(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 103782291 B
(45)授权公告日 2017.06.23

(21)申请号 201280036760.7
(22)申请日 2012.07.17
(65)同一申请的已公布的文献号
申请公布号 CN 103782291 A
(43)申请公布日 2014.05.07
(30)优先权数据
13/190,962 2011.07.26 US
(85)PCT国际申请进入国家阶段日
2014.01.24
(86)PCT国际申请的申请数据
PCT/US2012/047049 2012.07.17
(87)PCT国际申请的公布数据
WO2013/016071 EN 2013.01.31
(73)专利权人 国际商业机器公司
地址 美国纽约
(72)发明人 赵兵 V·卡斯泰利
(74)专利代理机构 北京市中咨律师事务所 11247
代理人 张亚非 于静
(51)Int.Cl.
G06F 17/27(2006.01)
(56)对比文件
US 2004215457 A1,2004.10.28,
CN 1591415 A,2005.03.09,
US 2004236580 A1,2004.11.25,
CN 1770107 A,2006.05.10,
CN 102084417 A,2011.06.01,
CN 101493830 A,2009.07.29,
审查员 赵会玲

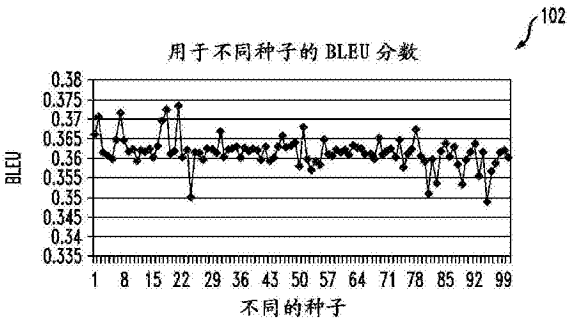
权利要求书3页 说明书9页 附图5页

(54)发明名称

定制自然语言处理引擎

(57)摘要

一种用于定制自然语言处理引擎的方法、装置和制造品。所述方法包括:使能选择希望的自然语言处理任务的一个或多个参数,所述一个或多个参数旨在由训练过的用户或未训练过的用户使用;将所述一个或多个选择的参数映射到优化算法的输入参数的一个或多个区间的集合;以及将具有所述输入参数的一个或多个区间的集合的所述优化算法应用于由自然语言处理引擎所使用的模型,以产生定制模型。



CN 103782291 B

1.一种用于定制自然语言处理引擎的方法,其中所述方法包含:

使能选择希望的自然语言处理任务的一个或多个参数,所述一个或多个参数旨在由训练过的用户和未训练过的用户使用;

使用可训练的映射方法将所述一个或多个选择的参数映射到优化算法的输入参数的一个或多个区间的集合;以及

将具有所述输入参数的一个或多个区间的集合的所述优化算法应用于由自然语言处理引擎所使用的模型,以产生定制模型;

其中由计算机设备实现所述步骤中的至少一个步骤;

其中希望的自然语言处理任务包含语言对之间的多语种翻译。

2.根据权利要求1所述的方法,其中使能选择希望的自然语言处理任务的一个或多个参数包含提供用户界面,以便用户选择希望的自然语言处理任务的一个或多个参数。

3.根据权利要求1所述的方法,其中所述一个或多个参数包含典型句子的长度、内容性质、散文质量、预期的感叹词数量、翻译的预期用途、以及文本输入的一般主题中的至少一个。

4.根据权利要求1所述的方法,其中所述一个或多个参数包含一个或多个预定值。

5.根据权利要求1所述的方法,其中使用从一个或多个专家用户收集的数据来训练所述映射方法。

6.根据权利要求1所述的方法,其中使能选择希望的自然语言处理任务的一个或多个参数包含使能调整所述自然语言处理任务中的一个或多个语法成分的相关性。

7.根据权利要求1所述的方法,其中将所述一个或多个选择的参数自动地转变成具有上界和下界的初始种子。

8.根据权利要求1所述的方法,还包含:

向用户提供反馈。

9.根据权利要求8所述的方法,其中向用户提供反馈包含以具有来自原始模型的翻译和来自所述定制模型的翻译的文档的形式提供反馈。

10.根据权利要求1所述的方法,还包含:

应用线性回归算法以转换人类输入,从而变换用于优化的上界和下界。

11.根据权利要求1所述的方法,还包含:

通过收集一个或多个文档的集合、针对每个文档调节优化器的至少一个参数、以及选择对应于每个文档的所述至少一个调节过的参数,来构建训练集。

12.根据权利要求1所述的方法,还包含:

自动分析一组一个或多个代表性文档,并为每个文档分配一个分数,所述分数将与映射方法结合使用以识别用于优化参数的一个或多个区间。

13.根据权利要求1所述的方法,还包含:

提供一种系统,其中所述系统包含至少一个不同的软件模块,每个不同的软件模块包含在有形的计算机可读可记录存储介质上,并且其中所述至少一个不同的软件模块包含在硬件处理器上运行的接口模块、转换模块以及优化模块。

14.一种计算机可读存储介质,其上存储有计算机可读程序,当执行所述计算机程序时,使得计算机执行多个方法步骤,包括:

使能选择希望的自然语言处理任务的一个或多个参数，所述一个或多个参数旨在由训练过的用户或未训练过的用户使用；

使用可训练的映射方法将所述一个或多个选择的参数映射到优化算法的输入参数的一个或多个区间的集合；以及

将具有所述输入参数的一个或多个区间的集合的所述优化算法应用于由自然语言处理引擎所使用的模型，以产生定制模型；

其中希望的自然语言处理任务包含语言对之间的多语种翻译。

15.根据权利要求14所述的计算机可读存储介质，其中使能选择希望的自然语言处理任务的一个或多个参数包含提供用户界面，以便用户选择希望的自然语言处理任务的一个或多个参数。

16.根据权利要求14所述的计算机可读存储介质，其中所述一个或多个参数包含典型句子的长度、内容性质、散文质量、预期的感叹词数量、翻译的预期用途、以及文本输入的一般主题中的至少一个。

17.根据权利要求14所述的计算机可读存储介质，其中使能选择希望的自然语言处理任务的一个或多个参数包含使能调整所述自然语言处理任务中的一个或多个语法成分的相关性。

18.根据权利要求14所述的计算机可读存储介质，其中将所述一个或多个选择的参数自动地转变成具有上界和下界的初始种子。

19.一种用于定制自然语言处理引擎的系统，包含：

至少一个不同的软件模块，每个不同的软件模块包含在有形的计算机可读介质上；

存储器；以及

至少一个处理器，其耦合到所述存储器并操作地用于：

使能选择希望的自然语言处理任务的一个或多个参数，所述一个或多个参数旨在由训练过的用户或未训练过的用户使用；

使用可训练的映射方法将所述一个或多个选择的参数映射到优化算法的输入参数的一个或多个区间的集合；以及

将具有所述输入参数的一个或多个区间的集合的所述优化算法应用于由自然语言处理引擎所使用的模型，以产生定制模型；

其中希望的自然语言处理任务包含语言对之间的多语种翻译。

20.根据权利要求19所述的系统，其中操作地用于使能选择希望的自然语言处理任务的一个或多个参数的耦合到所述存储器的所述至少一个处理器还操作地用于提供用户界面，以便用户选择希望的自然语言处理任务的一个或多个参数。

21.根据权利要求19所述的系统，其中所述一个或多个参数包含典型句子的长度、内容性质、散文质量、预期的感叹词数量、翻译的预期用途、以及文本输入的一般主题中的至少一个。

22.根据权利要求19所述的系统，其中耦合到所述存储器的操作地用于使能选择希望的自然语言处理任务的一个或多个参数的所述至少一个处理器，还操作地用于使能调整所述自然语言处理任务中的一个或多个语法成分的相关性。

23.根据权利要求19所述的系统，其中将所述一个或多个选择的参数自动地转变成具

有上界和下界的初始种子。

定制自然语言处理引擎

[0001] 政府合同

[0002] 本发明是在美国国防先进研究项目局(DARPA)授予的合同号:HR0011-08-C0110(全球自主语言利用(GALE))下的政府支持下做出的。政府对于本发明拥有特定权利。

技术领域

[0003] 本发明的实施例一般涉及信息技术,并且更具体地涉及自然语言处理系统。

背景技术

[0004] 统计机器翻译引擎使用对数线性框架将子模型组合在一起并将子代价(或分数)集成到单个代价/分数,以对翻译决策进行排名。此类框架对用于对数线性式组合的权重敏感,这使得翻译引擎不太适配于不同文体,因为翻译模型的误差表面(error surface)是崎岖的并且优化算法是脆弱的并容易遭受任何起始点(种子)影响,因此。为适配此类模型,优化算法的初始种子在优化成功中会起关键作用。在现有方法中,常常只通过对软件发布中已经提供的种子进行随机扰乱来获得此类初始种子。

[0005] 机器翻译系统的输出文本的翻译质量通常经由包括BLEU(双语评估替身)、TER(翻译编辑率)、WER(字错误率)、METEOR(用于具有明确排序的翻译评估的度量)、n-gram精度及其变种的自动度量进行测量。用于自然语言处理(NLP)的统计模型依赖于初始起始点,从初始起始点它们在给定数据的情况下优化目标函数。寻找最优解通常是困难的(NP-完全),并且优化器寻找高度依赖于初始种子的局部最优。因此,找到较好的初始种子会对结果的质量产生正面的影响,并且存在进行此类寻找的需求。

发明内容

[0006] 在本发明的一个方面中,提供了用于定制自然语言处理引擎的技术。一种用于定制自然语言处理引擎的示例性计算机实现的方法可包括以下步骤:使能选择希望的自然语言处理任务的一个或多个参数,所述一个或多个参数旨在由训练过的和未训练过的用户使用;将所述一个或多个选择的参数映射到优化算法的输入参数的一个或多个区间的集合;以及将具有所述输入参数的一个或多个区间的集合的所述优化算法应用于由自然语言处理引擎所使用的模型,以产生定制模型。

[0007] 本发明的另一个方面或其元素可以以有形地包含计算机可读指令的制造品(article of manufacture)的形式来实现,当执行所述计算机可读指令时,其使得计算机执行如本文所描述的多个方法步骤。此外,本发明的另一个方面或其元素可以以包括存储器和至少一个处理器(其耦合到所述存储器并操作地执行所述方法步骤)的装置的形式来实现。此外,本发明的另一个方面或其元素可以以用于执行本文所描述的所述方法步骤(或其元素)的部件的形式来实现;所述部件可包括(i)硬件模块,(ii)软件模块,或(iii)硬件模块和软件模块的组合;(i)至(iii)中的任何一个实现本文所阐述的特定技术,并且所述软件模块存储在有形计算机可读存储介质(或多个此类介质)中。

[0008] 从以下结合附图阅读的本发明的说明性实施例的详细描述中，本发明的这些以及其他目的、特征和优点将变得明显。

附图说明

[0009] 图1是示出根据本发明的一个实施例用于影响优化算法达到不同的局部最优的种子选择的示例的图；

[0010] 图2是示出根据本发明的一个方面的示例实施例的框图；

[0011] 图3是示出根据本发明的一个方面的示例实施例的图；

[0012] 图4A是示出根据本发明的一个方面的用于显示来自系统训练的预测量句子的用户界面的框图；

[0013] 图4B是示出根据本发明的一个方面的用于请求用户输入的用户界面的框图；

[0014] 图4C是示出根据本发明的一个方面的用于确认用户输入以及启动优化过程的用户界面的框图；

[0015] 图5是示出根据本发明的一个实施例的用于定制自然语言处理引擎的技术的流程图；以及

[0016] 图6是在其上可实现本发明的至少一个实施例的示例性计算机系统的系统图。

具体实施方式

[0017] 在本文中，将在一个或多个自动化机器翻译系统的上下文中说明性地描述本发明的原理。然而，应当了解，本发明的原理不限于任何特定的系统架构，并且更一般地适用于任何自然语言处理系统，其中，优化与自然语言处理系统相关联的一个或多个结果将是希望的。

[0018] 如本文所使用的，短语“自然语言处理”(NLP)一般指与计算机和人类(自然)语言之间的交互有关的计算机科学和语言学领域。因此，由于机器翻译系统是自然语言处理系统的一个示例，因此“机器翻译”一般指在计算机系统的控制下用于将第一自然语言(仅作为示例，英语语言)中的文本翻译成第二自然语言(仅作为示例，汉语语言家族中的一个或意大利语)中的文本的技术。

[0019] 还应当理解，可经由自动语音识别(ASR)系统(如已知的，其从说话者接收口语表达并将所述口语表达转换(译码)成文本)来生成机器翻译系统的文本输入。因此，说话者可以用第一自然语言说话，并且ASR生成的文本将作为机器翻译系统的输入。类似的，由机器翻译系统用第二自然语言输出的文本可作为自动文本到语音(TTS)系统(如已知的，其将文本转换成以第二自然语言可听见地呈现给听者的语音)的输入。然而，应当了解，本发明的原理集中于机器翻译系统(更一般地，自然语言处理系统)，而不是ASR或TTS系统。

[0020] 现有数据驱动的方法忽略了潜在用户的范围(它们不能将用户建模为属于离散群或属于分布)，这导致了对于潜在用户群来说较低的翻译质量。因此，例如，现有的方法不能既处理初级用户又处理老练用户，即系统是以特定用户群的需求为代价，为“一般用户”进行优化的。通过给予特定域/群更多权重的度量来取代平均度量的直接替换通常也不是希望的。例如，如果方法是为高级用户在复杂(困难的)材料(例如化学中的技术术语)上进行优化的，则存在降低由大多数用户所使用的简单材料的翻译质量的风险。最后，不加区别地

增加更多的数据既昂贵(收集平行语料库是劳动密集型工作,需要对同一文档进行多重人工翻译)又低效(收集足够的数据以覆盖老练用户感兴趣的所有可能场景,以及收集典型用户每天都接触的流动性极大的web/社区内容是实际上不可能的)。

[0021] 因此,如本文所描述的,本发明的一个方面包括经由手工调整基于句法的机器翻译引擎中的同步语法结构来降低适配代价。本方面的一个实施例包括基于终端用户输入寻找初始种子。对于翻译引擎来说,技术包括揭露关键语法成分并为用户提供用户界面(UI)机制以对所述成分的相关性进行调整。这自动地翻译成初始种子(具有上界和下界),以便优化算法加速改进域特定翻译。

[0022] 如本文所详细说明的,本发明的一个方面包括用于获得用于自动优化/适配的种子的人机界面。另外,本发明的另一个方面包括用于构建支持此类人机交互的翻译引擎的框架。

[0023] 可经由人机交互配置的同步语法通过提供用于针对用户数据对翻译引擎参数进行调整的权重的初始猜测值,提供了灵活性。用户可指定将被翻译的材料的特点(例如,对于语音交谈数据材料是否是内在地单调的,或者对于正式新闻或歌词是否是期望更多的重排序)。

[0024] 如本文所描述的,对于用户快速建立用于任何进一步适配和优化的更好的基线或起始点来说,有限的人机交互的数量可能是有用的。因此,用户和自动程序可节省收集用于以预定的方式适配翻译引擎的用户数据的代价,并且加速优化算法以达到更好的结果。

[0025] 本文所详细说明的技术可用于在语言对之间进行双语/多语翻译。此外,本发明的一个方面用于自然语言处理并且包括统计模型。此外,在本发明的一个或多个实施例中,用户不会看到与本文所描述的技术结合的任何规则。所述界面用于推断用于翻译的用户特定数据的难度,并且本发明的一个方面同时推断用于运行优化算法的种子和界限,以对参数进行调整。

[0026] 在一个说明性实施例中,优化算法使用比现有方法中所使用的那些算法更一般化的算法,即被称为单纯形下山(simplex-downhill)算法的算法。单纯形下山算法是基于试探法的线性搜索技术,并且被认为比标准的最小错误率训练或MER更有效。见,B.Zhao等,"A Simplex Armijo Downhill Algorithm for Optimizing Statistical Machine Translation Decoding Parameters"Proceedings of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies (NAACL HLT-2009),Denver,CO,USA,其公开的全部内容通过引用被包含于此。该算法从种子K维权重向量(对应于(K-1)维单纯形中的一个点)开始。所述技术还在每一维上进行循环,并且通过将第k维设置为其上界和下界,将这个原始种子变换成高维中的一个点。如此,在高维空间中就产生了"球"(或"雪球")。在优化期间使用这个"雪球",迭代地应用四种操作:扩展、收缩、反射以及Armijo线搜索--以将该雪球滚动到包含最优解的更好空间,并且使其收缩直到到达局部最优。

[0027] Armijo算法改变轨迹,以使单纯形收缩到局部最优,并且使得该算法能够有更好的机会走出由自动机器翻译(MT)评价度量计算的充满误差的表面。

[0028] 如所述的,本发明的优选实施例可用于翻译系统从一种人类语言(源语言)到一种不同的人类语言(目标语言)的域适配。在这种实施例中,已经存在从源语言到目标语言的

通用翻译系统，并且用户对改进系统在特定域(例如，翻译语音转录文本)上的性能感兴趣。考虑到语音转录文本的特点通常是比例如正式文档更短的句子，具有可能需要更少重排序的更简单结构。

[0029] 提供给用户的界面具有若干控件(例如，仪表盘)，其捕获将被翻译的数据的类型下列方面：例如，句子的典型长度(从碎片的到非常长)、内容的性质(从一般到非常特定于域)、散文的质量(从不合语法到教科书式)、是否有预期的感叹词(没有到许多)等。用户能够使用UI做出选择(例如，中等长度句子，一般内容、不合语法的句子以及碎片的散文)。这些值被映射到用于初始种子的区间(如本文所描述的)，并且使用生成的种子来适配模型。

[0030] 另外，本发明的另一个方面向用户提供反馈。例如，可以以具有来自原始系统的翻译和来自适配后的系统的翻译的文档的形式，提供反馈。使用此类反馈，用户可以决定对过程进行迭代。

[0031] 如以上所述的，本发明的一个方面包括构建引擎以支持手工可调整的方案，以便改进为任何预定适配过程提供初始种子。通过为用户提供不需要了解翻译算法是如何运行的控件来改进翻译质量。由用户选择的值被映射到用于由优化算法所使用的参数的值的范围，优化算法以这些参数作为种子，并且优化算法用于适配翻译模型。因此，使用用于优化算法的适当的种子改进了对于特定域的翻译质量。

[0032] 图1是示出根据本发明的一个实施例的用于影响优化算法到达不同局部最优的种子选择的示例的图102。如在图1中所说明的，坏种子选择可导致次优优化结果，混淆用户，以及使满意度受损。然而，好种子选择可导致更快地收敛到最优点，并且改进用户体验。

[0033] 本发明的一个方面包括在以人为中心的度量和用于优化算法的种子参数之间的映射。以人为中心的度量例如可以指由非技术、未训练过的人可容易理解的参数，包括但不限于：典型句子的长度、翻译的预期用途、文本的一般主题等。好种子可能已经涉及搜索努力，并且这个搜索过程可包括单调、hiero、树到串、串到数的概率同步上下文无关语法(PSCFG)的上界，下界以及相对长度。本发明的一个或多个实施例包括一组用于不同文体类型的预定的上界和下界，其是经由监督学习或非监督学习从用于系统构建的训练数据学习到的。然后，将人类输入映射到用于种子的预定范围。

[0034] 另外，本发明的一个方面包括回归/最小二乘估计，以用于将人类输入转换成与语法结构相关联的权重的上界和下界。在一个实施例中，用户例如选择分数，并且系统将这些分数映射到例如使用回归算法的区间。

[0035] 此外，本发明的一个方面包括学习线性回归算法以转换人类输入，从而变换用于优化的上/下界。因此，这可包括提供句子，以便用户用标度标记(例如，从[1-5]，对用户来说，1是最简单的句子，5是最困难的句子)。另外，这些标记过的句子可保存成向量α。内部可读性分数被计算并保存成向量β。通过以最小平方误差将β变换成α：$\alpha=\bar{\lambda}\beta+\varepsilon$，对回归或最小平方误差参数$\bar{\lambda}$进行比较，其中ε是预测的人类评分分数与内部机器评分分数之间的残余误差。同一参数$\bar{\lambda}$可用于预测或确定用于每个揭露的语法成分的上界和下界，并且从[下界，上界]所限定的种子可用于任何后续的优化。

[0036] 在本发明的一个或多个实施例中，通过收集文档集合、针对每个文档对优化器的参数进行调整(例如，经由专家)、以及使一组人单独地选择UI中的参数以描述他们对每个

文档的感觉，来构建训练集。然后，结合方法使用此收集的数据，以学习用户输入和用于参数的区间之间的映射(例如，诸如以上所描述的线性回归方法)。

[0037] 在另一个实施例中，用户向系统提供一组代表性文档，并且系统自动分析这些文档并给这些文档指定分数(例如，诸如Flesch-Kincaid年级水平、Gunning-Fog分数、Coleman-Liau索引以及SMOG索引)。然后，这些分数的分布结合映射方法(诸如本文所描述的映射方法)使用，以识别用于优化参数的区间。

[0038] 在这种实施例中，不要求用户提供人类可理解的参数的值，而是提供将被翻译的文档类型的特定代表性示例。使用这些文档，本发明的一个或多个实施例自动计算各种量，并从这些值的集合构建用于优化算法的参数的一组区间。另外，本领域的技术人员将了解，存在若干可能机制来指定示例文档，包括系统迭代地向用户提供附加建议以及用户选择或拒绝建议的交互式轮流方法。

[0039] 在本发明的另一个方面中，系统可从由先前用户或由软件的提供者构建的预先指定的多组用于参数的区间的集合开始，每一组输入对应于特定域。在这个实施例中，系统使用由用户提供的示例或人类指定的参数，来选择预先指定的多组参数中的一组参数。在这种实施例中，用户被确保最终获得具有已证明参数、并且可能需要来自用户的更少示例的系统。另外，如果示例的数目足够大，则系统可以构造一组新的用于参数的区间。

[0040] 图2是示出根据本发明的一个方面的一个示例实施例的框图。作为说明，图2示出了用于请求用户输入他或她感兴趣的数据的接口模块202。输入转换模块204将用户输入转换成用于优化的上界、下界、或起始点。此外，优化模块206以给定上/下界或初始种子执行优化算法，并且译码模块208应用适配过的权重，以用于在软件中进行译码。另外，用户的实际数据流210也被提供给翻译模块212，其使用适配过的系统来提供对用户数据的翻译，从而生成翻译输出214。

[0041] 图3是示出根据本发明的一个方面的一个示例实施例的图。作为说明，在步骤302中，系统(经由用户界面)显示在系统训练期间进行测量的若干句子。在步骤304中，用户根据他或她自己的判断，对这些句子从难到易进行测量。另外，在步骤306，系统请求用户对他或她的数据难度(给定所显示的句子)进行排名，例如从1至5。因此，在步骤308，系统将用户的(一个或多个)选择映射到用于为优化提供种子的译码参数的下/上界。

[0042] 在步骤310，系统生成更好的/更准确翻译种子并运行优化。此外，在步骤312，系统应用优化后的权重，并调整翻译引擎。

[0043] 图4A是示出根据本发明的一个方面的用于显示来自系统训练的预测量的句子的用户界面402的框图。作为说明，图4A示出指令组件404、句子查询406和408、以及排名组件410和412。

[0044] 图4B是示出根据本发明的一个方面的用于请求用户输入的用户界面402的框图。作为说明，图4B示出指令组件422、查询响应组件424、以及排名组件426。

[0045] 图4C是示出根据本发明的一个方面的用于确认用户输入以及启动优化过程的用户界面402的框图。作为说明，图4C示出指令组件432和运行优化提示组件434。

[0046] 图5是示出根据本发明的一个实施例的用于定制自然语言处理引擎的技术的流程图(其中至少一个步骤是由计算机设备来执行的)。步骤502包括使能选择希望的自然语言处理任务(例如，语言对之间的多语种翻译)的一个或多个参数，所述一个或多个参数旨在

由训练过的或未训练过的用户使用。这个步骤例如可以使用接口模块来执行。这些参数可包括可由未训练的用户理解的参数(例如,以人为中心的参数)。例如,参数可包括典型句子的长度、内容的性质、散文的质量、预期的感叹词的数量、翻译的预期用途、以及文本输入的一般主题。此外,参数可包括预定值。

[0047] 使能选择希望的自然语言处理任务的参数可包括提供用户界面,以便用户选择希望的自然语言处理任务的参数。此外,使能选择希望的自然语言处理任务的参数可包括使能对自然语言处理任务中的一个或多个语法成分的相关性进行调整。

[0048] 步骤504包括将一个或多个所选择的参数映射到优化算法的输入参数的一个或多个区间的集合。这个步骤例如可使用转换模块来执行。将所选择的参数映射到优化算法的输入参数的区间的集合可包括使用可训练的映射方法。可使用从一个或多个专家用户收集的数据来训练映射方法。另外,可自动地将所选择的参数转变成具有上界和下界的初始种子。

[0049] 步骤506包括将具有输入参数的一个或多个区间的集合的优化算法应用于由自然语言处理引擎所使用的模型,以产生定制模型(为终端用户)。这个步骤例如可使用优化模块来执行。

[0050] 在图5中所示出的技术还可包括例如以具有来自原始模型的翻译和来自定制模型的翻译的文档的形式向用户提供反馈。本发明的一个方面还可包括应用转换用户输入、以便变换用于优化的上界和下界的线性回归算法。

[0051] 此外,在图5中所示出的技术包括通过收集一个或多个文档的集合、针对每个文档调节优化器的至少一个参数、以及选择至少一个调节过的参数来对应每个文档,来构建训练集。另外,本发明的一个方面可包括自动分析一组代表性文档,以及给每个文档分配一个分数,其将与映射方法结合使用以识别用于优化参数的区间。

[0052] 如本文所描述的,图5中所示出的技术还可包括提供一种系统,其中所述系统包括不同的软件模块,所述不同软件模块中的每一个软件模块包含在计算机可读可记录的存储介质上。例如,所有模块(或其任何子集)可以在同一介质上,或每一个模块可以在不同的介质上。模块可包括在图中所示出的组件中的任何一个组件或全部组件。在本发明的一个方面中,所述模块包括:例如可在硬件处理器上运行的接口模块、输入转换模块、优化模块、译码模块、以及翻译模块。然后可使用在硬件处理器上执行的系统的不同模块(如以上所描述的)来实现所述方法步骤。此外,计算机程序产品可包括有形计算机可读可记录存储介质,其具有适用于被执行以实现本文所描述的至少一个方法步骤(包括提供具有不同软件模块的系统)的代码。

[0053] 另外,图5中所示出的技术可经由计算机程序产品来实现,其可包括存储在数据处理系统中的计算机可读存储介质中的计算机可使用的程序代码,并且其中所述计算机可使用的程序代码是通过网络从远程数据处理系统下载的。此外,在本发明的一个方面中,计算机程序产品可包括存储在服务器数据处理系统中的计算机可读存储介质中的计算机可使用的程序代码,并且其中计算机可使用的程序代码通过网络被下载到远程数据处理系统以便在远程系统的计算机可读存储介质中使用。

[0054] 如本领域的技术人员将了解,本发明的各方面可以体现为系统、方法或计算机程序产品。因此,本发明的各方面可采用全部硬件的实施例、全部软件的实施例(包括固件、常

驻软件、微代码等)或组合软件方面和硬件方面的实施例的方式,其在本文中通称为“电路”、“模块”或“系统”。此外,本发明的各方面可采用包含在具有包含在其上的计算机可读程序代码的计算机可读介质中的计算机程序产品的形式。

[0055] 本发明的一个方面或其元素可以以装置的形式来实现,所述装置包括存储器以及耦合到所述存储器并操作地执行示例性方法步骤的至少一个处理器。

[0056] 另外,本发明的一个方面可使用在通用计算机或工作站上运行的软件。参照图6,此类实现方式例如可使用例如:处理器602、存储器604、以及例如由显示器606和键盘608组成的输入/输出接口。如本文所使用的术语“处理器”旨在包括任何处理设备,例如诸如包括CPU(中央处理器)和/或其它形式的处理电路的处理设备。此外,术语“处理器”可指超过一个的单独处理器。术语“存储器”旨在包括与处理器或CPU相关联的存储器,例如诸如RAM(随机存取存储器)、ROM(只读存储器)、固定存储设备(例如硬盘驱动器)、可移动存储设备(例如,软盘)、闪存存储器等。另外,如本文所使用的短语“输入/输出接口”例如旨在包括用于将数据输入到处理单元的机制(例如,鼠标)、以及用于提供与所述处理单元相关联的结果的机制(例如,打印机)。处理器602、存储器604、和输入/输出接口(诸如显示器606和键盘608)例如可经由总线610进行互连,作为数据处理单元612的一部分。合适的互连(例如经由总线610)还可提供给网络接口614(诸如网卡),其可被提供以与计算机网络进行接口,以及提供给介质接口616(诸如软磁盘或CD-ROM驱动器),其可被提供以与介质618进行接口。

[0057] 因此,计算机软件包括用于执行如本文所描述的本发明的方法的指令或代码,可存储在相关联的存储设备(例如,ROM、固定或可移动存储器)中,并且当准备好被使用时,被部分或全部加载(例如,加载到RAM)并由CPU执行。此类软件可包括(但不限于)固件、常驻软件、微代码等。

[0058] 适合用于存储和/或执行程序代码的数据处理系统将包括通过系统总线610直接或间接耦合到存储单元604的至少一个处理器602。存储单元可包括在程序代码的实际执行期间所使用的本地存储器、大容量存储器、以及缓存存储器,其提供暂时存储至少一些程序代码,以便减少在执行期间必须从大容量存储器取回代码的次数。

[0059] 输入/输入或I/O设备(包括但不限于键盘608、显示器606、指点设备等)可直接(诸如经由总线610)或通过介于中间的I/O控制器(为清楚起见而省略)耦合到所述系统。

[0060] 网络适配器(诸如网络接口614)还可耦合到所述系统,以使能数据处理系统通过介于中间的私有或公用网络耦合到其它数据处理系统或远程打印机或存储设备。调制解调器、电缆调制解调器和以太网卡仅是若干当前可用类型的网络适配器。

[0061] 如本文(包括权利要求)所使用的,“服务器”包括运行服务器程序的物理数据处理系统(例如,如图6中所示出的系统612)。应当理解,此类物理服务器可包括或不包括显示器和键盘。

[0062] 如所指出的,本发明的各方面可采用包含在具有在其上包含的计算机可读程序代码的计算机可读介质中的计算机程序产品的形式。此外,可以采用一个或多个计算机可读介质的任意组合。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。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更具体的例子(非穷举的列表)将包括:具有一个或多个导线的电连接、便携式计算机盘、硬盘、随机访问存储器(RAM)、

只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、光纤、便携式紧凑盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本文件中，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。

[0063] 计算机可读的信号介质可以包括例如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。计算机可读的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。

[0064] 计算机可读介质上包含的程序代码可以用适当的介质传输，包括——但不限于——无线、有线、光缆、RF等等，或者上述的任意合适的组合。

[0065] 可以以至少一种程序设计语言的任意组合来编写用于执行本发明的各方面操作的计算机程序代码，所述程序设计语言包括面向对象的程序设计语言—诸如Java、Smalltalk、C++等，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——包括局域网(LAN)或广域网(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)。

[0066] 以上参照根据本发明的实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述了本发明。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机程序指令实现。这些计算机程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器，从而生产出一种机器，使得这些计算机程序指令在通过计算机或其它可编程数据处理装置的处理器执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。

[0067] 也可以把这些计算机程序指令存储在计算机可读介质中，这些指令使得计算机、其它可编程数据处理装置、或其他设备以特定方式工作，从而，存储在计算机可读介质中的指令就产生出包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的指令的制造品(article of manufacture)。因此，本发明的一个方面包括有形地包含计算机可读指令的制造品，当执行指令时，使得计算机执行如本文所描述的多个方法步骤。

[0068] 也可以把计算机程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程装置或其它设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机或其它可编程装置上执行的指令提供实现流程图和/或框图中的一个或多个方框中所指定的功能/动作的过程。

[0069] 附图中的流程图和框图显示了根据本发明的各种实施例的系统、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，所述模块、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，两个连续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。也要注意的是，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动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。

[0070] 需要注意的是，本文所描述的方法中的任何一个方法可包括提供在计算机可读存储介质上包含的不同软件模块的系统的附加步骤；所述模块例如可包括图2中所示出的组件中的任何组件或全部组件。然后，可使用在硬件处理器602上运行的如以上所述的系统的不同软件模块和/或子模块来执行所述方法步骤。此外，计算机程序产品可包括计算机可读存储介质，其具有适用于被执行以实现本文中所描述的至少一个方法步骤(包括提供具有不同的软件模块的系统)的代码。

[0071] 在任何情况下，应当理解，本文中所说明的组件可以以硬件、软件、或其组合的各种形式来实现；例如，专用集成电路(多个)(ASICS)、功能电路、具有相关联存储器的适当编程的通用数字计算机等。给定本文中所提供的本发明的教导，相关领域的普通技术人员将能够设想本发明的组件的其它实现方式。

[0072] 本文所使用的术语仅是出于描述特定实施例的目的，并且不是旨在限制本发明。如本文所使用的，单数形式“一个”、“一种”和“所述”旨在也包括复数形式，除非上下文中以其他方式清楚地指出。还应当理解，当在本说明书中使用时，术语“包含”和/或“包含有”指定存在所述的特征、整数、步骤、操作、元素、和/或组件，但是不排除存在或增加另外一个特征、整数、步骤、操作、元素组件、和/或其组合。

[0073] 在下面的权利要求中，所有装置或步骤加功能元件的对应的结构、材料、动作、以及等同物旨在包括用于与如明确要求的其它所要求的元件组合执行该功能的任何结构、材料、或动作。出于说明和描述的目的已经提供了本发明的描述，但是不是旨在是穷尽的或将本发明限制于所公开的形式。对于本领域的普通技术人员来说，许多修改和变化将是明显的，而不背离本发明的范围和精神。所选择和描述的实施例是为了更好地解释本发明的原理和实际应用，以及使能本领域的其它普通技术人员理解本发明具有如适用于特定预期使用的各种修改的各种实施例。

[0074] 本发明中的至少一个方面可提供有益效果，例如诸如减少收集用于以预定的方式适配翻译引擎的用户数据的代价，以及加速优化算法到达更好的结果。

[0075] 已经出于说明的目的提供了本发明的各种实施例的描述，但是不是旨在是穷尽的或限制于所公开的实施例。对于本领域的普通技术人员来说，许多修改和变化将是明显的，而不背离所能描述的实施例的范围和精神的情况。所选择的在本文中所使用的术语是为了更好地解释实施例的原理、实际应用或在市场上可以找到的技术上的技术改进，或使能本领域的其他普通技术人员理解本文中所公开的实施例。

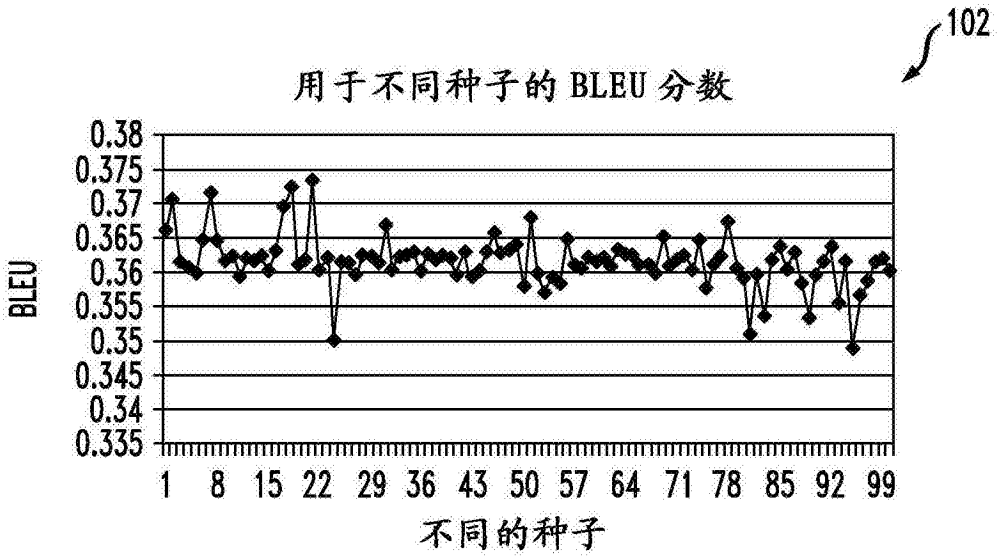

102
用于不同种子的 BLEU 分数
BLEU
0.38
0.375
0.37
0.365
0.36
0.355
0.35
0.345
0.34
0.335
1 8 15 22 29 36 43 50 57 64 71 78 85 92 99
不同的种子

图1

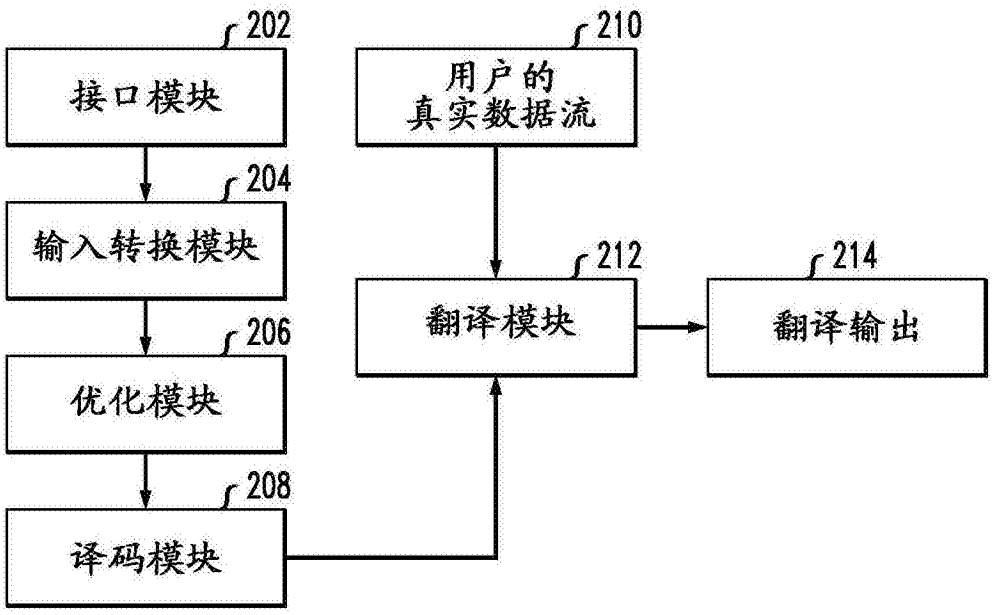

202
接口模块
204
输入转换模块
206
优化模块
208
译码模块
210
用户的
真实数据流
212
翻译模块
214
翻译输出

图2

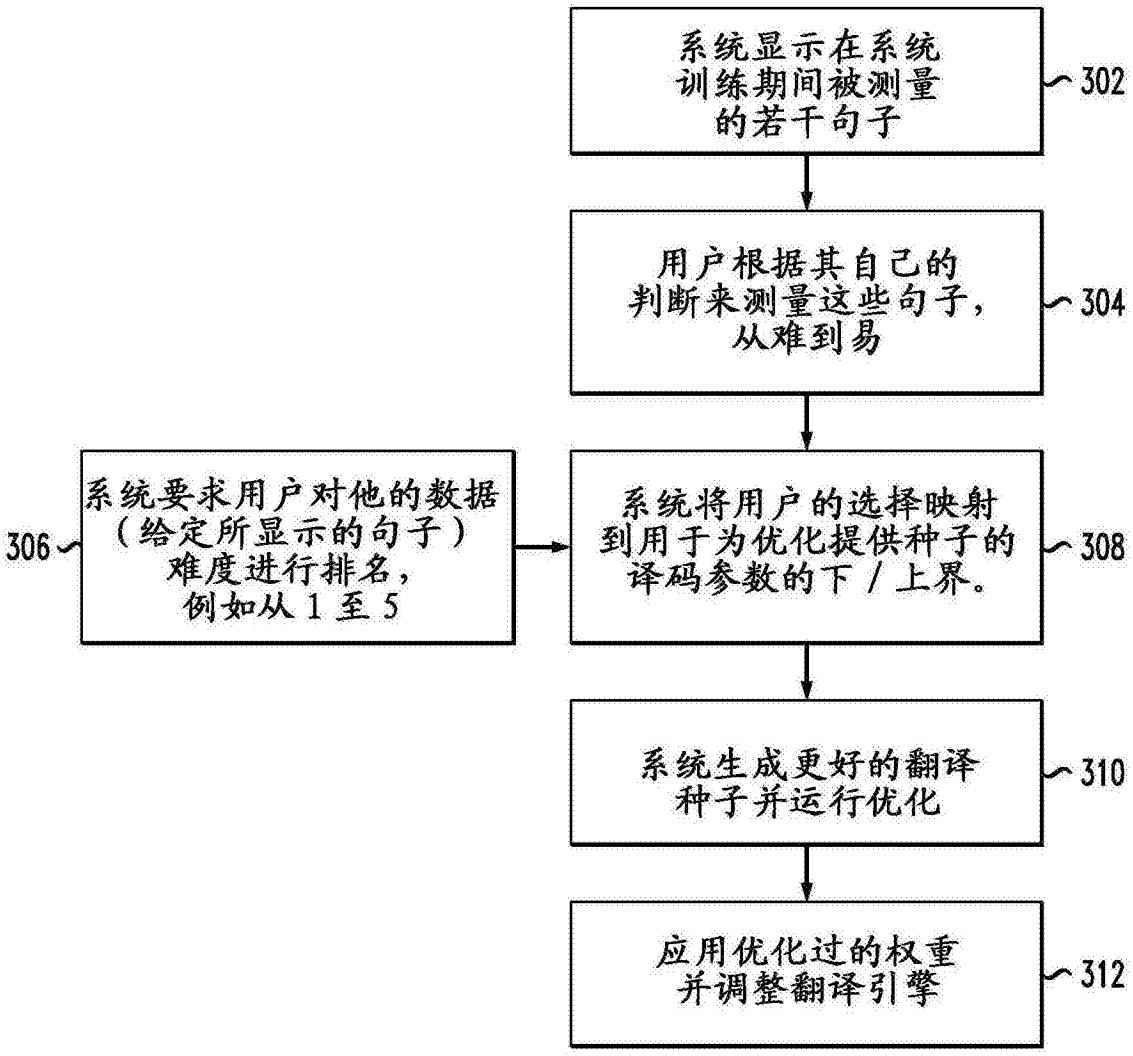

系统显示在系统训练期间被测量的若干句子
302
用户根据其自己的判断来测量这些句子，从难到易
304
306
系统要求用户对他的数据（给定所显示的句子）难度进行排名，例如从 1 至 5
系统将用户的选择映射到用于为优化提供种子的译码参数的下 / 上界。
308
系统生成更好的翻译种子并运行优化
310
应用优化过的权重并调整翻译引擎
312

图3

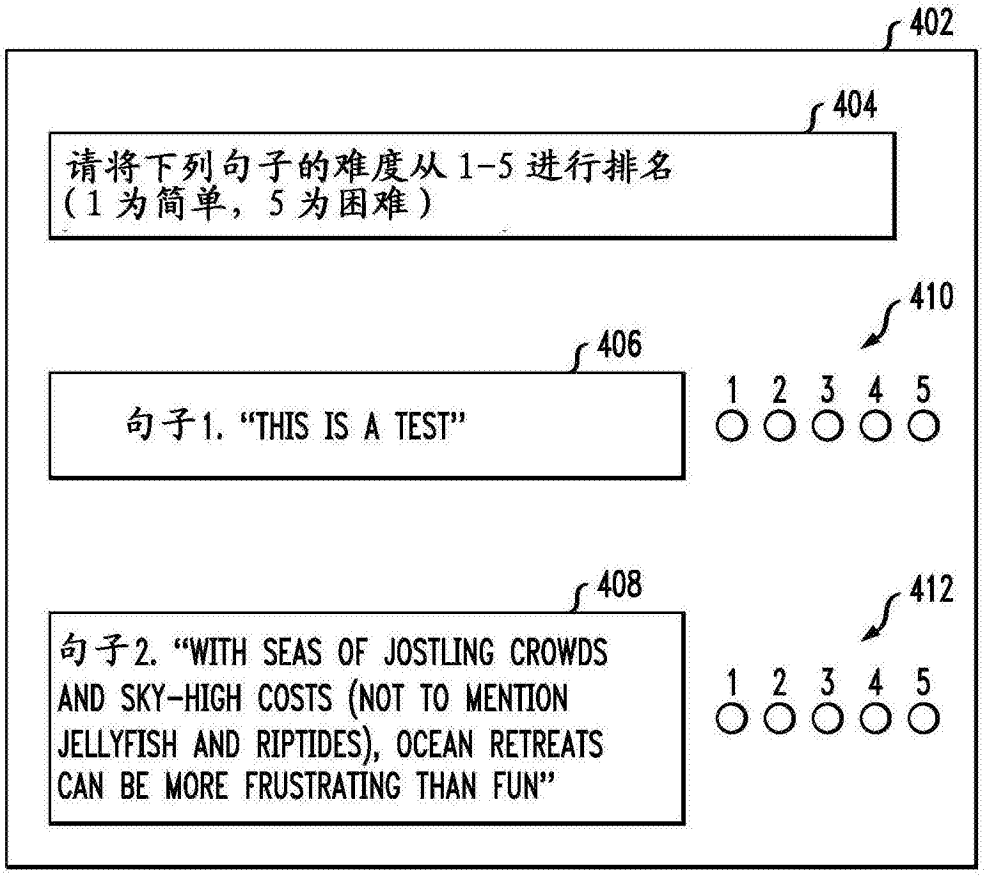

402
404
请将下列句子的难度从 1-5 进行排名
(1 为简单，5 为困难)
410
406
1 2 3 4 5
句子1. "THIS IS A TEST"
408
412
句子2. "WITH SEAS OF JOSTLING CROWDS AND SKY-HIGH COSTS (NOT TO MENTION JELLYFISH AND RIPTIDES), OCEAN RETREATS CAN BE MORE FRUSTRATING THAN FUN"
1 2 3 4 5

图4A

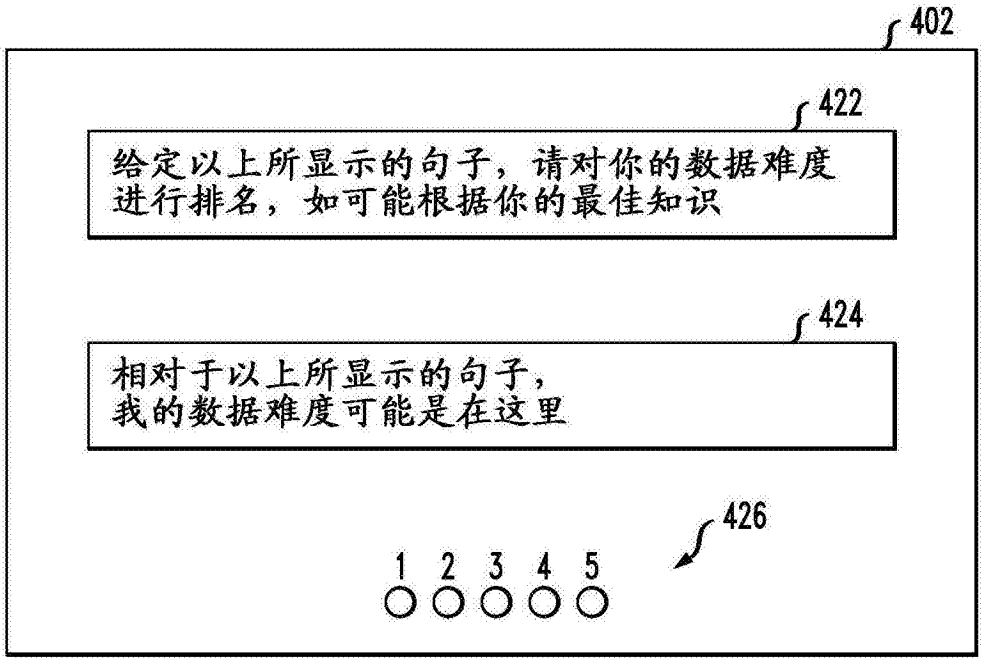

402
422
给定以上所显示的句子，请对你的数据难度进行排名，如可能根据你的最佳知识
424
相对于以上所显示的句子，
我的数据难度可能是在这里
426
1 2 3 4 5

图4B

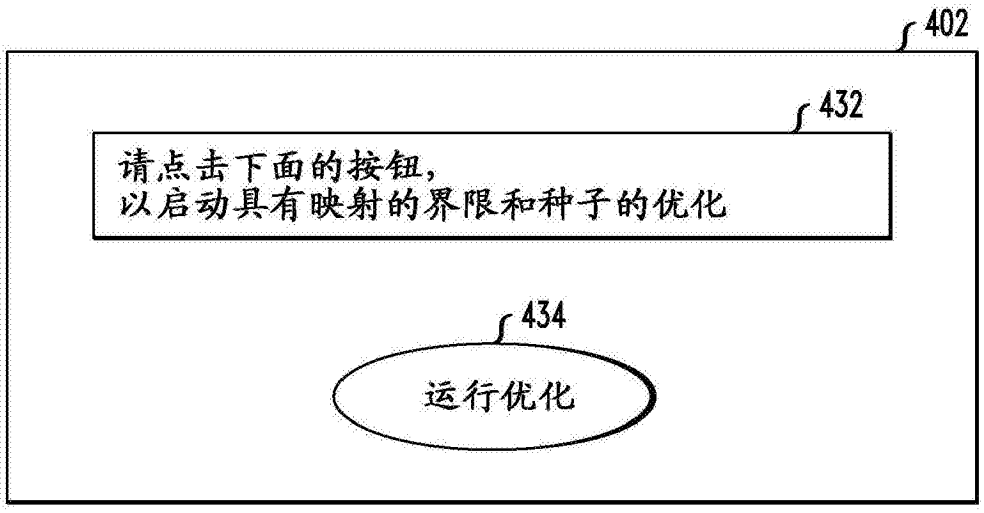

402
432
请点击下面的按钮，
以启动具有映射的界限和种子的优化
434
运行优化

图4C

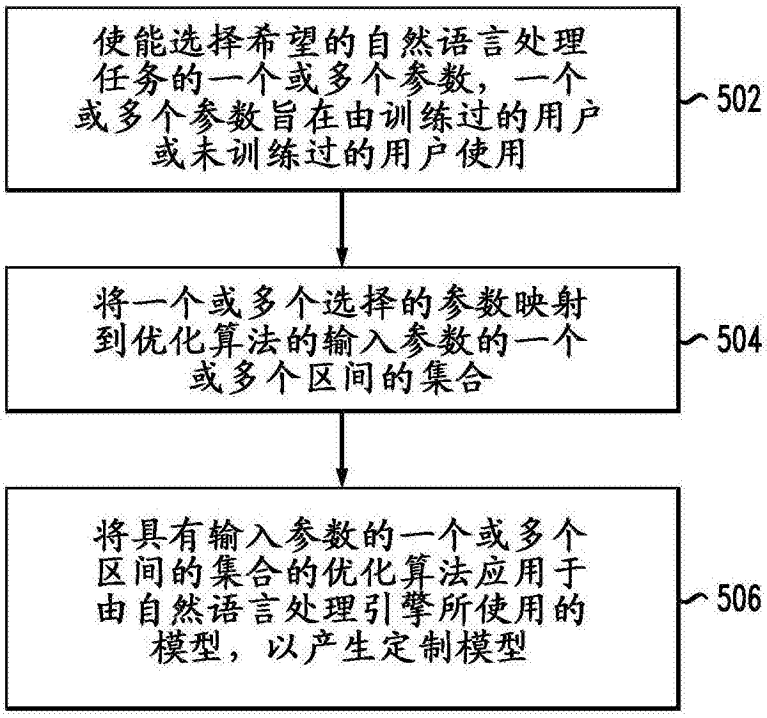

使能选择希望的自然语言处理任务的一个或多个参数，一个或多个参数旨在由训练过的用户或未训练过的用户使用
502
将一个或多个选择的参数映射到优化算法的输入参数的一个或多个区间的集合
504
将具有输入参数的一个或多个区间的集合的优化算法应用于由自然语言处理引擎所使用的模型，以产生定制模型
506

图5

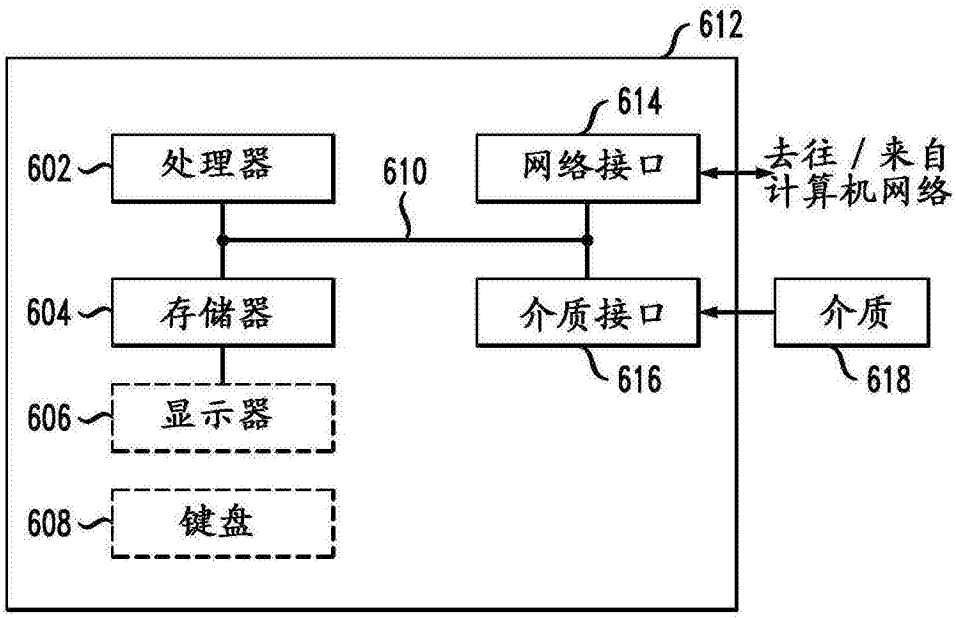

612
614
602
处理器
610
网络接口
去往 / 来自
计算机网络
604
存储器
介质接口
介质
616
618
606
显示器
608
键盘

图6