实施例的第四种可能的实现方式，所述对每一个所述候选词，确定反文档频率，以及确定与各释义的关联度之前，还包括：

[0024] 对每一个所述候选词，确定各释义的词向量；

[0025] 根据各释义的词向量之间的相似距离，对相应候选词的各释义进行合并。

[0026] 作为本申请实施例的第五种可能的实现方式，所述从所述源语言文本中，识别得到多义词，包括：

[0027] 根据多义词库，从所述源语言文本中，识别得到所述多义词；

[0028] 其中，所述多义词库是根据各单词的一词多义概率确定的；所述多义词所述一词多义概率大于设定阈值；

[0029] 所述一词多义概率包括相应单词e翻译为每一释义Ti的概率P(e|Ti)，以及包括每一种释义Ti用于翻译为相应单词e的概率P(Ti|e)；其中，i为所述多义词的释义序号，取值为1至n的自然数，n为所述多义词的释义总数。

[0030] 本申请第二方面实施例提出了一种多义词的翻译装置，包括：

[0031] 获取模块，用于获取源语言文本；

[0032] 识别模块，用于从所述源语言文本中，识别得到多义词；

[0033] 查询模块，用于查询所述多义词每一种释义对应的相关词；

[0034] 确定模块，用于根据所述源语言文本中包含的相关词，确定对应的目标释义；

[0035] 翻译模块，用于将所述多义词翻译为所述目标释义。

[0036] 本申请第三方面实施例提出了一种电子设备，包括：

[0037] 至少一个处理器；以及

[0038] 与所述至少一个处理器通信连接的存储器；其中，

[0039] 所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行上述实施例中所述的多义词的翻译方法。

[0040] 本申请第四方面实施例提出了一种存储有计算机指令的非瞬时计算机可读存储介质，所述计算机指令用于使所述计算机执行上述实施例中所述的多义词的翻译方法。

[0041] 上述申请中的一个实施例具有如下优点或有益效果：通过获取源语言文本，从源语言文本中，识别得到多义词，查询多义词每一种释义对应的相关词，根据源语言文本中包含的相关词，确定对应的目标释义，将多义词翻译为目标释义。该方法根据源语言文本中包含的多义词的释义对应的相关词，对多义词进行翻译，实现了根据源语言文本的上下文对多义词进行翻译，确保多义词能够被正确的翻译出来，从而避免错误释义产生的情况。

[0042] 上述可选方式所具有的其他效果将在下文中结合具体实施例加以说明。

附图说明

[0043] 附图用于更好地理解本方案,不构成对本申请的限定。其中:

[0044] 图1为本申请实施例提供的一种多义词的翻译方法的流程示意图;

[0045] 图2为本申请实施例提供的另一种多义词的翻译方法的流程示意图;

[0046] 图3为本申请实施例提供的又一种多义词的翻译方法的流程示意图;

[0047] 图4为本申请实施例提供的一种多义词的翻译装置的结构示意图;

[0048] 图5是用来实现本申请实施例的多义词的翻译方法的电子设备的框图。

具体实施方式

[0049] 以下结合附图对本申请的示范性实施例做出说明,其中包括本申请实施例的各种细节以助于理解,应当将它们认为仅仅是示范性的。因此,本领域普通技术人员应当认识到,可以对这里描述的实施例做出各种改变和修改,而不会背离本申请的范围和精神。同样,为了清楚和简明,以下的描述中省略了对公知功能和结构的描述。

[0050] 为了解决相关技术中对于文本中存在的多义词进行翻译时,翻译的准确率低的技术问题,本申请提出了一种多义词的翻译方法,通过获取源语言文本,从源语言文本中,识别得到多义词,查询多义词每一种释义对应的相关词,根据源语言文本中包含的相关词,确定对应的目标释义,将多义词翻译为目标释义。

[0051] 下面结合参考附图描述本申请实施例的行列式文本的存储方法、装置以及电子设备。

[0052] 图1为本申请实施例所提供的一种多义词的翻译方法的流程示意图。

[0053] 本申请实施例以该多义词的翻译方法被配置于多义词的翻译装置中来举例说明,该多义词的翻译装置可以应用于任一电子设备中,以使该电子设备可以执行多义词的翻译功能。

[0054] 其中,电子设备可以为个人电脑(Personal Computer,简称PC)、云端设备、移动设备等,移动设备例如可以为手机、平板电脑、个人数字助理、穿戴式设备、车载设备等具有各种操作系统的硬件设备。

[0055] 如图1所示,该多义词的翻译方法可以包括以下步骤:

[0056] 步骤101,获取源语言文本。

[0057] 在一种可能的情况下,源语言文本,可以为用户输入的源语言文本,例如,用户手动输入的源语言文本,或者通过语音的方式输入的源语言文本,等等,本申请实施例中对用户输入源语言文本的方式不做限定。

[0058] 在另一种可能的情况下,源语言文本,还可以为图片中的文本。例如,电子设备通过摄像头采集的图片中包含的文本,或者从服务器下载的图片中包含的文本,等等。

[0059] 本申请实施例中,源语言文本,为待翻译文本,如源语言文本为英语文本。当然也可以为其他语言的文本,在此不做限制。

[0060] 步骤102,从源语言文本中,识别得到多义词。

[0061] 其中,多义词,是具有两个或两个以上意思的词。例如,源语言文本是英语时,shot可以翻译为射击、开枪、镜头、照片,等等。

[0062] 本申请实施例中,获取到源语言文本后,可以对源语言文本进行识别,以得到多义

词。

[0063] 作为一种可能的实现方式，可以根据多义词库从源语言文本中识别得到多义词。其中，根据多义词库，是根据各单词的一词多义概率确定的；多义词的一词多义概率大于设定阈值。

[0064] 其中，一词多义概率包括相应单词e翻译为每一释义Ti的概率P(e|Ti)，以及包括每一种释义Ti用于翻译为相应单词e的概率P(Ti|e)。其中，i为多义词的释义序号，取值为1至n的自然数，n为多义词的释义总数。

[0065] 具体地，首先提取多义词库的短语表，进而根据短语表对源语言文本进行过滤，得到多义词。例如，可以根据多义词的一词多义概率对短语表进行筛选，得到一词多义概率大于设定阈值的多义词。

[0066] 需要说明的是，从源语言文本中，识别得到的多义词不限于一个，可以识别得到源语言文本中所有的多义词。

[0067] 步骤103，查询多义词每一种释义对应的相关词。

[0068] 其中，相关词，是指横向关联是存在并列概念的词。例如，国庆节与10月1日，电影和镜头。

[0069] 本申请实施例中，从源语言文本中识别得到多义词后，在语料库中查询得到多义词每一种释义对应的相关词。

[0070] 举例来说，源语言文本为英文文本，例如“A modern movie have something along the lines of three thousand shots.Each one of these shots are a few seconds long.But it would take designers the whole time of film making to create these shots”。从源语言文本中识别得到的多义词为“shot”，对应的释义可以为射击、开枪、镜头、照片、射中，等等。镜头这一种释义对应的相关词可以为movie、film。射击、开枪这一释义对应的相关词可以为murder dead gun。

[0071] 作为一种可能的实现方式，可以从语料库的各样本中，确定原文包括多义词的目标样本，根据目标样本的原文中除多义词以外的词，确定多个候选词，确定每一个候选词语各释义的关联度，进而从各候选词中，确定语料库中包含的相应释义对应的相关词。进而，在相应释义对应的相关词中查询多义词每一种释义对应的相关词。

[0072] 步骤104，根据源语言文本中包含的相关词，确定对应的目标释义。

[0073] 其中，目标释义，为多义词在源语言文本中对应的释义。例如。多义词shot，对应的释义可以为射击、开枪、镜头、照片、射中，根据源语言文本中包含的相关词，确定对应的目标释义为镜头。

[0074] 本申请实施例，确定多义词的每一种释义对应的相关词后，根据源语言文本中包含的相关词，确定对应的目标释义。

[0075] 具体地，确定多义词的每一种释义对应的相关词后，查询源语言文本中包含的每一种释义对应的相关词，确定源语言文本中存在其中一种释义对应的相关词后，根据源语言文本中包含的相关词，可以确定多义词对应的目标释义。

[0076] 步骤105，将多义词翻译为目标释义。

[0077] 本申请实施例中，根据源语言文本中包含的相关词，确定对应的目标释义后，进而将多义词翻译为目标释义。

[0078] 作为一种可能的实现方式，在对整篇源语言文本进行翻译时，从源语言文本中识别得到多义词及其释义对应的相关词后，将多义词翻译为根据源语言文本中包含的相关词确定的目标释义。

[0079] 本申请实施例的多义词的翻译方法，通过获取源语言文本，从源语言文本中，识别得到多义词，查询多义词每一种释义对应的相关词，根据源语言文本中包含的相关词，确定对应的目标释义，将多义词翻译为目标释义。该方法根据源语言文本中包含的多义词的释义对应的相关词，对多义词进行翻译，实现了根据源语言文本的上下文对多义词进行翻译，确保多义词能够被正确的翻译出来，从而避免错误释义产生的情况。

[0080] 在上述实施例的基础上，在上述步骤103中查询多义词每一种释义对应的相关词之前，需要从语料库中确定每一种释义对应的相关词。下面结合图2对上述过程进行详细介绍，图2为本申请实施例提供的另一种多义词的翻译方法的流程示意图。

[0081] 如图2所示，该翻译方法还可以包括以下步骤：

[0082] 步骤201，从语料库的各样本中，确定原文包含多义词的目标样本。

[0083] 其中，语料库中存放的是在语言的实际使用中真实出现过的语言材料。

[0084] 本申请实施例中，语料库中包含有各种文本样本，在从源语言文本中识别得到多义词后，可以进一步的，从语料库的各样本中，确定各样本的原文包括多义词的目标样本。

[0085] 可以理解为，将语料库的各样本的原文中包含多义词的样本称为目标样本。

[0086] 步骤202，根据目标样本的原文中除多义词以外的词，确定多个候选词。

[0087] 本申请实施例中，从语料库中确定包含多义词的目标样本后，可以进一步的，根据目标样本的原文中除多义词以外的词，确定多个候选词。

[0088] 举例来说，目标样本的原文中有20个词，其中一个为多义词，在目标样本中除多义词以外的19个词中确定多个候选词。例如，确定3个候选词。

[0089] 步骤203，对每一个候选词，确定反文档频率，以及确定与各释义的关联度。

[0090] 其中，与各释义的关联度，是指原文中包含多义词和相应的候选词，且在对应译文中包含相应释义的概率。

[0091] 需要解释的是，词频-反文档频率(term frequency-inverse document frequency，以下简称TF-IDF)是一种用于资讯检索与资讯探勘的常用加权技术。TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

[0092] 本申请实施例中，对于每一个候选词，计算各候选词的反文档频率。

[0093] 作为一种可能的实现方式，对一个候选词，统计语料库，确定原文中包含该候选词的样本个数，以及语料库中包含的样本总数。根据样本总数与原文中包含该候选词的样本个数之间的比值，确定反文档频率。

[0094] 举例来说，假设候选词A，语料库中包括的样本总数为20万条，语料库中各样本的原文中包含候选词A的样本个数为1万条，样本总数与原文中包含候选词A的样本个数之间的比值为200000/10000。由此，可以确定候选词A的反文档频率IDF＝log(200000/10000)。

[0095] 本申请实施例中，对于每一个候选词，确定与各释义的关联度。也就说，确定原文中包含多义词和相应的候选词，且在对应译文中包含相应释义的概率。

[0096] 作为一种可能的实现方式，对一个候选词，统计语料库，确定原文中包含多义词和该候选词，且在相应译文中包含释义Ti的样本个数yi。其中，i为所述多义词的释义序号，取值为1至n的自然数，n为多义词的释义总数。同时，确定原文中包含多义词，且在相应译文中包含释义Ti的样本个数Yi。进而，根据训练样本个数yi与训练样本个数Yi之间的比值，确定该候选词与释义Ti的关联度。

[0097] 继续以上述示例为例，对于候选词A，统计语料库，确定原文中包含多义词B和候选词A，且在相应译文中包含释义C的样本个数为yi为150条。确定原文中包括多义词B，且在相应译文中包含释义C的样本个数Yi为200条。计算得到训练样本个数yi与训练样本个数Yi之间的比值为150/200。进而可以确定候选词A与释义C的关联度为150/200。

[0098] 需要说明的是，对于从目标样本中确定的多个候选词，均可以根据上述方法确定每一个候选词的反文档频率以及与各释义的关联度。

[0099] 步骤204，对每一种释义，根据各候选词的反文档频率，以及根据各候选词与相应释义的关联度，从各候选词中，确定相应释义对应的相关词。

[0100] 本申请实施例中，对每一个候选词，确定反文档频率以及与各释义的关联度后，可以从多个候选词中，确定相应释义对应的相关词。

[0101] 可以理解为，可以根据各候选词的反文档频率以及与各释义的关联度对各候选词进行筛选，以将候选词中反文档频率以及与相应释义的关联度最高的候选词作为相应释义的相关词。

[0102] 本申请实施例的多义词翻译方法，通过从语料库的各样本中，确定原文包含多义词的目标样本，根据目标样本的原文中除多义词以外的词，确定多个候选词，对每一个候选词，确定反文档频率，以及确定与各释义的关联度，对每一种释义，根据各候选词的反文档频率，以及根据各候选词与相应释义的关联度，从各候选词中，确定相应释义对应的相关词。由此，在语料库中包含多义词的目标样本中确定多个候选词，以从多个候选词中确定相应释义对应的相关词，以实现对整篇文章进行句内多义词翻译，进而提高多义词翻译的准确率。

[0103] 在上述实施例的基础上，在上述步骤203中对每一个候选词，确定反文档频率，以及确定与各释义的关联度之前，还可以对每一个候选词确定对应各释义的词向量，以根据各释义的词向量之间的相似距离，对相应候选词的各释义进行合并。由此，通过对多义词的释义进行合并，能够提高文本翻译的速率。下面结合图3对上述过程进行详细介绍，图3为本申请实施例提供的又一种多义词的翻译方法的流程示意图。

[0104] 如图3所示，该多义词的翻译方法还可以包括以下步骤:

[0105] 步骤301，对每一个候选词，确定各释义的词向量。

[0106] 其中，词向量(Word embedding)，又叫Word嵌入是自然语言处理(NLP)中的一组语言建模和特征学习技术的统称，其中来自词汇表的单词或短语被映射到实数的向量。

[0107] 需要说明的是，对于相似的词，其对应的词向量也相近。

[0108] 本申请实施例中，从语料库中原文包含多义词的目标样本中确定多个候选词后，每一个候选词对应于多义词的一个或多个释义。对每一个候选词，确定各释义的词向量。

[0109] 作为一种可能的实现方式，可以基于语言模型的方法确定各释义的词向量，具体地，将每一个候选词的各释义输入训练神经网络语言模型(Neural Network Language

Model,简称NNLM),词向量作为语言模型的附带产出。

[0110] 步骤302,根据各释义的词向量之间的相似距离,对相应候选词的各释义进行合并。

[0111] 其中,各释义的词向量之间的相似距离,是指各释义之间的相似度。

[0112] 例如,射击和开枪对应的词向量之间的相似距离,大于射击和镜头之间的相似距离。

[0113] 本申请实施例中,对于每一个候选词,确定各释义的词向量后,计算各释义的词向量之间的相似距离,以将相同意思的各释义进行合并。

[0114] 作为一种可能的实现方式,可以采用计算各释义的词向量的欧式距离的方法,确定各释义的词向量之间的相似距离。例如,采用如下公式计算各释义的词向量之间的相似距离。

[0115] $$d = \sqrt{\sum_{i=1}^{N}(X_{1i} - X_{2i})}$$

[0116] 其中,d为释义的词向量之间的相似距离,X_{1i}与X_{2i}分别为两个释义对应的词向量。

[0117] 作为另一种可能的实现方式,还可以通过计算各释义的词向量之间的夹角余弦值来评估各释义的词向量之间的相似距离。

[0118] 本申请实施例的多义词翻译方法,通过对每一个候选词,确定各释义的词向量,根据各释义的词向量之间的相似距离,对相应候选词的各释义进行合并。由此,通过对多义词的释义进行合并,能够提高文本翻译的速率。

[0119] 为了实现上述实施例,本申请提出了一种多义词的翻译装置。

[0120] 图4为本申请实施例提出的一种多义词翻译装置的结构示意图。

[0121] 如图4所示,该多义词的翻译装置400,可以包括:获取模块410、识别模块420、查询模块430、确定模块440以及翻译模块450。

[0122] 获取模块410,用于获取源语言文本。

[0123] 识别模块420,用于从源语言文本中,识别得到多义词。

[0124] 查询模块430,用于查询多义词每一种释义对应的相关词。

[0125] 确定模块440,用于根据源语言文本中包含的相关词,确定对应的目标释义。

[0126] 翻译模块450,用于将多义词翻译为目标释义。

[0127] 作为一种可能的情况,该多义词的翻译装置400,还可以包括:

[0128] 样本处理模块,用于从语料库的各样本中,确定原文包含多义词的目标样本。

[0129] 选择模块,用于根据目标样本的原文中除所述多义词以外的词,确定多个候选词。

[0130] 计算模块,用于对每一个候选词,确定反文档频率,以及确定与各释义的关联度;其中,与各释义的关联度,是指原文中包含多义词和相应的候选词,且在对应译文中包含相应释义的概率。

[0131] 关联模块,用于对每一种释义,根据各候选词的反文档频率,以及根据各候选词与相应释义的关联度,从各候选词中,确定相应释义对应的相关词。

[0132] 作为另一种可能的情况,计算模块,还可以用于:

[0133] 对一个候选词,统计语料库,确定原文中包含多义词和一个候选词,且在相应译文中包含释义Ti的样本个数yi;其中,i为多义词的释义序号,取值为1至n的自然数,n为多义

词的释义总数；

[0134] 确定原文中包含多义词，且在相应译文中包含释义Ti的样本个数Yi；

[0135] 根据训练样本个数yi与训练样本个数Yi之间的比值，确定一个候选词与释义Ti的关联度。

[0136] 作为另一种可能的情况，计算模块，还可以用于：

[0137] 对一个候选词，统计语料库，确定原文中包含一个候选词的样本个数，以及语料库中包含的样本总数；

[0138] 根据样本总数与原文中包含一个候选词的样本个数之间的比值，确定反文档频率。

[0139] 作为另一种可能的情况，该多义词的翻译装置400，还可以包括：

[0140] 词向量确定模块，用于对每一个所述候选词，确定各释义的词向量；

[0141] 合并模块，用于根据各释义的词向量之间的相似距离，对相应候选词的各释义进行合并。

[0142] 作为另一种可能的情况，识别模块420，还可以用于：

[0143] 根据多义词库，从所述源语言文本中，识别得到所述多义词；

[0144] 其中，所述多义词库是根据各单词的一词多义概率确定的；所述多义词所述一词多义概率大于设定阈值；

[0145] 所述一词多义概率包括相应单词e翻译为每一释义Ti的概率P(e|Ti)，以及包括每一种释义Ti用于翻译为相应单词e的概率P(Ti|e)；其中，i为所述多义词的释义序号，取值为1至n的自然数，n为所述多义词的释义总数。

[0146] 需要说明的是，前述对多义词的翻译方法实施例的解释说明也适用于该多义词的翻译装置，此处不再赘述。

[0147] 本申请实施例的多义词的翻译装置，通过获取源语言文本，从源语言文本中，识别得到多义词，查询多义词每一种释义对应的相关词，根据源语言文本中包含的相关词，确定对应的目标释义，将多义词翻译为目标释义。该方法根据源语言文本中包含的多义词的释义对应的相关词，对多义词进行翻译，实现了根据源语言文本的上下文对多义词进行翻译，确保多义词能够被正确的翻译出来，从而避免错误释义产生的情况。

[0148] 为了实现上述实施例，本申请实施例提出了一种计算机设备，包括：

[0149] 至少一个处理器；以及

[0150] 与所述至少一个处理器通信连接的存储器；其中，

[0151] 所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行上述实施例中所述的多义词的翻译方法。

[0152] 为了实现上述实施例，本申请实施例提出了一种存储有计算机指令的非瞬时计算机可读存储介质，所述计算机指令用于使所述计算机执行上述实施例中所述的多义词的翻译方法。

[0153] 根据本申请的实施例，本申请还提供了一种电子设备和一种可读存储介质。

[0154] 如图5所示，是根据本申请实施例的多义词的翻译方法的电子设备的框图。电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助

理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本申请的实现。

[0155] 如图5所示，该电子设备包括：一个或多个处理器501、存储器502，以及用于连接各部件的接口，包括高速接口和低速接口。各个部件利用不同的总线互相连接，并且可以被安装在公共主板上或者根据需要以其它方式安装。处理器可以对在电子设备内执行的指令进行处理，包括存储在存储器中或者存储器上以在外部输入/输出装置(诸如，耦合至接口的显示设备)上显示GUI的图形信息的指令。在其它实施方式中，若需要，可以将多个处理器和/或多条总线与多个存储器和多个存储器一起使用。同样，可以连接多个电子设备，各个设备提供部分必要的操作(例如，作为服务器阵列、一组刀片式服务器、或者多处理器系统)。图5中以一个处理器501为例。

[0156] 存储器502即为本申请所提供的非瞬时计算机可读存储介质。其中，所述存储器存储有可由至少一个处理器执行的指令，以使所述至少一个处理器执行本申请所提供的多义词的翻译方法。本申请的非瞬时计算机可读存储介质存储计算机指令，该计算机指令用于使计算机执行本申请所提供的多义词的翻译方法。

[0157] 存储器502作为一种非瞬时计算机可读存储介质，可用于存储非瞬时软件程序、非瞬时计算机可执行程序以及模块，如本申请实施例中的多义词的翻译的方法对应的程序指令/模块(例如，附图4所示的获取模块410、识别模块420、查询模块430、确定模块440和翻译模块450)。处理器501通过运行存储在存储器502中的非瞬时软件程序、指令以及模块，从而执行服务器的各种功能应用以及数据处理，即实现上述方法实施例中的多义词的翻译方法。

[0158] 存储器502可以包括存储程序区和存储数据区，其中，存储程序区可存储操作系统、至少一个功能所需要的应用程序；存储数据区可存储根据电子设备的使用所创建的数据等。此外，存储器502可以包括高速随机存取存储器，还可以包括非瞬时存储器，例如至少一个磁盘存储器件、闪存器件、或其他非瞬时固态存储器件。在一些实施例中，存储器502可选包括相对于处理器501远程设置的存储器，这些远程存储器可以通过网络连接至多义词的翻译电子设备。上述网络的实例包括但不限于互联网、企业内部网、局域网、移动通信网及其组合。

[0159] 多义词的翻译方法的电子设备还可以包括：输入装置503和输出装置504。处理器501、存储器502、输入装置503和输出装置504可以通过总线或者其他方式连接，图5中以通过总线连接为例。

[0160] 输入装置503可接收输入的数字或字符信息，以及产生与多义词的翻译电子设备的用户设置以及功能控制有关的键信号输入，例如触摸屏、小键盘、鼠标、轨迹板、触摸板、指示杆、一个或者多个鼠标按钮、轨迹球、操纵杆等输入装置。输出装置504可以包括显示设备、辅助照明装置(例如，LED)和触觉反馈装置(例如，振动电机)等。该显示设备可以包括但不限于，液晶显示器(LCD)、发光二极管(LED)显示器和等离子体显示器。在一些实施方式中，显示设备可以是触摸屏。

[0161] 此处描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系

统、专用ASIC(专用集成电路)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括:实施在一个或者多个计算机程序中,该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释,该可编程处理器可以是专用或者通用可编程处理器,可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令,并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0162] 这些计算程序(也称作程序、软件、软件应用、或者代码)包括可编程处理器的机器指令,并且可以利用高级过程和/或面向对象的编程语言、和/或汇编/机器语言来实施这些计算程序。如本文使用的,术语"机器可读介质"和"计算机可读介质"指的是用于将机器指令和/或数据提供给可编程处理器的任何计算机程序产品、设备、和/或装置(例如,磁盘、光盘、存储器、可编程逻辑装置(PLD)),包括,接收作为机器可读信号的机器指令的机器可读介质。术语"机器可读信号"指的是用于将机器指令和/或数据提供给可编程处理器的任何信号。

[0163] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0164] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)和互联网。

[0165] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。

[0166] 根据本申请实施例的技术方案,通过获取源语言文本,从源语言文本中,识别得到多义词,查询多义词每一种释义对应的相关词,根据源语言文本中包含的相关词,确定对应的目标释义,将多义词翻译为目标释义。由此,提高了多义词翻译的准确度。该方法根据源语言文本中包含的多义词的释义对应的相关词,对多义词进行翻译,实现了根据源语言文本的上下文对多义词进行翻译,确保多义词能够被正确的翻译出来,从而避免错误释义产生的情况。

[0167] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本发申请中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行,只要能够实现本申请公开的技术方案所期望的结果,本文在此不进行限制。

[0168] 上述具体实施方式,并不构成对本申请保护范围的限制。本领域技术人员应该明

白的是，根据设计要求和其他因素，可以进行各种修改、组合、子组合和替代。任何在本申请的精神和原则之内所作的修改、等同替换和改进等，均应包含在本申请保护范围之内。

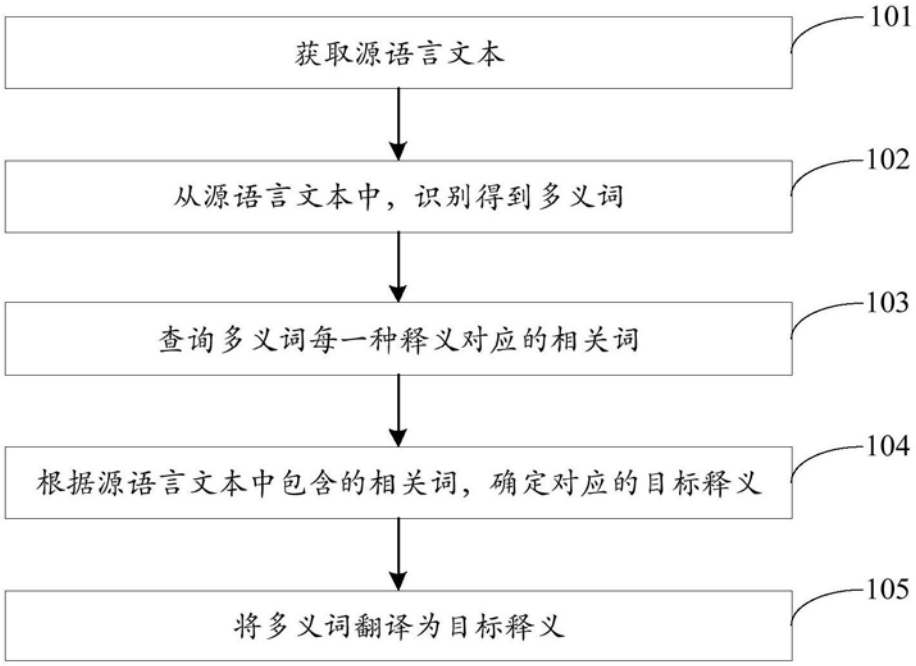

获取源语言文本
101
从源语言文本中，识别得到多义词
102
查询多义词每一种释义对应的相关词
103
根据源语言文本中包含的相关词，确定对应的目标释义
104
将多义词翻译为目标释义
105

图1

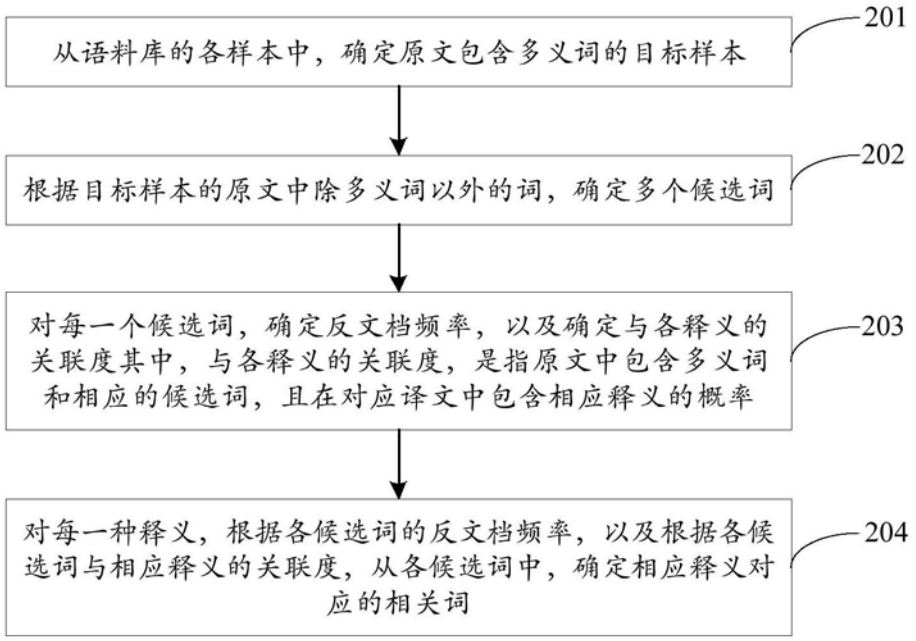

从语料库的各样本中，确定原文包含多义词的目标样本
201
根据目标样本的原文中除多义词以外的词，确定多个候选词
202
对每一个候选词，确定反文档频率，以及确定与各释义的关联度其中，与各释义的关联度，是指原文中包含多义词和相应的候选词，且在对应译文中包含相应释义的概率
203
对每一种释义，根据各候选词的反文档频率，以及根据各候选词与相应释义的关联度，从各候选词中，确定相应释义对应的相关词
204

图2

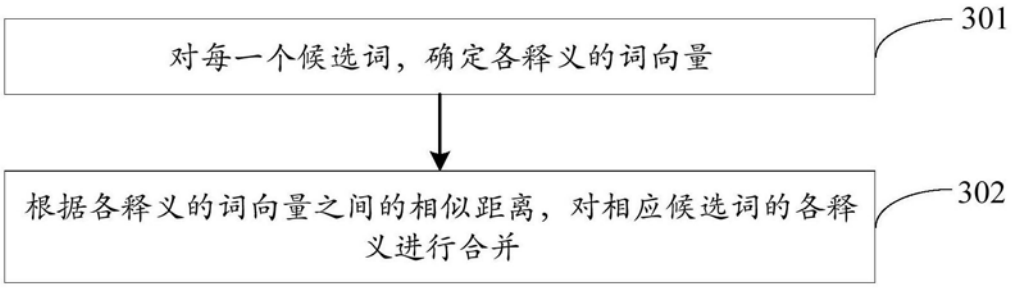

对每一个候选词，确定各释义的词向量
301
根据各释义的词向量之间的相似距离，对相应候选词的各释义进行合并
302

图3


多义词的翻译装置400
获取模块410
识别模块420
查询模块430
确定模块440
翻译模块450

图4

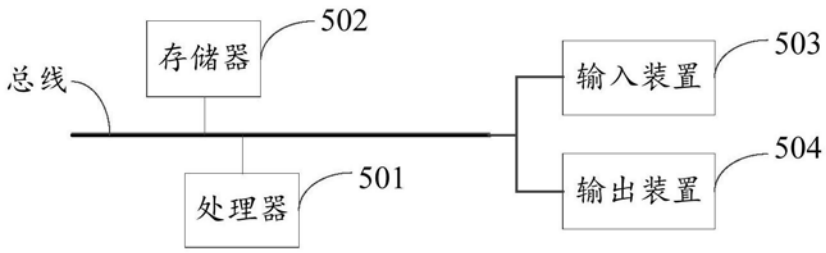

总线
存储器
502
处理器
501
输入装置
503
输出装置
504

图5