(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 110929504 B
(45)授权公告日 2023.05.30

(21)申请号 201811117612.3
(22)申请日 2018.09.20
(65)同一申请的已公布的文献号
申请公布号 CN 110929504 A
(43)申请公布日 2020.03.27
(73)专利权人 阿里巴巴集团控股有限公司
地址 英属开曼群岛大开曼资本大厦一座四层847号邮箱
(72)发明人 李辰 周君沛 刘恒友 包祖贻 徐光伟 李林琳
(74)专利代理机构 北京博浩百睿知识产权代理有限责任公司 11134
专利代理师 褚敏 宋子良
(51)Int.Cl.
G06F 40/253(2020.01)
G06F 40/284(2020.01)
G06F 40/44(2020.01)
G06F 40/58(2020.01)
(56)对比文件
CN 102799579 A,2012.11.28
CN 103488488 A,2014.01.01
CN 105068998 A,2015.11.18
CN 107122346 A,2017.09.01
CN 107622053 A,2018.01.23
CN 107807915 A,2018.03.16
CN 107861954 A,2018.03.30
US 2017139906 A1,2017.05.18
US 2018018577 A1,2018.01.18
韦向峰;张全;熊亮.一种基于语义分析的汉语语音识别纠错方法.计算机科学.2006,(第10期),第152-155页.
审查员 徐雯晖

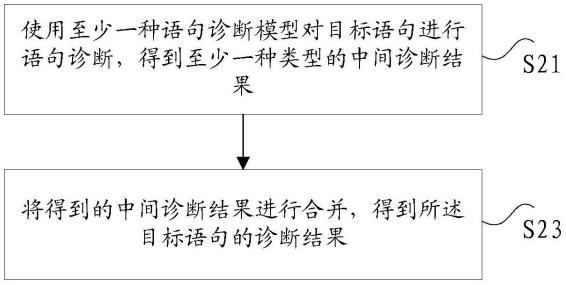
权利要求书4页 说明书15页 附图4页

(54)发明名称

语句诊断方法、装置和系统

(57)摘要

本发明公开了一种语句诊断方法、装置和系统。其中,该方法包括:使用至少一种语句诊断模型对目标语句进行语句诊断,得到至少一种类型的中间诊断结果,其中,语句诊断模型至少包括如下至少一个:基于规则的模型、基于统计机器翻译的模型和基于神经机器翻译的模型;将得到的中间诊断结果进行合并,得到目标语句的诊断结果。本发明解决了现有技术中语句诊断系统的效果不佳的技术问题。



CN 110929504 B

1.一种语句诊断方法,其特征在于,包括:

使用至少两种语句诊断模型对目标语句进行语句诊断,得到至少两种类型的中间诊断结果,其中,所述语句诊断模型包括如下至少两个:基于规则的模型、基于统计机器翻译的模型和基于神经机器翻译的模型;

将得到的中间诊断结果进行合并,得到所述目标语句的诊断结果;

其中,使用至少两种语句诊断模型对目标语句进行语句诊断,得到至少两种类型的中间诊断结果,包括:使用所述基于规则的模型对所述目标语句进行语句诊断,得到第一中间诊断结果;使用所述基于统计机器翻译的模型对所述目标语句进行语句诊断,得到多个第二中间诊断结果;使用所述基于神经机器翻译的模型对所述目标语句进行语句诊断,得到多个第三中间诊断结果;

所述方法还包括:通过预定合并方式对多个所述第二中间诊断结果进行合并得到第一合并结果;和/或,通过所述预定合并方式对多个所述第三中间诊断结果进行合并得到第二合并结果;

将得到的中间诊断结果进行合并,得到所述目标语句的诊断结果,包括:将所述第一中间诊断结果、所述第一合并结果和所述第二合并结果中的至少两个进行合并,得到所述诊断结果。

2.根据权利要求1所述的方法,其特征在于,使用所述基于规则的模型对所述目标语句进行语句诊断,得到第一中间诊断结果,包括:

获取预先构建的词语库;

将所述目标语句进行分词处理,得到所述目标语句对应的分词结果;

将所述分词结果中的每个词与所述词语库中的候选词进行比对,查找出不属于所述词语库中的目标词;

将不属于所述词语库中的目标词进行变形处理,得到所述第一中间诊断结果。

3.根据权利要求2所述的方法,其特征在于,将不属于所述词语库中的目标词进行变形处理,得到所述第一中间诊断结果,包括:

将所述目标词进行多种变形处理,得到对应的多种候选诊断结果;

从所述候选诊断结果中选择所述第一中间诊断结果。

4.根据权利要求3所述的方法,其特征在于,将所述目标词进行多种变形处理,得到对应的多种候选诊断结果,包括如下任意一种或多种:

如果所述目标词中包括两个及以上的字符,且将所述目标词中的字符改变顺序后,属于所述词语库中的词语,确定改变字符顺序后的目标词为所述候选诊断结果;

如果所述目标词与所述目标语句中的前一个词语或后一个词语连接得到的第一组合属于所述词语库,确定所述第一组合为所述候选诊断结果;

生成所述目标词中每个字符对应的相似字,所述相似字包括:形近字或音近字,如果所述相似字与所述目标语句中的前一个词语或后一个词语的第二组合属于所述词语库,确定所述第二组合为所述候选诊断结果。

5.根据权利要求3所述的方法,其特征在于,从所述候选诊断结果中选择所述第一中间诊断结果,包括:

使用预设的语言模型对每个候选诊断结果进行打分,其中,所述打分用于表示所述候

选诊断结果的准确程度；

选择得分最高的候选诊断结果为所述第一中间诊断结果。

6.根据权利要求1所述的方法，其特征在于，使用所述基于统计机器翻译的模型对所述目标语句进行语句诊断，得到第二中间诊断结果，包括：

基于预设的短语表获取所述目标语句对应的候选语句；

基于预设的语言模型确定所述候选语句的得分，其中，所述候选语句的得分用于表示所述候选语句的准确程度；

将所述目标语句分别和每个所述候选语句组合，构成多个句子对；

基于预设的翻译模型获取所述句子对的条件概率，其中，所述句子对的条件概率用于表示在所述候选语句生成的情况下，生成所述目标语句的概率；

使用束搜索根据所述得分和所述条件概率得到所述第二中间诊断结果。

7.根据权利要求6所述的方法，其特征在于，基于预设的短语表获取所述目标语句对应的候选语句，包括：

对所述目标语句以预设粒度进行分割，得到所述目标语句对应的多个单位对象；

在预设短语表中查找与所述单位对象对应的内容，并将与每个所述单位对象对应的内容进行排列组合，得到与所述目标语句对应的候选语句。

8.根据权利要求1所述的方法，其特征在于，使用所述基于神经机器翻译的模型对所述目标语句进行语句诊断，得到第三中间诊断结果，包括：

通过所述基于神经机器翻译的模型中的编码器对所述目标语句进行编码，以预测出所述目标语句对应的状态向量；

通过所述基于神经机器翻译的模型中的解码器对所述状态向量进行解码，以预测所述目标语句对应的所述第三中间诊断结果。

9.根据权利要求1所述的方法，其特征在于，将得到的中间诊断结果进行合并，得到所述目标语句的诊断结果，包括：

对所述第一中间诊断结果、多个所述第二中间诊断结果和多个所述第三中间诊断结果进行合并，得到所述目标语句的诊断结果。

10.根据权利要求9所述的方法，其特征在于，所述基于神经机器翻译的模型具有多种不同的配置参数，每个配置参数对应一个第三中间诊断结果。

11.根据权利要求10所述的方法，其特征在于，基于统计机器翻译的模型对所述目标语句进行语句诊断时，所述目标语句分别以词粒度和字粒度被分割，每种分割的粒度对应一个所述第二中间诊断结果。

12.根据权利要求11所述的方法，其特征在于，对所述第一中间诊断结果、多个所述第二中间诊断结果和多个所述第三中间诊断结果进行合并，得到所述目标语句的诊断结果，包括：

通过预定合并方式对多个第二中间诊断结果进行合并得到第一合并结果，并通过所述预定合并方式并对多个第三中间诊断结果进行合并得到第二合并结果；

将所述第一中间诊断结果、所述第一合并结果和所述第二合并结果进行合并，得到所述诊断结果。

13.根据权利要求1所述的方法，其特征在于，所述预定合并方式包括如下任意一种：

确定预先规定的优先级最高的诊断结果为合并后的诊断结果；

确定所有诊断结果的并集为合并后的诊断结果；

确定所有诊断结果的交集为合并后的诊断结果；

确定得分最高的结果为合并后的诊断结果，其中，通过预设语言模型对合并结果进行打分。

14.根据权利要求12所述的方法，其特征在于，将所述第一中间诊断结果、所述第一合并结果和所述第二合并结果进行合并，得到所述诊断结果，包括：

判断所述第一中间诊断结果、所述第一合并结果和所述第二合并结果是否冲突；

如果所述第一中间诊断结果、所述第一合并结果和所述第二合并结果中的任意两项冲突，按照预定合并方式进行合并；

如果所述第一中间诊断结果、所述第一合并结果和所述第二合并结果均冲突，则保持所述目标语句不进行纠正。

15.一种语句诊断装置，其特征在于，包括：

诊断模块，用于使用至少两种语句诊断模型对目标语句进行语句诊断，得到至少两种类型的中间诊断结果，其中，所述语句诊断模型包括如下至少两个：基于规则的模型、基于统计机器翻译的模型和基于神经机器翻译的模型；

合并模块，用于将得到的中间诊断结果进行合并，得到所述目标语句的诊断结果；

其中，所述诊断模块用于通过执行如下步骤来使用至少两种语句诊断模型对目标语句进行语句诊断，得到至少两种类型的中间诊断结果：使用所述基于规则的模型对所述目标语句进行语句诊断，得到第一中间诊断结果；使用所述基于统计机器翻译的模型对所述目标语句进行语句诊断，得到多个第二中间诊断结果；使用所述基于神经机器翻译的模型对所述目标语句进行语句诊断，得到多个第三中间诊断结果；

所述装置还用于执行如下步骤：通过预定合并方式对多个所述第二中间诊断结果进行合并得到第一合并结果；和/或，通过所述预定合并方式对多个所述第三中间诊断结果进行合并得到第二合并结果；

所述合并模块用于通过执行如下步骤来得到所述诊断结果：将所述第一中间诊断结果、所述第一合并结果和所述第二合并结果中的至少两个进行合并，得到所述诊断结果。

16.一种语句诊断系统，包括：

处理器；以及

存储器，与所述处理器连接，用于为所述处理器提供处理以下处理步骤的指令：

使用至少两种语句诊断模型对目标语句进行语句诊断，得到至少两种类型的中间诊断结果，其中，所述语句诊断模型包括如下至少两个：基于规则的模型、基于统计机器翻译的模型和基于神经机器翻译的模型；

将得到的中间诊断结果进行合并，得到所述目标语句的诊断结果；

其中，使用至少两种语句诊断模型对目标语句进行语句诊断，得到至少两种类型的中间诊断结果，包括：使用所述基于规则的模型对所述目标语句进行语句诊断，得到第一中间诊断结果；使用所述基于统计机器翻译的模型对所述目标语句进行语句诊断，得到多个第二中间诊断结果；使用所述基于神经机器翻译的模型对所述目标语句进行语句诊断，得到多个第三中间诊断结果；

所述处理步骤还包括:通过预定合并方式对多个所述第二中间诊断结果进行合并得到第一合并结果;和/或,通过所述预定合并方式对多个所述第三中间诊断结果进行合并得到第二合并结果;

将得到的中间诊断结果进行合并,得到所述目标语句的诊断结果,包括:将所述第一中间诊断结果、所述第一合并结果和所述第二合并结果中的至少两个进行合并,得到所述诊断结果。

语句诊断方法、装置和系统

技术领域

[0001] 本发明涉及语言信息处理领域，具体而言，涉及一种语句诊断方法、装置和系统。

背景技术

[0002] 语法诊断在写作和翻译中，用于高速用户语句中可能存在的错误以及对应的修改方式，这在母语的写作上、外国外对中文的学习等多方面都具有很强的需求。

[0003] 中文有各种复杂的变化，很容易出现语法错误，例如：(1)中国已成了世界拥有最多“烟民”的国家。错误：缺字，应为：中国已成了世界上拥有最多“烟民”的国家。(2)孩子的教育不能只靠一个学校老师。错误：多字，应为：孩子的教育不能只靠一个老师。(3)父母对孩子的爱情是最重要的。错误：用词错误，应为：父母对孩子的关爱是最重要的。(4)生产率较低，那肯定价格要上升。错误：词语顺序错误，应为：生产率较低，那价格肯定要上升。

[0004] 传统的语法诊断方法通常是针对某一类特定的错误来以语言学的知识来设计规则，例如：针对语序方面就会根据每个词的词性来进行语法诊断，如果遇到动词后接介词那么就判断这个语法是错误的。这种判断方式存在的问题是，灵活性不够高，而且每种模型只能针对一种特定类型的错误，以及会有较高的误伤率。

[0005] 还有一种传统的语法诊断方法是基于分类器的方法，这种方法虽然可以训练多个分类器来针对多种语法错误，但是它是假设所有的错误之间没有关联，也就是对于较为复杂的错误，比如两个错误之间是有关联的情况，这种传统的基于分类器的方法无法识别。

[0006] 针对上述的问题，目前尚未提出有效的解决方案。

发明内容

[0007] 本发明实施例提供了一种语句诊断方法、装置和系统，以至少解决现有技术中语句诊断系统的效果不佳的技术问题。

[0008] 根据本发明实施例的一个方面，提供了一种语句诊断方法，包括：使用至少一种语句诊断模型对目标语句进行语句诊断，得到至少一种类型的中间诊断结果，其中，所述语句诊断模型至少包括如下至少一个：基于规则的模型、基于统计机器翻译的模型和基于神经机器翻译的模型；将得到的中间诊断结果进行合并，得到所述目标语句的诊断结果。

[0009] 根据本发明实施例的另一方面，还提供了一种语言诊断装置，包括：诊断模块，用于使用至少一种语句诊断模型对目标语句进行语句诊断，得到至少一种类型的中间诊断结果，其中，所述语句诊断模型至少包括如下至少一个：基于规则的模型、基于统计机器翻译的模型和基于神经机器翻译的模型；合并模块，用于将得到的中间诊断结果进行合并，得到所述目标语句的诊断结果。

[0010] 根据本发明实施例的另一方面，还提供了一种语句诊断系统，包括：处理器；以及存储器，与所述处理器连接，用于为所述处理器提供处理以下处理步骤的指令：使用至少一种语句诊断模型对目标语句进行语句诊断，得到至少一种类型的中间诊断结果，其中，所述语句诊断模型至少包括如下至少一个：基于规则的模型、基于统计机器翻译的模型和基于

神经机器翻译的模型;将得到的中间诊断结果进行合并,得到所述目标语句的诊断结果。

[0011] 现有技术中使用单一的语句诊断方式对语句进行诊断,只能识别或纠正一种特定类型的错误,且当句子较为复杂时,还会有较高的误伤率。上述方案使用多种语句诊断模型并行对目标语句进行诊断,并将诊断结果进行结合,由于三种不同的诊断模型能够诊断出目标语句多种类型的错误,因此将三种语句诊断模型的诊断结果进行合并,不仅能够发现目标语句中更多的较为复杂的错误,还能减少对目标的误伤率,因此具有较高的灵活程度和语句诊断效果。

[0012] 由此,本申请上述实施例现有技术中语句诊断系统的效果不佳的技术问题。

附图说明

[0013] 此处所说明的附图用来提供对本发明的进一步理解,构成本申请的一部分,本发明的示意性实施例及其说明用于解释本发明,并不构成对本发明的不当限定。在附图中:

[0014] 图1示出了一种用于实现语句诊断方法的计算机终端(或移动设备)的硬件结构框图;

[0015] 图2是根据本申请实施例1的语句诊断方法的流程图;

[0016] 图3是根据本申请实施例1的一种语句诊断的示意图;

[0017] 图4是根据本申请实施例1的一种语句诊断模型的示意图;

[0018] 图5是根据本申请实施例1的一种合并中间诊断结果的示意图;

[0019] 图6是根据本申请实施例2的一种语句诊断装置的示意图;以及

[0020] 图7是根据本申请实施例4的一种计算机终端的结构框图。

具体实施方式

[0021] 为了使本技术领域的人员更好地理解本发明方案,下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分的实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都应当属于本发明保护的范围。

[0022] 需要说明的是,本发明的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的数据在适当情况下可以互换,以便这里描述的本发明的实施例能够以除了在这里图示或描述的那些以外的顺序实施。此外,术语“包括”和“具有”以及他们的任何变形,意图在于覆盖不排他的包含,例如,包含了一系列步骤或单元的过程、方法、系统、产品或设备不必限于清楚地列出的那些步骤或单元,而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0023] 首先,在对本申请实施例进行描述的过程中出现的部分名词或术语适用于如下解释:

[0024] Chineseas Foreign Language(CFL):非中文母语的人,将中文当作外语在学习。

[0025] Grammatical Error Correction(GEC):语法诊断,纠正语法中的错误。

[0026] Statistical Machine Translation(SMT):统计机器翻译模型,是一种基于统计

方法的翻译模型。

[0027] Neural Machine Translation(NMT):神经机器翻译模型,是一种基于神经网络的翻译模型。

[0028] Encoder:编码器,本申请中的Encoder将序列编码为一个状态向量。

[0029] Decoder:解码器,本申请中的Decoder将一个状态向量解码为一个词表中的词。

[0030] 实施例1

[0031] 根据本发明实施例,还提供了一种语句诊断方法的实施例,需要说明的是,在附图的流程图示出的步骤可以在诸如一组计算机可执行指令的计算机系统中执行,并且,虽然在流程图中示出了逻辑顺序,但是在某些情况下,可以以不同于此处的顺序执行所示出或描述的步骤。

[0032] 本申请实施例一所提供的方法实施例可以在移动终端、计算机终端或者类似的运算装置中执行。图1示出了一种用于实现语句诊断方法的计算机终端(或移动设备)的硬件结构框图。如图1所示,计算机终端10(或移动设备10)可以包括一个或多个(图中采用102a、102b,……,102n来示出)处理器102(处理器102可以包括但不限于微处理器MCU或可编程逻辑器件FPGA等的处理装置)、用于存储数据的存储器104、以及用于通信功能的传输模块106。除此以外,还可以包括:显示器、输入/输出接口(I/O接口)、通用串行总线(USB)端口(可以作为I/O接口的端口中的一个端口被包括)、网络接口、电源和/或相机。本领域普通技术人员可以理解,图1所示的结构仅为示意,其并不对上述电子装置的结构造成限定。例如,计算机终端10还可包括比图1中所示更多或者更少的组件,或者具有与图1所示不同的配置。

[0033] 应当注意到的是上述一个或多个处理器102和/或其他数据处理电路在本文中通常可以被称为“数据处理电路”。该数据处理电路可以全部或部分的体现为软件、硬件、固件或其他任意组合。此外,数据处理电路可为单个独立的处理模块,或全部或部分的结合到计算机终端10(或移动设备)中的其他元件中的任意一个内。如本申请实施例中所涉及到的,该数据处理电路作为一种处理器控制(例如与接口连接的可变电阻终端路径的选择)。

[0034] 存储器104可用于存储应用软件的软件程序以及模块,如本发明实施例中的语句诊断方法对应的程序指令/数据存储装置,处理器102通过运行存储在存储器104内的软件程序以及模块,从而执行各种功能应用以及数据处理,即实现上述的应用程序的漏洞检测方法。存储器104可包括高速随机存储器,还可包括非易失性存储器,如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中,存储器104可进一步包括相对于处理器102远程设置的存储器,这些远程存储器可以通过网络连接至计算机终端10。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。

[0035] 传输装置106用于经由一个网络接收或者发送数据。上述的网络具体实例可包括计算机终端10的通信供应商提供的无线网络。在一个实例中,传输装置106包括一个网络适配器(Network Interface Controller,NIC),其可通过基站与其他网络设备相连从而可与互联网进行通讯。在一个实例中,传输装置106可以为射频(Radio Frequency,RF)模块,其用于通过无线方式与互联网进行通讯。

[0036] 显示器可以例如触摸屏式的液晶显示器(LCD),该液晶显示器可使得用户能够与计算机终端10(或移动设备)的用户界面进行交互。

[0037] 此处需要说明的是,在一些可选实施例中,上述图1所示的计算机设备(或移动设备)可以包括硬件元件(包括电路)、软件元件(包括存储在计算机可读介质上的计算机代码)、或硬件元件和软件元件两者的结合。应当指出的是,图1仅为特定具体实例的一个实例,并且旨在示出可存在于上述计算机设备(或移动设备)中的部件的类型。

[0038] 在上述运行环境下,本申请提供了如图2所示的语句诊断方法。图2是根据本申请实施例1的语句诊断方法的流程图。

[0039] 步骤S21,使用至少一种语句诊断模型对目标语句进行语句诊断,得到至少一种类型的中间诊断结果,其中,语句诊断模型至少包括如下至少一个:基于规则的模型、基于统计机器翻译的模型和基于神经机器翻译的模型。

[0040] 具体的,上述目标语句可以是需要被诊断的中文语句,例如,非中文母语的人在学习中文时书写的中文语句,待出版或待发送的文件中的中文语句等。

[0041] 上述语句诊断模型为预先获取的诊断模型,用于对目标语句的语法、错字、别字等多种语种中的错误进行诊断,其中,上述基于规则的模型(Rule-based)依赖于预先设定的规则,可以是由语言学家等专业人士开发的语句诊断规则,基于获取的规则对语句在词汇以及句法上进行诊断。统计机器翻译(Statistical Machine Translation,SMT)的方式可以是通过大量的平行语料进行统计分析,构建翻译模型,进而使用此模型进行翻译。神经机器翻译(Neural Machine Translation,NMT)直接采用神经网络以端到端方式进行翻译建模的机器翻译方法。在本申请中,将基于统计机器翻译的模型和基于神经机器翻译的模型都用在语句诊断中,即基于统计机器翻译的模型和基于神经机器翻译的模型在训练时均使用单语语料(可以是中文语料)进行训练,输出结果也为与单语语料相同语种的语句。

[0042] 在一种可选的实施例中,可以并行的同时使用基于规则的模型、基于统计机器翻译的模型和基于神经机器翻译的模型对目标语句进行诊断,得到多个中间诊断结果。

[0043] 在另一种可选的实施例中,可以并行的使用上述三种模型中规定任意两种模型对目标语句进行诊断,得到多个中间诊断结果。

[0044] 步骤S23,将得到的中间诊断结果进行合并,得到目标语句的诊断结果。

[0045] 在上述步骤中,在得到多个中间诊断结果的情况下,将多个诊断结果进行合并,得到目标语句的诊断结果。目标语句的诊断结果是将目标语句中词汇或句法的错误进行修正后的语句。

[0046] 对于目标语句,使用不同的诊断模型输出的中间诊断结果可能不同,将多个中间诊断结果进行合并,指的是根据多种中间诊断结果对目标语句的修正,得到对目标语言的最佳修正结果。

[0047] 在合并过程中存在如下两种情况,一种是多种中间诊断结果不存在冲突的情况,一种是多种中间诊断结果存在冲突的情况。首先,在第一种情况下,多种中间诊断结果不存在冲突的情况,基于不同的诊断模型,纠正了目标语句中不同类型的错误,可以将多种中间诊断结果的并集作为多种中间诊断结果的合并结果;在第二种情况下,多种中间诊断结果存在冲突,因此需要选择其中一种作为合并结果。

[0048] 图3是根据本申请实施例1的一种语句诊断的示意图,如图3所示,将目标语句输入至语句诊断模型,语句诊断模型中包括基于规则的模型、基于统计机器翻译的模型和基于神经机器翻译的模型这三种诊断模型,每种诊断模型输出至少一个对应的中间诊断结果,

将多个中间诊断结果输入至语法诊断结果合并模块进行合并，得到最终的语法诊断结果，即目标语句对应的诊断结果。

[0049] 本申请对于写文章的中文母语者来说，可以对他们写的文章进行语法诊断，找出错别字或语病错误；对CFL人群来说，可以帮助他们在中文写作中进行语法诊断，从而帮助他们更好地学习中文；对使用搜索引擎的用户来说，可以诊断出他们在输入查询语句的存在的错误。

[0050] 此处需要说明的是，现有技术中使用单一的语句诊断方式对语句进行诊断，只能识别或纠正一种特定类型的错误，且当句子较为复杂时，还会有较高的误伤率。上述方案使用多种语句诊断模型并行对目标语句进行诊断，并将诊断结果进行结合，由于三种不同的诊断模型能够诊断出目标语句多种类型的错误，因此将三种语句诊断模型的诊断结果进行合并，不仅能够发现目标语句中更多的较为复杂的错误，还能减少对目标的误伤率，因此具有较高的灵活程度和语句诊断效果。

[0051] 由此，本申请上述实施例现有技术中语句诊断系统的效果不佳的技术问题。

[0052] 作为一种可选的实施例，使用至少一种语句诊断模型对目标语句进行语句诊断，得到至少一种类型的中间诊断结果，包括如下至少之一：使用基于规则的模型对目标语句进行语句诊断，得到第一中间诊断结果；使用基于统计机器翻译的模型对目标语；针对现有技术中语句诊断系统的效果不佳的问题，目前尚未提出有效的解决方案。

[0053] 下面，对使用上述三种语句诊断模型对语句进行诊断进行说明。

[0054] 作为一种可选的实施例，使用基于规则的模型对目标语句进行语句诊断，得到第一中间诊断结果，包括：获取预先构建的词语库；将目标语句进行分词处理，得到目标语句对应的分词结果；将分词结果中的每个词与词语库中的候选词进行比对，查找出不属于词语库中的目标词；将不属于词语库中的目标词进行变形处理，得到第一中间诊断结果。

[0055] 具体的，上述词语库可以是语言专家预先构建的，符合词汇使用规则以及中文语法的词语库。使用词语库进行语句诊断是基于超出该词语库的词语即为错误词语这一原则进行的。如果在词语库中无法查找到语句中的词语，则对查找不到的词语进行变形，以对查找不到的词语进行纠正。

[0056] 在上述方案中，以语句中的词语为粒度对语句进行诊断，基于词语库筛选出认为存在错误的词语，再通过变形对筛选出的词语进行纠正，从而得到第一诊断结果。

[0057] 图4是根据本申请实施例1的一种语句诊断模型的示意图。结合图4所示，使用基于规则的模型对语句进行诊断时，首先对语句进行拆分词，得到语句中包括的每个词，再根据词语库判断是否需要对词语进行变形，最后将需要变形的词语进行变形，变形方式具体可以包括：改变词语中汉字的顺序、与前向词语组合或与后向词语组合等方式。

[0058] 作为一种可选的实施例，将不属于词语库中的目标词进行变形处理，得到第一中间诊断结果，包括：将目标词进行多种变形处理，得到对应的多种候选诊断结果；从候选诊断结果中选择第一中间诊断结果。

[0059] 具体的，对目标词进行变形处理的方式包括多种，在上述方案中，将预设的多种变形处理方式都应用于目标词，得到多个候选诊断结果，然后从多个候选诊断结果中选择其中一个候选诊断结果作为第一中间诊断结果。

[0060] 在一种可选的实施例中，可以使用预设的语言模型对多种候选诊断结果进行打

分，选择其中的得分最高的候选诊断结果作为第一中间诊断结果。

[0061] 作为一种可选的实施例，将目标词进行多种变形处理，得到对应的多种候选诊断结果，包括如下任意一种或多种：如果目标词中包括两个及以上的字符，且将目标词中的字符改变顺序后，属于词语库中的词语，确定改变字符顺序后的目标词为候选诊断结果；如果目标词与目标语句中的前一个词语或后一个词语连接得到的第一组合属于词语库，确定第一组合为候选诊断结果；生成目标词中每个字符对应的相似字，相似字包括：形近字或音近字，如果相似字与目标语句中的前一个词语或后一个词语的第二组合属于词语库，确定第二组合为候选诊断结果。

[0062] 上述方案提出了三种对目标词进行变形处理的方式，下面依次进行说明。

[0063] 第一种方式，将目标词中的字符调整顺序。例如，目标语句为“汉字序顺并不定一影响阅读”，基于词语库对该语句进行分析，得到其中“序顺”并不属于词语库，因此将“序顺”作为目标词。将“序顺”中的两个汉字更换位置得到“顺序”，而“顺序”属于词语库，因此候选诊断结果为“顺序”。

[0064] 第二种方式，将目标词和前向的词语或后向的词语进行组合。例如：目标语句为“南京市长江大桥”，进行分析后得到的分词结果为“南京”、“市长”、“江大桥”，基于词语库对该语句进行分析，得到其中“江大桥”并不属于词语库，因此将“江大桥”作为目标词。将“江大桥”与语句中的前向字符“长”进行组合，得到“长江大桥”，而“长江大桥”属于词语库，因此候选诊断结果为“长江大桥”。

[0065] 第三种方式，基于字符的形近字或音近字对目标词进行变形。例如，目标语句为“小明的成绩非常郝”，基于词语库对该语句进行分析，得到其中“郝”并不属于词语库，因此将“郝”作为目标词。将“郝”的同音字“好”代替“郝”，将“好”与前向词语“非常”连接后得到“非常好”，而“非常好”属于词语库，因此候选诊断结果为“非常好”。

[0066] 如果目标词仅经过上述一种变形后属于词语库，则将这一种变形方式的变形结果作为第一中间诊断结果，而如果目标词经过上述多种变形后都属于词语库，则需要从中进行选择，作为一种可选的实施例，从所述候选诊断结果中选择所述第一中间诊断结果，包括：使用预设的语言模型对每个候选诊断结果进行打分，其中，打分用于表示候选诊断结果的准确程度；选择得分最高的候选诊断结果为第一中间诊断结果。

[0067] 在上述方案中，选择得分最高的候选诊断结果作为第一中间诊断结果，即为选择了准确程度最高的候选诊断结果作为第一中间诊断结果。

[0068] 上述预设的语言模型应用于各种自然语言处理问题，如语音识别、机器翻译、分词、词性标注，等等。语言模型实际为用来计算一个语句的概率的模型，即$P(W_1,W_2,...W_k)$，其中，$W_1,W_2,...W_k$用于表示句子中的每个分词。利用语言模型，可以确定哪个词序列的可能性更大。给定候选诊断结果构成的语句序列，根据语言模型即得到的打分即可以为：$P(S)=P(W_1,W_2,...W_k)=P(W_1)P(W_2|W_1)...P(W_k)P(W_2|W_1,W_2,...W_{k-1})$。

[0069] 作为一种可选的实施例，使用基于统计机器翻译的模型对目标语句进行语句诊断，得到第二中间诊断结果，包括：基于预设的短语表获取目标语句对应的候选语句；基于预设的语言模型确定候选语句的得分，其中，候选语句的得分用于表示候选语句的准确程度；将所述目标语句分别和每个所述候选语句组合，构成多个句子对；基于预设的翻译模型获取所述句子对的条件概率，其中，所述句子对的条件概率用于表示在所述候选语句生成

的情况下,生成所述目标语句的概率;;使用束搜索根据所述得分和所述条件概率得到所述第二中间诊断结果。

[0070] 基于统计机器翻译的模型在进行处理时,其实质的原理是,一个语言T由于经过一个噪声信道发生变形,从而在信道的另一端呈现出另一种语言S,基于统计机器翻译的模型根据已知的S,将S恢复为最有可能的T。应用在本申请中,可以将目标语句作为经过噪声信道而发生变形的S,基于统计机器翻译的模型需要对目标语句进行诊断,将其还原成正确的语句T。

[0071] 上述方案基于预设的短语表获取目标语句对应的候选语句,然后确定每种候选语句的概率,选择概率最大的一个候选语句作为中间第二诊断结果。在基于统计机器翻译的模型在进行处理时,每种候选语句的概率为$P(T|S)=\frac{P(T)\ P(S|T)}{P(S)}$。

[0072] 如图4所示,在一种可选的实施例中,首先在一个非常大的单语语料上训练一个语言模型,对任一个可能的候选语句e,语言模型会赋予这个句子一个分数P(e),该分数P(e)即为上述P(T)。再使用平行语料(即上述训练数据)训练翻译模型,该翻译模型会对句子对给出P(f|e),即为上述P(S|T)。最后使用噪声通道模型将语言模型和翻译模型组合起来,并且使用束搜索来求取atgmaxP(e)P(f|e),得到最终的语法诊断结果。

[0073] 作为一种可选的实施例,基于预设的短语表获取目标语句对应的候选语句,包括:对目标语句以预设粒度进行分割,得到目标语句对应的多个单位对象;在预设短语表中查找与单位对象对应的内容,并将与每个单位对象对应的内容进行排列组合,得到与目标语句对应的近似候选语句。

[0074] 具体是,上述预设粒度可以包括词粒度和字粒度,对应的单位对象即为词语和汉字,在短语表中查找词语或汉字,并将查找的结果进行排列组合,即可得到多个候选语句。

[0075] 作为一种可选的实施例,使用基于神经机器翻译的模型对目标语句进行语句诊断,得到第三中间诊断结果,包括:通过基于神经机器翻译的模型中的编码器对目标语句进行编码,以预测出目标语句对应的状态向量;通过基于神经机器翻译的模型中的解码器对状态向量进行解码,以预测目标语句对应的第三中间诊断结果。

[0076] 具体的,上述编码器(Encoder)和解码器(Decoder)均为神经网络,作为编码器的神经网络用语将目标语句中的词语编码成为稠密向量,作为解码器的神经网络用于根据稠密向量,解码出目标语句对应的正确句子,即第三中间诊断结果。

[0077] 在一种可选的实施例中,结合图4所示,基于神经机器翻译的模型在处理时,采用一种简单直观的方法完成语句的诊断,首先使用一个称为编码器(Encoder)的神经网络将目标语句编码为一个稠密向量,然后使用一个称为解码器(Decoder)的神经网络从该向量中解码出正确的语言句子,即第三中间诊断结果。

[0078] 作为一种可选的实施例,将得到的中间诊断结果进行合并,得到所述目标语句的诊断结果,包括:对第一中间诊断结果、第二中间诊断结果和第三中间诊断结果进行合并,得到目标语句的诊断结果。

[0079] 作为一种可选的实施例,所述基于神经机器翻译的模型具有多种不同的配置参数,每个配置参数对应一个第三中间诊断结果。

[0080] 具体的,上述配置参数可以为编码器和/或解码器的参数,在上述方案中,基于神

经机器翻译的模型中的编码器和解码器均为神经网络，这两个神经网络都可以具有不同的网络参数，例如：神经网络是单向还是双向、神经网络是否进行预训练，以及用于训练神经网络的训练数据。神经网络因此在给定神经网络不同的网络参数时，会得到不同的第三中间诊断结果。因此在上述方案中，赋予基于神经机器翻译的模型以多种不同的神经网络参数，得到多个第三中间诊断结果。

[0081] 在一种可选的实施例中，如图4所示，采用四种不同的配置参数，得到四种第三中间诊断结果，分别为Cn_1,Cn_2,Cn_3,Cn_4。

[0082] 作为一种可选的实施例，基于统计机器翻译的模型对目标语句进行语句诊断时，目标语句分别以词粒度和字粒度被分割，每种分割的粒度对应一个第二中间诊断结果。

[0083] 在一种可选的实施例中，如图4所示，采用两种的粒度进行分词，从而得到两种第二中间诊断结果，分别为Cs_char和Cs_word。

[0084] 作为一种可选的实施例，对所述第一中间诊断结果、第二中间诊断结果和第三中间诊断结果进行合并，得到所述目标语句的诊断结果，包括：通过预定合并方式对多个第二中间诊断结果进行合并得到第一合并结果，并通过预定合并方式并对多个第三中间诊断结果进行合并得到第二合并结果；将第一中间诊断结果、第一合并结果和第二合并结果进行合并，得到所述目标语句的诊断结果。

[0085] 在上述方案中，第二中间诊断结果和第三中间诊断结果都为多个，因此，首先需要将多个第二中间诊断结果进行合并，并将多个第三中间诊断结果进行合并，再将多个第二中间诊断结果的合并结果、多个第三中间诊断结果的合并结果以及第一中间诊断结果进行合并。

[0086] 图5是根据本申请实施例1的一种合并中间诊断结果的示意图，在一种可选的实施例中，如图5所示，第一中间诊断结果为Cr，两个第二中间诊断结果分别为Cs_char和Cs_word，四个第三中间诊断结果分别为Cn_1,Cn_2,Cn_3,Cn_4。

[0087] 首先将Cs_char和Cs_word进行低层次合并，得到第一合并结果Cs，并将Cn_1,Cn_2,Cn_3,Cn_4进行低层次合并，即上述第一合并，得到第而合并结果Cn。再将Cr、Cs和Cn进行高层次合并，得到最终的合并结果，即最终语法诊断结果。

[0088] 作为一种可选的实施例，预定合并方式包括如下任意一种：确定预先规定的优先级最高的诊断结果为合并后的诊断结果；确定所有诊断结果的并集为合并后的诊断结果；确定所有诊断结果的交集为合并后的诊断结果；确定得分最高的结果为合并后的诊断结果，其中，通过预设语言模型对合并结果进行打分。

[0089] 上述方案提供了四种对诊断结果进行合并的方式，下面以对两种第二中间诊断结果进行合并，得到第一合并结果来进行说明。两个第而中间诊断结果分别为Cs_char和Cs_word。

[0090] 在第一种方式中，如果预先规定按词语作为分词粒度的基于统计机器翻译的模型具有最高的优先级，按词语作为分词粒度的基于统计机器翻译的模型得到的第二中间诊断结果为Cs_char，则将Cs_char作为第二合并结果。

[0091] 在第二种方式中，将Cs_char和Cs_word的并集作为第二合并结果。例如，Cs_char将“汉字的序顺并不定一能影阅响读”纠正为“汉字的顺序并不定一能影阅响读”；Cs_word将“汉字的序顺并不定一能影阅响读”纠正为“汉字的序顺并不一定能影响阅读”，则得到的

并集为“汉字的顺序并不一定能影响阅读”。

[0092] 在第三种方式中，将Cs_char和Cs_word的交集作为第二合并结果。例如，Cs_char将“汉字的序顺并不定一能影阅响读”纠正为“汉字的顺序并不一定能影响阅读”；Cs_word将“汉字的序顺并不定一能影阅响读”纠正为“汉字的顺序不一定能影响阅读”，则得到的并集为“汉字的顺序并不定一能影响阅读”。

[0093] 在第四种方式中，对Cs_char和Cs_word进行打分，打分仍然可以使用语言模型P(W_1,W_2,...W_k)，选择得分最高的第二中间诊断结果为第一合并结果。

[0094] 作为一种可选的实施例，将所述第一中间诊断结果、所述第一合并结果和所述第二合并结果进行合并，得到所述目标语句的诊断结果，包括：判断第一中间诊断结果、第一合并结果和第二合并结果是否冲突；如果第一中间诊断结果、第一合并结果和第二合并结果中的任意两项冲突，按照预定合并方式进行合并；如果第一中间诊断结果、第一合并结果和第二合并结果均冲突，则保持目标语句不进行纠正。

[0095] 在低层次组合的基础上将三个模型的中间诊断结果进行组合，即进行高层次合并。结合图5所示，即为对基于规则模型的语法诊断结果Cr、基于统计机器翻译模型的语法诊断结果Cs、基于神经机器翻译模型的语法诊断结果Cn，这三个模型进行最终的组合。在进行高层次的合并时，仍然包括存在冲突和不存在冲突两种情况。

[0096] 如果第一中间诊断结果、第一合并结果和第二合并结果不存在冲突，则将这三个诊断结果的并集作为目标语句的诊断结果。

[0097] 如果第一中间诊断结果、第一合并结果和第二合并结果存在冲突，则按照如下方案执行：如果三个语句诊断模型的诊断结果中，两个模型的诊断结果存在冲突，可以使用在低层次组合时使用的预定合并方式；如果三个语句诊断模型的结果均冲突，则认为这三个模型对目标语句进行纠错的置信度较低，因此可以保留目标语句原句，不进行纠错。

[0098] 需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本发明并不受所描述的动作顺序的限制，因为依据本发明，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于优选实施例，所涉及的动作和模块并不一定是本发明所必须的。

[0099] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到根据上述实施例的方法可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件，但很多情况下前者是更佳的实施方式。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质(如ROM/RAM、磁碟、光盘)中，包括若干指令用以使得一台终端设备(可以是手机，计算机，服务器，或者网络设备等)执行本发明各个实施例所述的方法。

[0100] 实施例2

[0101] 根据本发明实施例，还提供了一种用于实施上述语句诊断方法的语句诊断装置，

[0102] 图6是根据本申请实施例2的一种语句诊断装置的示意图，如图6所示，该装置600包括：

[0103] 诊断模块602，用于使用至少一种语句诊断模型对目标语句进行语句诊断，得到至少一种类型的中间诊断结果，其中，语句诊断模型至少包括如下至少一个：基于规则的模

型、基于统计机器翻译的模型和基于神经机器翻译的模型。

[0104] 合并模块604，用于将得到的中间诊断结果进行合并，得到目标语句的诊断结果。

[0105] 此处需要说明的是，上述诊断模块602和合并模块604对应于实施例1中的步骤S21至步骤S23，两个模块与对应的步骤所实现的实例和应用场景相同，但不限于上述实施例一所公开的内容。需要说明的是，上述模块作为装置的一部分可以运行在实施例一提供的计算机终端10中。

[0106] 作为一种可选的实施例，诊断模块包括如下至少之一：第一诊断子模块，用于使用基于规则的模型对目标语句进行语句诊断，得到第一中间诊断结果；第二诊断子模块，用于使用基于统计机器翻译的模型对目标语句进行语句诊断，得到第二中间诊断结果；第三诊断子模块，用于使用基于神经机器翻译的模型对目标语句进行语句诊断，得到第三中间诊断结果。

[0107] 作为一种可选的实施例，第一诊断子模块包括：获取单元，用于获取预先构建的词语库；处理单元，用于将目标语句进行分词处理，得到目标语句对应的分词结果；查找模块，用于将分词结果中的每个词与词语库中的候选词进行比对，查找出不属于词语库中的目标词；变形单元，用于将不属于词语库中的目标词进行变形处理，得到第一中间诊断结果。

[0108] 作为一种可选的实施例，变形单元包括：变形子单元，用于将目标词进行多种变形处理，得到对应的多种候选诊断结果；选择子单元，用于从候选诊断结果中选择第一中间诊断结果。

[0109] 作为一种可选的实施例，变形子单元包括如下任意一种或多种：第一确定子单元，用于如果目标词中包括两个及以上的字符，且将目标词中的字符改变顺序后，属于词语库中的词语，确定改变字符顺序后的目标词为候选诊断结果；第二确定子单元，用于如果目标词与目标语句中的前一个词语或后一个词语连接得到的第一组合属于词语库，确定第一组合为候选诊断结果；第三确定子单元，用于生成目标词中每个字符对应的相似字，相似字包括：形近字或音近字，如果相似字与目标语句中的前一个词语或后一个词语的第二组合属于词语库，确定第二组合为候选诊断结果。

[0110] 作为一种可选的实施例，选择子单元包括：打分子单元，用于使用预设的语言模型对每个候选诊断结果进行打分，其中，打分用于表示候选诊断结果的准确程度；诊断结果选择子单元，用于选择得分最高的候选诊断结果为第一中间诊断结果。

[0111] 作为一种可选的实施例，第二诊断子模块包括：第一获取单元，用于基于预设的短语表获取目标语句对应的候选语句；确定单元，用于基于预设的语言模型确定候选语句的得分，其中，候选语句的得分用于表示候选语句的准确程度；组合单元，用于将目标语句分别和每个候选语句组合，构成多个句子对；第二获取单元，用于基于预设的翻译模型获取句子对的条件概率，其中，句子对的条件概率用于表示在候选语句生成的情况下，生成目标语句的概率；第三获取单元，用于使用束搜索根据得分和条件概率得到第二中间诊断结果。

[0112] 作为一种可选的实施例，第一获取单元包括：分割子单元，用于对目标语句以预设粒度进行分割，得到目标语句对应的多个单位对象；排列子单元，用于在预设短语表中查找与单位对象对应的内容，并将与每个单位对象对应的内容进行排列组合，得到与目标语句对应的候选语句。

[0113] 作为一种可选的实施例，第三诊断子模块包括：编码单元，用于通过基于神经机器

翻译的模型中的编码器对目标语句进行编码，以预测出目标语句对应的状态向量；解码单元，用于通过基于神经机器翻译的模型中的解码器对状态向量进行解码，以预测目标语句对应的第三中间诊断结果。

[0114] 作为一种可选的实施例，合并模块还用于对第一中间诊断结果、第二中间诊断结果和第三中间诊断结果进行合并，得到目标语句的诊断结果。

[0115] 作为一种可选的实施例，基于神经机器翻译的模型具有多种不同的配置参数，每个配置参数对应一个第三中间诊断结果。

[0116] 作为一种可选的实施例，基于统计机器翻译的模型对目标语句进行语句诊断时，目标语句分别以词粒度和字粒度被分割，每种分割的粒度对应一个第二中间诊断结果。

[0117] 作为一种可选的实施例，合并模块包括：第一合并子模块，用于通过预定合并方式对多个第二中间诊断结果进行合并得到第一合并结果，并通过预定合并方式并对多个第三中间诊断结果进行合并得到第二合并结果；第二合并子模块，用于将第一中间诊断结果、第一合并结果和第二合并结果进行合并，得到目标语句的诊断结果。

[0118] 作为一种可选的实施例，预定合并方式包括如下任意一种：确定预先规定的优先级最高的诊断结果为合并后的诊断结果；确定所有诊断结果的并集为合并后的诊断结果；确定所有诊断结果的交集为合并后的诊断结果；确定得分最高的结果为合并后的诊断结果，其中，通过预设语言模型对合并结果进行打分。

[0119] 作为一种可选的实施例，第二合并子模块包括：判断单元，用于判断第一中间诊断结果、第一合并结果和第二合并结果是否冲突；合并单元，用于如果第一中间诊断结果、第一合并结果和第二合并结果中的任意两项冲突，按照预定合并方式进行合并；保持单元，用于如果第一中间诊断结果、第一合并结果和第二合并结果均冲突，则保持目标语句不进行纠正。

[0120] 实施例3

[0121] 本发明的实施例可以提供一种语句诊断系统，包括：

[0122] 处理器；以及

[0123] 存储器，与处理器连接，用于为处理器提供处理以下处理步骤的指令：使用至少一种语句诊断模型对目标语句进行语句诊断，得到至少一种类型的中间诊断结果，其中，语句诊断模型至少包括如下至少一个：基于规则的模型、基于统计机器翻译的模型和基于神经机器翻译的模型；将得到的中间诊断结果进行合并，得到目标语句的诊断结果。

[0124] 进一步地，存储器还为处理器提供了处理实施例1中其他步骤的指令，此处不再赘述。

[0125] 实施例4

[0126] 本发明的实施例可以提供一种计算机终端，该计算机终端可以是计算机终端群中的任意一个计算机终端设备。可选地，在本实施例中，上述计算机终端也可以替换为移动终端等终端设备。

[0127] 可选地，在本实施例中，上述计算机终端可以位于计算机网络的多个网络设备中的至少一个网络设备。

[0128] 在本实施例中，上述计算机终端可以执行应用程序的漏洞检测方法中以下步骤的程序代码：使用至少一种语句诊断模型对目标语句进行语句诊断，得到至少一种类型的中

间诊断结果，其中，语句诊断模型至少包括如下至少一个：基于规则的模型、基于统计机器翻译的模型和基于神经机器翻译的模型；将得到的中间诊断结果进行合并，得到目标语句的诊断结果。

[0129] 可选地，图7是根据本申请实施例4的一种计算机终端的结构框图。如图7所示，该计算机终端A可以包括：一个或多个(图中仅示出一个)处理器702、存储器704、以及外设接口706。

[0130] 其中，存储器可用于存储软件程序以及模块，如本发明实施例中的安全漏洞检测方法和装置对应的程序指令/模块，处理器通过运行存储在存储器内的软件程序以及模块，从而执行各种功能应用以及数据处理，即实现上述的系统漏洞攻击的检测方法。存储器可包括高速随机存储器，还可以包括非易失性存储器，如一个或者多个磁性存储装置、闪存、或者其他非易失性固态存储器。在一些实例中，存储器可进一步包括相对于处理器远程设置的存储器，这些远程存储器可以通过网络连接至终端A。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。

[0131] 处理器可以通过传输装置调用存储器存储的信息及应用程序，以执行下述步骤：使用至少一种语句诊断模型对目标语句进行语句诊断，得到至少一种类型的中间诊断结果，其中，语句诊断模型至少包括如下至少一个：基于规则的模型、基于统计机器翻译的模型和基于神经机器翻译的模型；将得到的中间诊断结果进行合并，得到目标语句的诊断结果。

[0132] 可选的，上述处理器还可以执行如下步骤的程序代码：使用至少一种语句诊断模型对目标语句进行语句诊断，得到至少一种类型的中间诊断结果，包括如下至少之一：使用基于规则的模型对目标语句进行语句诊断，得到第一中间诊断结果；使用基于统计机器翻译的模型对目标语句进行语句诊断，得到第二中间诊断结果；使用基于神经机器翻译的模型对目标语句进行语句诊断，得到第三中间诊断结果。

[0133] 可选的，上述处理器还可以执行如下步骤的程序代码：获取预先构建的词语库；将目标语句进行分词处理，得到目标语句对应的分词结果；将分词结果中的每个词与词语库中的候选词进行比对，查找出不属于词语库中的目标词；将不属于词语库中的目标词进行变形处理，得到第一中间诊断结果。

[0134] 可选的，上述处理器还可以执行如下步骤的程序代码：将目标词进行多种变形处理，得到对应的多种候选诊断结果；从候选诊断结果中选择第一中间诊断结果。

[0135] 可选的，上述处理器还可以执行如下任意一种或多种步骤的程序代码：如果目标词中包括两个及以上的字符，且将目标词中的字符改变顺序后，属于词语库中的词语，确定改变字符顺序后的目标词为候选诊断结果；如果目标词与目标语句中的前一个词语或后一个词语连接得到的第一组合属于词语库，确定第一组合为候选诊断结果；生成目标词中每个字符对应的相似字，相似字包括：形近字或音近字，如果相似字与目标语句中的前一个词语或后一个词语的第二组合属于词语库，确定第二组合为候选诊断结果。

[0136] 可选的，上述处理器还可以执行如下步骤的程序代码：使用预设的语言模型对每个候选诊断结果进行打分，其中，打分用于表示候选诊断结果的准确程度；选择得分最高的候选诊断结果为第一中间诊断结果。

[0137] 可选的，上述处理器还可以执行如下步骤的程序代码：基于预设的短语表获取目

标语句对应的候选语句;基于预设的语言模型确定候选语句的得分,其中,候选语句的得分用于表示候选语句的准确程度;将目标语句分别和每个候选语句组合,构成多个句子对;基于预设的翻译模型获取句子对的条件概率,其中,句子对的条件概率用于表示在候选语句生成的情况下,生成目标语句的概率;使用束搜索根据得分和条件概率得到第二中间诊断结果。

[0138] 可选的,上述处理器还可以执行如下步骤的程序代码:对目标语句以预设粒度进行分割,得到目标语句对应的多个单位对象;在预设短语表中查找与单位对象对应的内容,并将与每个单位对象对应的内容进行排列组合,得到与目标语句对应的候选语句。

[0139] 可选的,上述处理器还可以执行如下步骤的程序代码:通过基于神经机器翻译的模型中的编码器对目标语句进行编码,以预测出目标语句对应的状态向量;通过基于神经机器翻译的模型中的解码器对状态向量进行解码,以预测目标语句对应的第三中间诊断结果。

[0140] 可选的,上述处理器还可以执行如下步骤的程序代码:对第一中间诊断结果、第二中间诊断结果和第三中间诊断结果进行合并,得到目标语句的诊断结果。

[0141] 可选的,上述处理器还可以执行如下步骤的程序代码:基于神经机器翻译的模型具有多种不同的配置参数,每个配置参数对应一个第三中间诊断结果。

[0142] 可选的,上述处理器还可以执行如下步骤的程序代码:基于统计机器翻译的模型对目标语句进行语句诊断时,目标语句分别以词粒度和字粒度被分割,每种分割的粒度对应一个第二中间诊断结果。

[0143] 可选的,上述处理器还可以执行如下步骤的程序代码:通过预定合并方式对多个第二中间诊断结果进行合并得到第一合并结果,并通过预定合并方式并对多个第三中间诊断结果进行合并得到第二合并结果;将第一中间诊断结果、第一合并结果和第二合并结果进行合并,得到目标语句的诊断结果。

[0144] 可选的,上述处理器还可以执行如下步骤的程序代码:预定合并方式包括如下任意一种:确定预先规定的优先级最高的诊断结果为合并后的诊断结果;确定所有诊断结果的并集为合并后的诊断结果;确定所有诊断结果的交集为合并后的诊断结果;确定得分最高的结果为合并后的诊断结果,其中,通过预设语言模型对合并结果进行打分。

[0145] 可选的,上述处理器还可以执行如下步骤的程序代码:判断第一中间诊断结果、第一合并结果和第二合并结果是否冲突;如果第一中间诊断结果、第一合并结果和第二合并结果中的任意两项冲突,按照预定合并方式进行合并;如果第一中间诊断结果、第一合并结果和第二合并结果均冲突,则保持目标语句不进行纠正。

[0146] 此处需要说明的是,现有技术中使用单一的语句诊断方式对语句进行诊断,只能识别或纠正一种特定类型的错误,且当句子较为复杂时,还会有较高的误伤率。上述方案使用多种语句诊断模型并行对目标语句进行诊断,并将诊断结果进行结合,由于三种不同的诊断模型能够诊断出目标语句多种类型的错误,因此将三种语句诊断模型的诊断结果进行合并,不仅能够发现目标语句中更多的较为复杂的错误,还能减少对目标的误伤率,因此具有较高的灵活程度和语句诊断效果。

[0147] 由此,本申请上述实施例现有技术中语句诊断系统的效果不佳的技术问题。

[0148] 本领域普通技术人员可以理解,图7所示的结构仅为示意,计算机终端也可以是智

能手机(如Android手机、iOS手机等)、平板电脑、掌声电脑以及移动互联网设备(Mobile Internet Devices,MID)、PAD等终端设备。图7其并不对上述电子装置的结构造成限定。例如,计算机终端A还可包括比图7中所示更多或者更少的组件(如网络接口、显示装置等),或者具有与图7所示不同的配置。

[0149] 本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令终端设备相关的硬件来完成,该程序可以存储于一计算机可读存储介质中,存储介质可以包括:闪存盘、只读存储器(Read-Only Memory,ROM)、随机存取器(Random Access Memory,RAM)、磁盘或光盘等。

[0150] 实施例5

[0151] 本发明的实施例还提供了一种存储介质。可选地,在本实施例中,上述存储介质可以用于保存上述实施例一所提供的语句诊断方法所执行的程序代码。

[0152] 可选地,在本实施例中,上述存储介质可以位于计算机网络中计算机终端群中的任意一个计算机终端中,或者位于移动终端群中的任意一个移动终端中。

[0153] 可选地,在本实施例中,存储介质被设置为存储用于执行以下步骤的程序代码:使用至少一种语句诊断模型对目标语句进行语句诊断,得到至少一种类型的中间诊断结果,其中,语句诊断模型至少包括如下至少一个:基于规则的模型、基于统计机器翻译的模型和基于神经机器翻译的模型;将得到的中间诊断结果进行合并,得到目标语句的诊断结果。

[0154] 上述本发明实施例序号仅仅为了描述,不代表实施例的优劣。

[0155] 在本发明的上述实施例中,对各个实施例的描述都各有侧重,某个实施例中没有详述的部分,可以参见其他实施例的相关描述。

[0156] 在本申请所提供的几个实施例中,应该理解到,所揭露的技术内容,可通过其它的方式实现。其中,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,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接口,单元或模块的间接耦合或通信连接,可以是电性或其它的形式。

[0157] 所述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。

[0158] 另外,在本发明各个实施例中的各功能单元可以集成在一个处理单元中,也可以是各个单元单独物理存在,也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现,也可以采用软件功能单元的形式实现。

[0159] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可为个人计算机、服务器或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、只读存储器(ROM,Read-Only Memory)、随机存取存

储器(RAM,Random Access Memory)、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

[0160] 以上所述仅是本发明的优选实施方式,应当指出,对于本技术领域的普通技术人员来说,在不脱离本发明原理的前提下,还可以做出若干改进和润饰,这些改进和润饰也应视为本发明的保护范围。

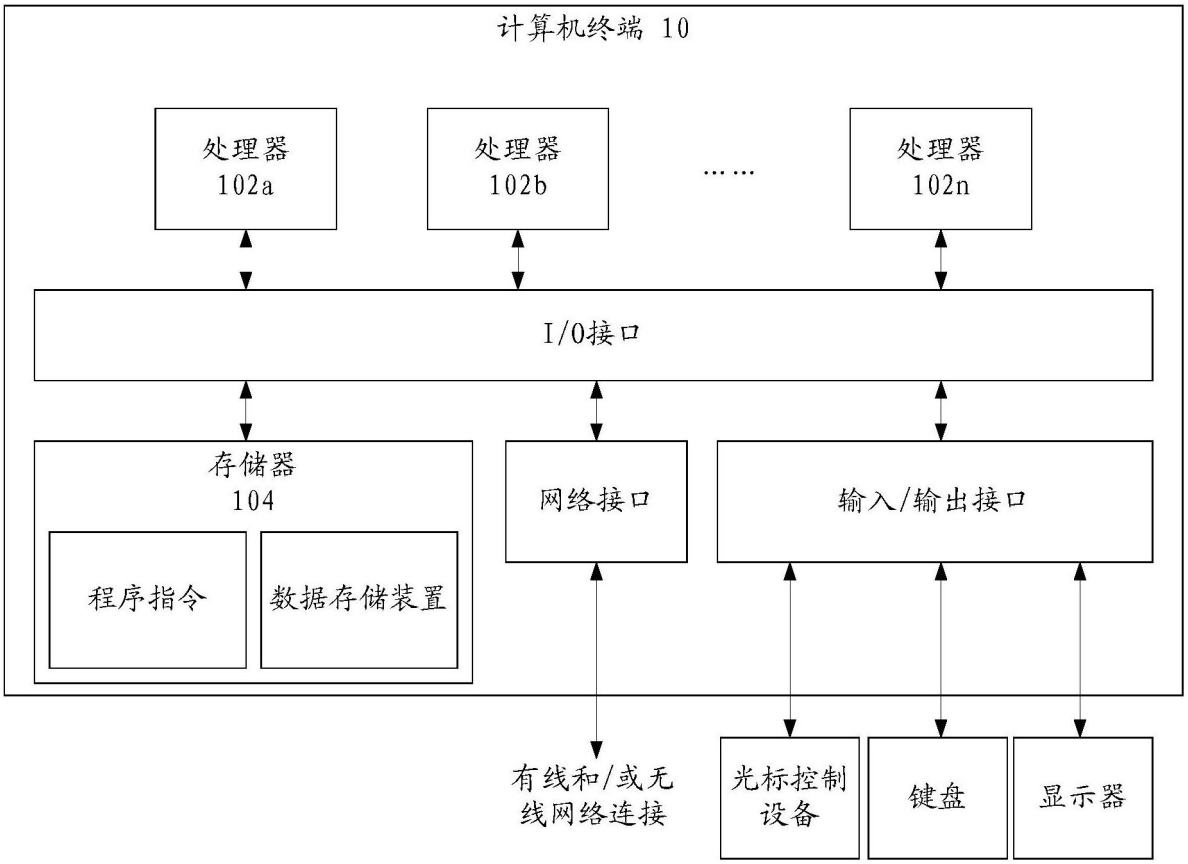

计算机终端 10
处理器
102a
处理器
102b
……
处理器
102n
I/O接口
存储器
104
程序指令
数据存储装置
网络接口
输入/输出接口
有线和/或无线网络连接
光标控制设备
键盘
显示器

图1

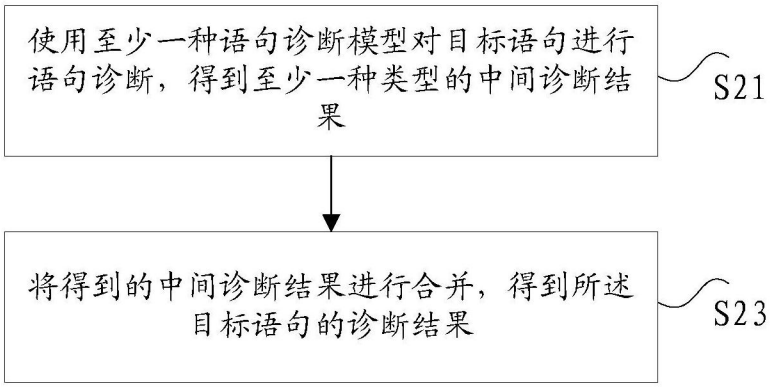

使用至少一种语句诊断模型对目标语句进行语句诊断，得到至少一种类型的中间诊断结果
S21
将得到的中间诊断结果进行合并，得到所述目标语句的诊断结果
S23

图2

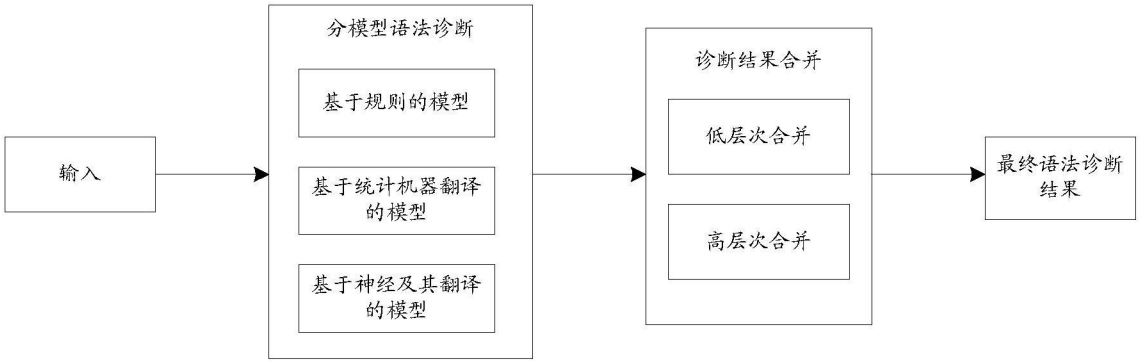

输入
分模型语法诊断
基于规则的模型
基于统计机器翻译的模型
基于神经及其翻译的模型
诊断结果合并
低层次合并
高层次合并
最终语法诊断结果

图3

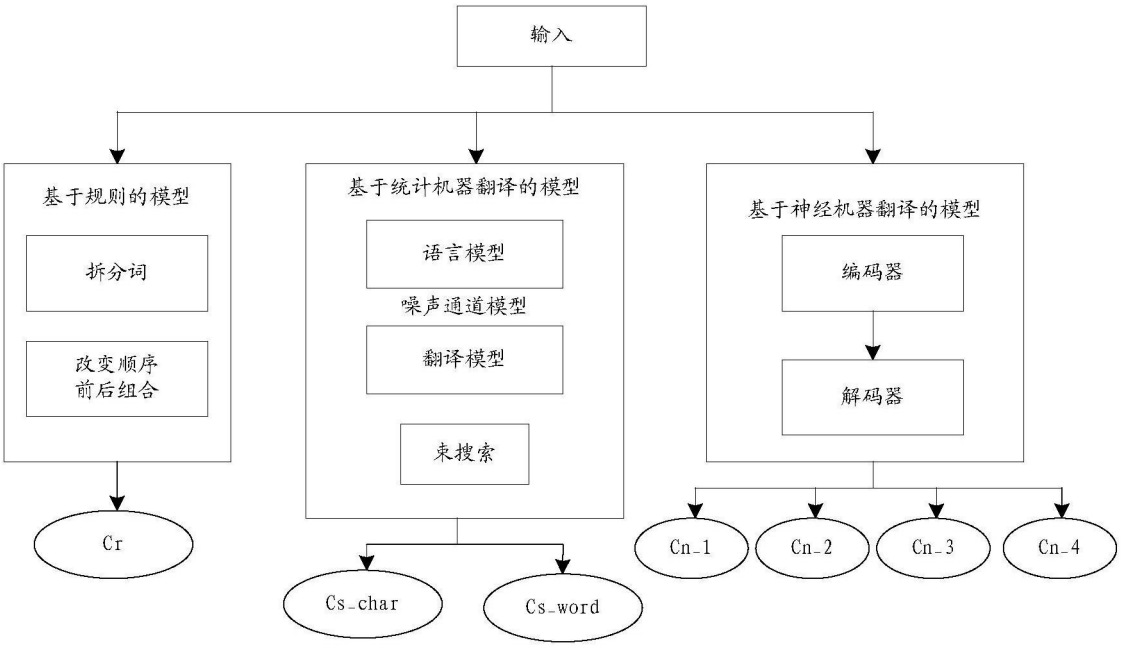

输入
基于规则的模型
拆分词
改变顺序
前后组合
Cr
基于统计机器翻译的模型
语言模型
噪声通道模型
翻译模型
束搜索
Cs_char
Cs_word
基于神经机器翻译的模型
编码器
解码器
Cn_1
Cn_2
Cn_3
Cn_4

图4

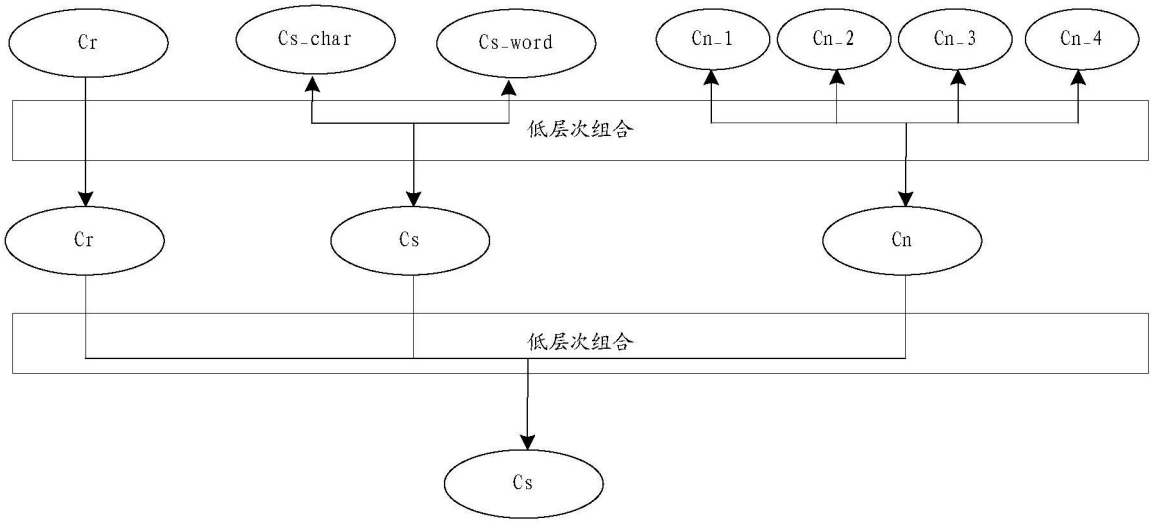

Cr
Cs_char
Cs_word
Cn_1
Cn_2
Cn_3
Cn_4
低层次组合
Cr
Cs
Cn
低层次组合
Cs

图5

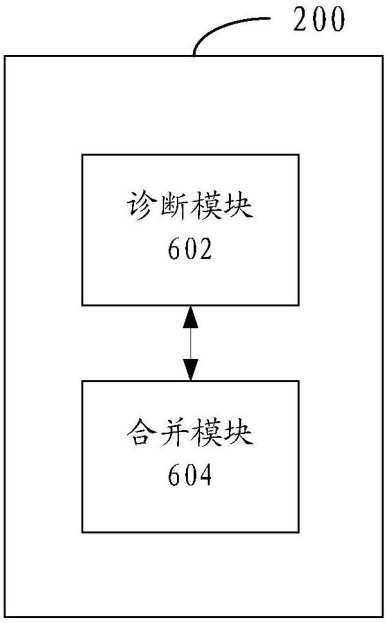

200
诊断模块
602
合并模块
604

图6

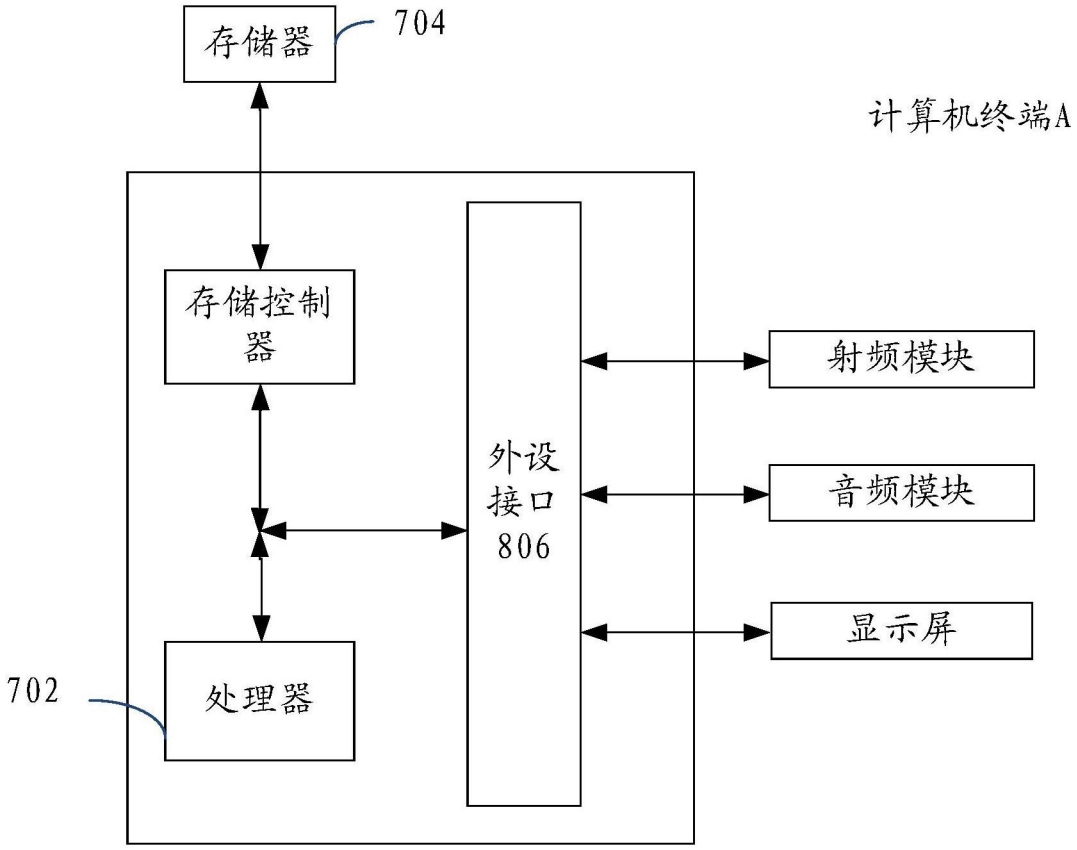


图7