(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111967266 A
(43)申请公布日 2020.11.20

(21)申请号 202010943147.X

(22)申请日 2020.09.09

(71)申请人 中国人民解放军国防科技大学
地址 410000 湖南省长沙市开福区德雅路109号

(72)发明人 王会梅 郭望舒 鲜明 刘建

(74)专利代理机构 上海上谷知识产权代理有限公司 31342
代理人 陈婷婷

(51)Int.Cl.
G06F 40/295(2020.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)
G06N 7/00(2006.01)

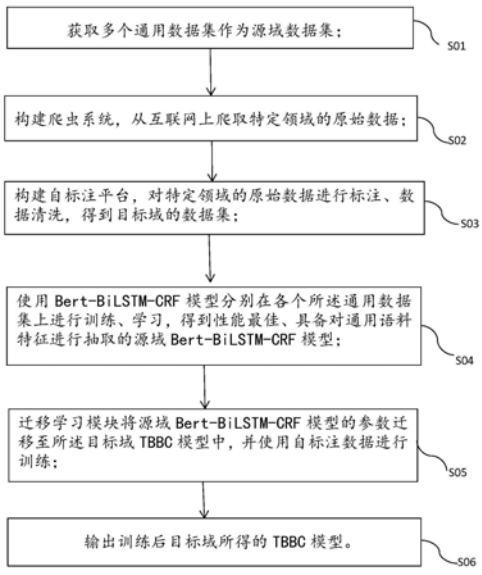
权利要求书2页 说明书9页 附图6页

(54)发明名称

中文命名实体识别模型及其构建方法和应用

(57)摘要

本发明提供一种中文命名实体识别模型及其创建方法以及应用于网络空间安全领域的方法。所述中文命名实体识别模型的应用基于迁移学习和深度神经网络，首先在中文命名实体识别领域公认的四大通用数据集上训练Bert-BiLSTM-CRF模型，充分学习到通用知识特征；而后进行模型迁移，将迁移学习后的TBBC(Trans-Bert-BiLSTM-CRF)模型在自标注的网络空间安全领域数据集上再进行训练，学习得到该领域知识的特征后并输出模型，最终得到有实际应用价值的TBBC模型，再进行中文命名实体识别。经测试可知本发明所得的TBBC模型的准确率、召回率和F1值提升明显，中文命名实体识别性能大大提高，可有效缓解在特定领域进行命名实体识别任务时训练数据不足、识别性能较低的现实困境。



CN 111967266 A

1.一种中文命名实体识别模型,其特征在于,

所述识别模型是基于Bert-BiLSTM-CRF模型的基础上增加了迁移学习模块的TBBC模型;所述Bert-BiLSTM-CRF模型从输入到输出方向依次包括Bert语言预训练模型、双向长短期记忆网络BiLSTM和条件随机场CRF层;所述迁移学习模块作用于所述Bert-BiLSTM-CRF模型;

所述Bert语言预训练模型,用于将中文词句进行词/字向量化,转化为机器可读的形式;

所述双向长短期记忆网络BiLSTM,用于将所述词/字向量进一步训练处理;

所述条件随机场CRF层用于对所述双向长短期记忆网络BiLSTM的输出结果进行解码以得到预测标注序列;

所述迁移学习模块,用于将基于通用语料训练的网络模型参数迁移至特定目标领域的新模型,并用以训练。

2.一种构建权利要求1所述的中文命名实体识别模型的方法,其特征在于,包括如下步骤:

获取多个通用数据集作为源域数据集;

在scrapy框架基础上构建爬虫系统,从互联网上爬取特定领域的原始数据;

构建自标注平台,对所爬取的特定领域的原始数据进行数据清洗,而后进行标注,得到目标域的数据集;

使用Bert-BiLSTM-CRF模型分别在所述源域数据集中的各个所述通用数据集上进行训练,充分学习到通用知识特征,得到训练后性能最佳、具备对通用语料特征进行抽取的源域Bert-BiLSTM-CRF模型;

所述迁移学习模块将所述源域Bert-BiLSTM-CRF模型的参数迁移至所述目标域TBBC模型中,然后使用自标注数据进行训练;

输出所述目标域训练后所得的TBBC模型。

3.根据权利要求2所述的构建中文命名实体识别模型的方法,其特征在于,所述自标注平台基于BRAT标注工具构建,对所述特定领域的数据进行标注的规则依据是BIO体系、BIOE体系以及BIOES体系中的一种或多种的结合。

4.根据权利要求2所述的构建中文命名实体识别模型的方法,其特征在于,所述迁移学习模块迁移过程具体为:

通过所述源域Bert-BiLSTM-CRF模型中的Bert语言预训练模型对输入的所述目标域的数据集进行词嵌入,得到其所有句子中的每个字向量;

通过将所述源域Bert-BiLSTM-CRF模型的神经网络参数迁移至所述目标域的TBBC模型的双向长短期记忆网络BiLSTM,然后将所述字向量输入所述目标域的TBBC模型中进行训练;

通过所述源域Bert-BiLSTM-CRF模型的特征标签参数迁移至所述目标域的TBBC模型的所述条件随机场CRF层,所述条件随机场CRF层将所述目标域的输出结果进行解码以得到一个预测标注序列。

5.一种将权利要求1所述的中文命名实体识别模型应用于网络空间安全领域的方法,其特征在于,包括如下步骤:

获取多个通用数据集作为源域数据集；

在scrapy框架基础上构建爬虫系统，从互联网上爬取网络空间安全领域的原始数据；

构建基于BRAT标注工具的自标注平台，并按照BIO体系对所述网络空间安全领域数据进行清洗，而后标注数据，得到目标域的数据集；

使用Bert-BiLSTM-CRF模型分别在所述源域数据集中的各个所述通用数据集上进行训练，充分学习通用知识特征，得到训练后性能最佳、具备对通用语料特征进行抽取的源域Bert-BiLSTM-CRF模型；

通过所述源域Bert-BiLSTM-CRF模型中所述Bert语言预训练模型对输入的所述目标域的数据集进行词嵌入，得到其所有句子中的每个字向量；

将所述源域Bert-BiLSTM-CRF模型里深度神经网中的神经元参数迁移至所述目标域的TBBC模型的所述双向长短期记忆网络BiLSTM中，然后将所述Bert语言预训练模型输出的字向量输入迁移后的所述目标域的TBBC模型中进行训练；

调整所述目标域的TBBC模型的所述条件随机场CRF层的输出，将所述双向长短期记忆网络BiLSTM输出的特征向量通过所述条件随机场CRF层解码为一个最优的标记序列，作为最后的预测标签输出；

对所述标记序列中的各个实体进行提取分类，完成中文实体识别。

6.根据权利要求5所述的应用方法，其特征在于，多个所述通用数据集包括人民日报数据集、微博数据集、微软亚洲研究院MSRA数据集和Chinese Literature数据集；所述中文命名实体识别模型的识别性能及迁移学习后的性能与所述通用数据集的语料类型多少和丰富程度呈正相关关系。

7.一种中文命名实体识别设备，其特征在于：包括：

一个或多个处理器；

存储器，用于存储一个或多个程序；

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求2-6中任一所述的方法。

8.一种计算机存储介质，其特征在于，所述计算机存储介质上存储有计算机程序指令，所述计算机程序指令被处理器执行时实现如权利要求2-6任意一项所述的方法。

中文命名实体识别模型及其构建方法和应用

技术领域

[0001] 本发明涉及自然语言处理领域，具体地，涉及一种基于迁移学习和深度神经网络模型的中文命名实体识别模型及其构建方法和应用。

背景技术

[0002] 命名实体识别一直以来都是信息抽取、自然语言处理等领域中重要的研究任务，主要任务是从非结构化文本中提取能体现现实世界中已存在的具体实体或者抽象实体的单词或者词组，例如人名、地名和组织机构名等，当前主流的识别目标实体为“三大类(实体类、时间类和数字类)、七小类(人名、地名、组织名、机构名、时间、日期、货币和百分比)。命名实体识别技术发展至今，已经在信息抽取、信息检索、机器翻译、问答系统、文本理解、舆情分析和知识图谱构建等诸多领域得到了广泛应用。

[0003] 命名实体识别技术发端于英文命名实体识别，而中文文本中没有词语边界符号，实体识别的效果很大程度上受制于自动分词的效果，因此，提高中文的分词效果是中文命名实体识别的前置条件。

[0004] 目前，中文命名实体识别主要有三种方法：

[0005] 一、基于规则的方法。该方法诞生于上世纪90年代，主要通过人工方法构建有限的规则库，再从待识别文本中通过规则匹配的方式识别出实体。后期研究者试图基于机器(如Bootstrapp i ng方法)自动发现和生产规则，提高制定规则的效率和效果。该方法规则制定的成本较高，因为如果要提高识别效果要求大量的规则，但显然有限的规则库无法囊括所有的实体；另外，规则对领域知识依赖极大，这使得不同领域的规则库无法移植迭代。

[0006] 二、基于统计机器学习的方法。本世纪初机器学习在自然语言处理领域兴起，为解决命名实体识别任务，研究学者提出了诸多方法，如：经典马尔科夫法(HMM)、最大熵法(ME)、条件随机场法(CRF)和支持矢量机法(SVM)，以及综合了前面几种方法的层叠马尔科夫方法、多层条件随机场方法等。统计机器学习方法存在的主要问题是识别准确率低，训练容易过拟合。

[0007] 三、基于深度神经网络的方法。近年来在解决命名实体识别任务方面，主流方法是采用深度神经网络方法。尤其是采用词向量的方法后，对自然语言处理领域的发展起到了强大的助推作用。当前针对命名实体识别任务进行研究的方法主要有基于卷积神经网络(RNN)、基于循环神经网络(RNN)、基于长短期记忆网络(LSTM)和基于图神经网络(GRU)等，并在部分领域取得了良好的效果。

[0008] 但在中文领域命名实体识别方面，训练有效的深度神经网络依赖大规模高质量的领域数据，当前在开源互联网能够获得通用数据集，但并无“网络空间安全”领域的训练数据，而直接使用现成的识别模型在通用数据集上进行训练，所得的模型识别效果并不理想，不具有应用价值。

发明内容

[0009] 针对网络空间安全领域中存在的中文命名实体识别数据缺乏、识别性能差等问题，本发明提出了一种基于迁移学习和深度神经网络的中文命名实体识别模型及其构建方法，以及其应用于网络空间安全领域进行中文命名实体识别的方法。本发明将在大规模通用数据集上训练并充分学习通用知识特征，通过迁移学习后在自标注的网络空间安全领域数据上进行训练并学习得到该领域知识的特征，所得模型识别性能明显提升，有效解决了在网络空间安全领域进行命名实体识别任务训练数据不足的现实困境。

[0010] 具体技术方案如下：

[0011] 所述识别模型是基于Bert-BiLSTM-CRF模型的基础上增加了迁移学习模块的TBBC模型；所述Bert-BiLSTM-CRF模型从输入到输出方向依次包括Bert语言预训练模型、双向长短期记忆网络BiLSTM和条件随机场CRF层；所述迁移学习模块作用于所述Bert-BiLSTM-CRF模型；

[0012] 所述Bert语言预训练模型，用于将中文词句进行词/字向量化，转化为机器可读的形式；

[0013] 所述双向长短期记忆网络BiLSTM，用于将所述词/字向量进一步训练处理；

[0014] 所述条件随机场CRF层用于对所述双向长短期记忆网络BiLSTM的输出结果进行解码以得到预测标注序列；

[0015] 所述迁移学习模块，用于将基于通用语料训练的网络模型参数迁移至特定目标领域的新模型，并用以训练。

[0016] 本发明还提供构建上述的中文命名实体识别模型的方法，具体包括如下步骤：

[0017] 获取多个通用数据集作为源域数据集；

[0018] 在scrapy框架基础上构建爬虫系统，从互联网上爬取特定领域的原始数据；

[0019] 构建自标注平台，对所述特定领域的原始数据进行清洗，而后进行数据标注，得到目标域的数据集；

[0020] 使用Bert-BiLSTM-CRF模型分别在所述源域数据集中的各个所述通用数据集上进行训练，充分学习到通用知识特征，得到训练后性能最佳、具备对通用语料特征进行抽取的源域Bert-BiLSTM-CRF模型；

[0021] 所述迁移学习模块将所述源域Bert-BiLSTM-CRF模型的参数迁移至所述目标域TBBC模型中，然后使用自标注数据进行训练；

[0022] 输出所述目标域训练后所得的TBBC模型。

[0023] 进一步的，所述自标注平台基于BRAT标注工具构建，对所述特定领域的数据进行标注的规则依据是BIO体系、BIOE体系以及BIOES体系中的一种或多种的结合。

[0024] 进一步的，所述迁移学习模块迁移过程具体为：

[0025] 通过所述源域Bert-BiLSTM-CRF模型中的Bert语言预训练模型对输入的所述目标域的数据集进行词嵌入，得到其所有句子中的每个字向量；

[0026] 通过将所述源域Bert-BiLSTM-CRF模型的神经网络参数迁移至所述目标域的TBBC模型的双向长短期记忆网络BiLSTM，然后将所述字向量输入所述目标域的TBBC模型中进行训练；

[0027] 通过所述源域Bert-BiLSTM-CRF模型的特征标签参数迁移至所述目标域的TBBC模

型的所述条件随机场CRF层,所述条件随机场CRF层将所述目标域的输出结果进行解码以得到一个预测标注序列。

[0028] 本发明还提供一种应用于网络空间安全领域的中文命名实体识别方法,包括如下步骤:

[0029] 获取多个通用数据集作为源域数据集;

[0030] 在scrapy框架基础上构建爬虫系统,从互联网上爬取网络空间安全领域的原始数据;

[0031] 构建基于BRAT标注工具的自标注平台,并按照BIO体系对所述网络空间安全领域数据进行清洗,而后标注数据,得到目标域的数据集;

[0032] 使用Bert-BiLSTM-CRF模型分别在所述源域数据集中的各个所述通用数据集上进行训练,充分学习通用知识特征,得到训练后性能最佳、具备对通用语料特征进行抽取的源域Bert-BiLSTM-CRF模型;

[0033] 通过所述源域Bert-BiLSTM-CRF模型中所述Bert语言预训练模型对输入的所述目标域的数据集进行词嵌入,得到其所有句子中的每个字向量;

[0034] 将所述源域Bert-BiLSTM-CRF模型里深度神经网中的神经元参数迁移至所述目标域的TBBC模型的所述双向长短期记忆网络BiLSTM中,然后将所述Bert语言预训练模型输出的字向量输入迁移后的所述目标域的TBBC模型中进行训练;

[0035] 调整所述目标域的TBBC模型的所述条件随机场CRF层的输出,将所述双向长短期记忆网络BiLSTM输出的特征向量通过所述条件随机场CRF层解码为一个最优的标记序列,作为最后的预测标签输出;

[0036] 对所述标记序列中的各个实体进行提取分类,完成中文实体识别。

[0037] 进一步的,所述通用数据集包括人民日报数据集、微博数据集、微软亚洲研究院MSRA数据集和ChineseLiterature数据集;所述中文命名实体识别模型的识别性能及迁移学习后性能与所述通用数据集的语料类型多少和丰富程度呈正相关关系。

[0038] 本发明还提供一种中文命名实体识别设备,包括:

[0039] 一个或多个处理器;

[0040] 存储器,用于存储一个或多个程序;

[0041] 当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现上述的方法。

[0042] 本发明还提供的一种计算机存储介质,所述计算机存储介质上存储有计算机程序指令,所述计算机程序指令被处理器执行时实现上述的方法。

[0043] 与现有技术相比,本发明的有益效果有:

[0044] 与单纯的BiLSTM-CRF模型以及使用word2vect词向量模型进行词向量化的模型相比,使用Bert语言预训练模型进行词向量化能在嵌入层就学习得到词句的语义特征,为双向长短期记忆网络BiLSTM进行特征学习奠定了基础,最终为提高识别性能起到明显作用;同时由于使用通用数据训练后再迁移至新网络,与直接使用自标注领域数据进行训练相比,提高了网络的“热启动”能力和泛化能力,进而提高了模型整体的识别性能。

附图说明

[0045] 为了更清楚地说明本发明实施例中的技术方案，下面将对实施例描述中所使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它的附图，其中：

[0046] 图1是本发明提供的中文命名实体识别模型的网络结构示意图；

[0047] 图2是所述中文命名实体识别模型的构建方法流程图；

[0048] 图3是本发明使用的的Bert语言预训练模型功能原理示意图；

[0049] 图4是LSTM模型功能原理示意图；

[0050] 图5是本发明提供的迁移学习模块功能原理示意图；

[0051] 图6是本发明提供的一个在网络空间安全领域的实施例的Trans-Bert-BiLSTM-CRF模型的构建及应用流程图；

[0052] 图7是图6中涉及的两组对比实验的结果图；其中，子图(a)为不同模型识别性能F1值变化曲线，子图(b)为不同源域数据集迁移学习性能F1值变化曲线。

具体实施方式

[0053] 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，所描述的实施例仅是本发明的一部分实施例，而不是全部的实施例。

[0054] 下面结合附图对本发明的实施方式进行详细说明。

[0055] 请参阅图1，是本发明提供的中文命名实体识别模型的网络结构示意图。所述中文命名实体识别模型是在Bert-BiLSTM-CRF模型的基础上增加了迁移学习模块的Trans-Bert-BiLSTM-CRF模型，简称TBBC模型；所述Bert-BiLSTM-CRF模型从输入到输出方向具体包括Bert语言预训练模型(简称“Bert”)、双向长短期记忆网络BiLSTM(简称“BiLSTM层”)和条件随机场CRF层(简称“CRF输出层”)；所述迁移学习模块作用于所述Bert-BiLSTM-CRF模型。所述Bert语言预训练模型，用于将中文词句进行词/字向量化，转化为机器可读的形式，便于后续网络的处理；所述双向长短期记忆网络BiLSTM，负责将所述词/字向量进一步处理；所述迁移学习模块负责将基于通用语料训练的网络模型参数迁移至新模型并用以训练；所述条件随机场CRF层负责对BiLSTM模块的输出结果进行解码以得到一个预测标注序列。

[0056] 其中，所述条件随机场CRF层概率预测模型的公式如下：

[0057] $$p(y|z;W,b)=\frac{\prod_{i=1}^{n}\varphi(y_{i-1},y_i,z)}{\sum_{y'\in y(z)}\prod_{i=1}^{n}\varphi(y'_{i-1},y'_i,z)}$$

[0058] 其中，$\varphi(y',y,z)=\exp(W_{y',y}^{T}z_i+b_{y',y})$，$W_{y',y}^{T}$和$b_{y',y}$分别表示由标签y′转移为标签y的权向量和偏差。

[0059] 请参阅图2，是所述中文命名实体识别模型的构建方法。具体包括为：

[0060] S01:获取多个通用数据集作为源域数据集；

[0061] S02:在scrapy框架基础上构建爬虫系统,从互联网上爬取特定领域的原始数据;

[0062] S03:基于BRAT标注工具构建自标注平台,对所述特定领域的原始数据按照BIO体系、BIOE体系以及BIOES体系中的一种或多种的结合进行标注,而后进行数据清洗,得到目标域的数据集;

[0063] S04:使用Bert-BiLSTM-CRF模型分别在所述源域数据集中的各个通用数据集上进行训练,充分学习到通用知识特征,得到训练后性能最佳、具备对通用语料特征进行抽取的源域Bert-BiLSTM-CRF模型;

[0064] S05:所述迁移学习模块将所述源域Bert-BiLSTM-CRF模型的参数迁移至所述目标域TBBC模型中,并使用自标注数据进行训练如下:

[0065] ①通过所述源域Bert-BiLSTM-CRF模型中的Bert语言预训练模型对输入的所述目标域的数据集进行词嵌入,得到其所有句子中的每个字向量;

[0066] ②通过将所述源域Bert-BiLSTM-CRF模型的参数迁移至所述目标域的TBBC模型的BiLSTM,然后将所述目标域的字向量输入TBBC模型中进行训练;

[0067] ③通过所述源域Bert-BiLSTM-CRF模型的参数迁移至所述目标域的TBBC模型的CRF,然后将上述目标域的输出结果进行解码以得到一个预测标注序列。

[0068] S06:输出所述目标域所得的TBBC模型,该模型具有现实的应用意义,且性能极大提高。

[0069] 图3是本发明使用的的Bert语言预训练模型功能原理示意图。本发明主要使用的是Google开源的Chinese base model通用语料模型对词进行向量化,该模型从输入到输出方向包含词嵌入层、编码层、模型层,一般有110M的参数。其中编码层最重要的自注意力公式如下:

[0070] $$Attention(Q,K,V)=soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

[0071] 其中Q,K,V均是输入字向量矩阵,d_k为输入向量维度。

[0072] 将Bert语言预训练模型应用到中文实体识别中,语言预训练是作为中文实体识别的上游任务,它把预训练出来的结果作为下游任务BiLSTM-CRF的输入,这就意味着下游主要任务是对预训练出来的词向量进行分类即可,它不仅减少了下游任务的工作量,而且能够得到更好的效果;Bert语言预训练模型不同于传统的预训练模型,Bert预训练出来的是动态词向量,能够在不同语境中表达不同的语义,相较于传统的语言预训练模型训练出来的静态词向量(无法表征一词多义),在中文实体识别中具有更大的优势。

[0073] 请参阅图4,是LSTM模型功能原理示意图。LSTM(Long-Short Term Memory,长短期记忆网络),是循环神经网络(RNN)的一种变体。它解决了RNN训练时所产生的梯度爆炸或梯度消失。LSTM模型巧妙地运用门控概念实现长期记忆,同时它也能够捕捉序列信息。LSTM模型由一个记忆单元、更新门(Update gate)、输出门(Output gate)和遗忘门(Forget gate)构成,其中记忆单元的作用是对信息进行管理和保存,更新门(Update gate)、输出门(Output gate)和遗忘门(Forget gate)的作用是控制记忆单元中信息的更新、衰减、输入和输出等动作。核心是通过学习LSTM模型中三个门的参数来管理记忆单元中的信息,从而使有用的信息经过较长的序列仍能保存在记忆单元中。LSTM模型的结构用公式表达如下

(BiLSTM同理)：

[0074] $i_t=\sigma(W_v[h_{t-1},x_t]+b_v)$

[0075] $f_t=\sigma(W_f[h_{t-1},x_t]+b_f)$

[0076] $o_t=\sigma(W_o[h_{t-1},x_t]+b_o)$

[0077] $$\tilde{c}=\tanh(W_c[h_{t-1},x_t]+b_c)$$

[0078] $$c_t=i_t^*\tilde{c}+f_t^*c_{t-1}$$

[0079] $h_t=o_t^*\tanh(c_t)$

[0080] LSTM模型在t时刻的输入由输入层X_t序列中前一个单元的隐含层h_{t-1}和记忆单元c_{t-1}两部分构成，在t时刻的输出为该单元的隐含层h_t和记忆单元c_t。首先通过计算三个门的信息输出以控制记忆单元的信息，然后计算记忆单元内的信息，最后，使用记忆单元值和输出门计算该时刻隐含层的值。σ表示sigmoid激活函数，tanh表示双曲正切激活函数，所有的W和b均为参数，i_t,f_t,o_t分别是输入门、遗忘门及输出门的输出结果。

[0081] BiLSTM(Bidirectional Long-Short Term Memory，双向长短期记忆网络)，其基本思想就是对每个词序列分别采取前向和后向LSTM模型，然后将同一个时刻的输出进行合并。因此对于每一个时刻而言，都对应着前向与后向的信息。

[0082] 请参阅图5，是本发明提供的迁移学习模块功能原理示意图。迁移学习中，本发明主要迁移了经过源域数据训练TBBC模型的所述双向长短期记忆网络BiLSTM和所述Bert语言预训练模型的所有参数，并迁移条件随机场CRF层的特征参数，修改输出层维度使其与自标注数据的标签类型个数相等。迁移学习的数学公式可简记为：

[0083] $P_t=\sum W_t(D_t)\times P_sT=\sum W_t(D_t)\sum W_s(D_s)T$

[0084] 其中，D_s、W_s分别为源域训练数据和网络训练函数，D_t、W_t为目标域训练数据和网络训练函数，T为迁移学习矩阵。

[0085] 请参阅图6，是本发明提供的一个应用在网络空间安全领域的实施例的Trans-Bert-BiLSTM-CRF模型的构建及应用流程图。

[0086] S1：首先从开源互联网上获取人民日报数据集(Github开源获取)、微博数据集、MSRA数据集(Github开源获取)和Chinese Literature数据集四个通用数据集。

[0087] S2：在scrapy框架基础上构建爬虫服务，在开源互联网上爬取与网络空间安全领域相关的文本。文本构成主要是门户网站新闻、百科网站和网络空间安全领域网站科普文章，共获取纯文本数据50M。数据经清洗后(去除无关内容、广告、乱码字符等与网络空间安全无关的文本)，按每句话为一条数据计数，共有约50720条文本数据。

[0088] S3：构建基于BRAT标注工具的自标注平台，在爬取并清洗后的数据中精选约5000条数据按照BIO体系进行标注，将标注数据处理为可训练数据。为保证数据的准确性，按10:1:1的比例随机将数据集划分为训练集、验证集和测试集以便后续进行训练、验证和测试。

[0089] S4：分别构建BiLSTM-CRF、Trans-BiLSTM-CRF、Bert-BiLSTM-CRF和Trans-Bert-BiLSTM-CRF四个网络模型，在相同的条件下(网络超参数初始值、实验硬件环境均相同)，将人民日报数据集分别在四个模型中进行训练、验证和测试，得出如图7中子图(a)的实验结果，从图7(a)可知，本发明的TBBC模型的测试集F1值最高，达到了0.9085，识别性能最佳。具体其他实施例中也可以应用其他的通用数据集，所得出的结果与人民日报数据集是一致

的。

[0090] S5:以所述四个通用数据集为源域数据集,分别在源域Bert-BiLSTM-CRF模型训练,得到四个不同的训练后的网络。将四个不同的训练后网络的参数分别迁移到目标域TBBC模型(无网络参数)中,选出迁移学习性能最优的源域数据集。从图7(b)可知,以微博数据集为源域进行迁移学习的模型F1值上升最稳定,性能最优。F1值是为了能够评价不同表征模型的性能,在精确率和召回率的基础上提出的概念,来对精确率和召回率进行整体评价,所以是行业内比较模型识别性能的主流参数。

[0091] S6:使用本发明的TBBC模型和微博源域数据集,在相同的条件下(网络超参数初始值、实验硬件环境均相同),使用自标注的网络空间安全领域数据集进行训练、验证和测试,得到有应用效益的TBBC模型及其参数,如图7中子图(b)的实验结果。从图7中子图(b)可知,以微博数据集为源域进行迁移学习的模型F1值上升最稳定,且最终测试集的F1值为最佳的0.9467,足见应用效果显著。通过观察分析其语料数据可知,微博数据集的语料实体类型一共有7种,而人民日报、MSRA和Chinese Literature三个数据集的标签分别只有3种、3种和6种,因此通过微博数据集训练出来的网络模型分类能力要好于其他三者。同时,微博数据集因既包含如人民日报里的官方语句,也包含了当下网络媒体的非正式语句,自标注的网络空间安全领域数据均从互联网上爬取,两者词句特征最接近,因此通过该语料训练的网络的识别性能也最佳。

[0092] S7:利用S6所得模型及其参数对各个实体进行提取分类,完成中文实体识别。

[0093] 实施本发明的实施例的示例设备可以包括一个或多个中央处理单元(CPU),其可以根据存储在只读存储器(ROM)中的计算机程序指令或者从存储单元加载到随机访问存储器(RAM)中的计算机程序指令,来执行各种适当的动作和处理。在RAM中,还可存储设备操作所需的各种程序和数据。CPU、ROM以及RAM通过总线彼此相连。输入/输出(I/O)接口也连接至总线。

[0094] 设备中的多个部件连接至I/O接口,包括:输入单元,例如键盘、鼠标等;输出单元,例如各种类型的显示器、扬声器等;存储单元,例如磁盘、光盘等;以及通信单元,例如网卡、调制解调器、无线通信收发机等。通信单元允许设备通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

[0095] 上文所描述的方法例如可由设备的处理单元执行。例如,在一些实施例中,方法可被实现为计算机软件程序,其被有形地包含于机器可读介质,例如存储单元。在一些实施例中,计算机程序的部分或者全部可以经由ROM和/或通信单元而被载入和/或安装到设备上。当计算机程序被加载到RAM并由CPU执行时,可以执行上文描述的方法的一个或多个动作。

[0096] 然而本领域技术人员可以理解,方法的步骤的执行并不局限于图中所示和以上所述的顺序,而是可以以任何其他合理的顺序来执行,或者可以并行执行。此外,设备也不必须包含上述所有组件,其可以仅仅包含执行本发明中所述的功能所必须的其中一些组件,并且这些组件的连接方式也可以形式多样。例如,在设备是诸如手机之类的便携式设备的情况下,可以具有与上述相比不同的结构。

[0097] 利用本发明的方案,使用Bert语言预训练模型进行词向量化能在嵌入层就学习得到词句的语义特征,为双向长短期记忆网络BiLSTM进行特征学习奠定了基础,最终为提高识别性能起到明显作用;同时由于使用通用数据训练后再迁移至新网络,与直接使用自标

注领域数据进行训练相比，提高了网络的“热启动”能力和泛化能力，进而提高了模型整体的识别性能。

[0098] 本发明可以是方法、装置、系统和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质，其上载有用于执行本发明的各个方面的计算机可读程序指令。

[0099] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括：便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身，诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波(例如，通过光纤电缆的光脉冲)、或者通过电线传输的电信号。

[0100] 这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备，或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令，并转发该计算机可读程序指令，以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0101] 用于执行本发明操作的计算机程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码，所述编程语言包括面向对象的编程语言-诸如Smalltalk、C++等，以及常规的过程式编程语言-诸如"C"语言或类似的编程语言。计算机可读程序指令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)。在一些实施例中，通过利用计算机可读程序指令的状态信息来个性化定制电子电路，例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA)，该电子电路可以执行计算机可读程序指令，从而实现本发明的各个方面。

[0102] 这些计算机可读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单元，从而生产出一种机器，使得这些指令在通过计算机或其它可编程数据处理装置的处理单元执行时，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。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0103] 也可以把计算机可读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、其它可编程数据处理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、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0104] 附图中的流程图和框图显示了根据本发明的多个实施例的系统、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、程序段或指令的一部分，所述模块、程序段或指令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。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。

[0105] 以上所述的本发明实施方式，并不构成对本发明保护范围的限定，任何在本发明精神和原则之内所作的修改、等同替换和改进等，均应包含在本发明的权利要求保护范围之内。

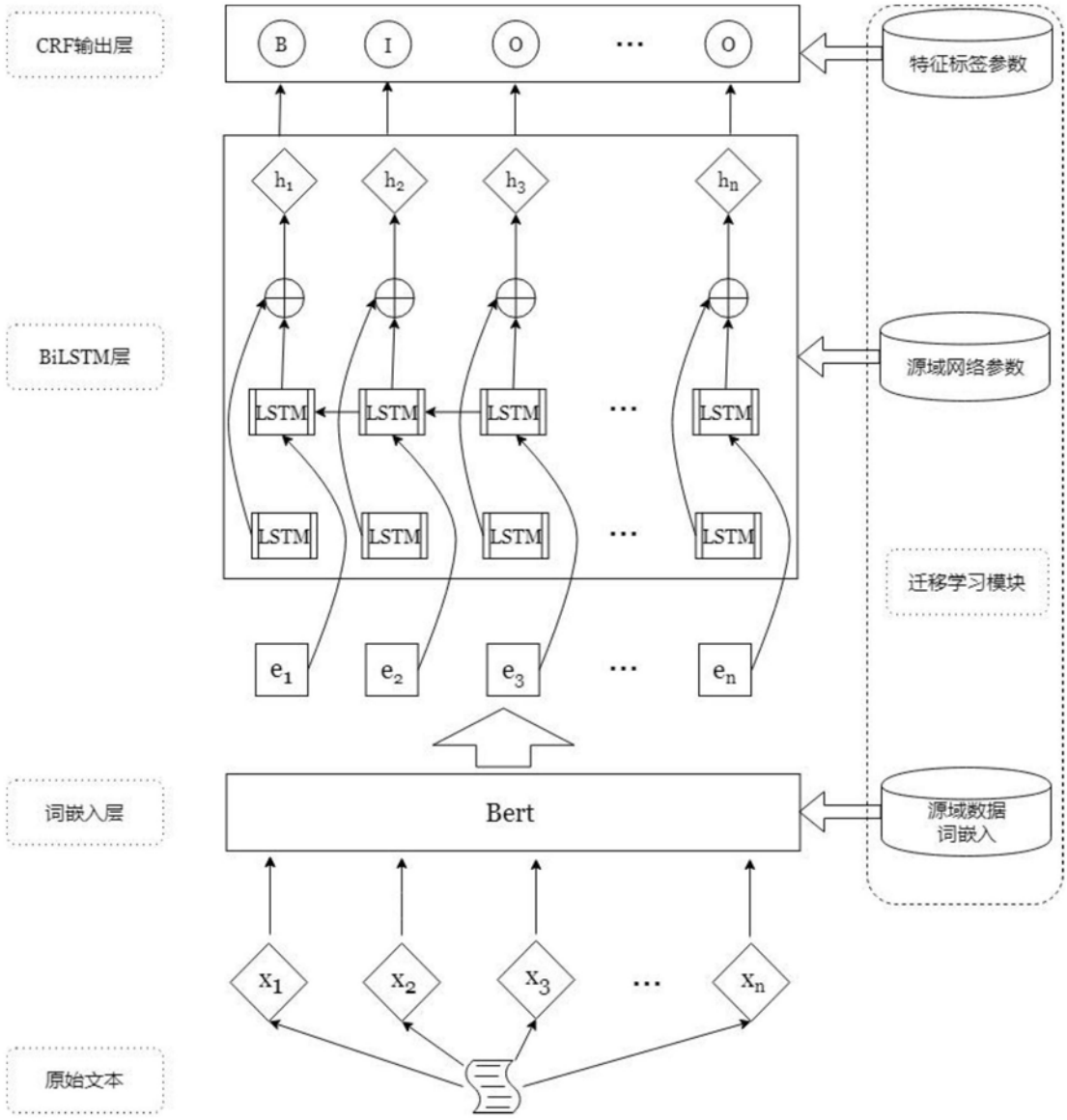

CRF输出层
B
I
O
…
O
特征标签参数
BiLSTM层
h_1
h_2
h_3
h_n
LSTM
源域网络参数
迁移学习模块
e_1
e_2
e_3
e_n
词嵌入层
Bert
源域数据
词嵌入
x_1
x_2
x_3
x_n
原始文本

图1

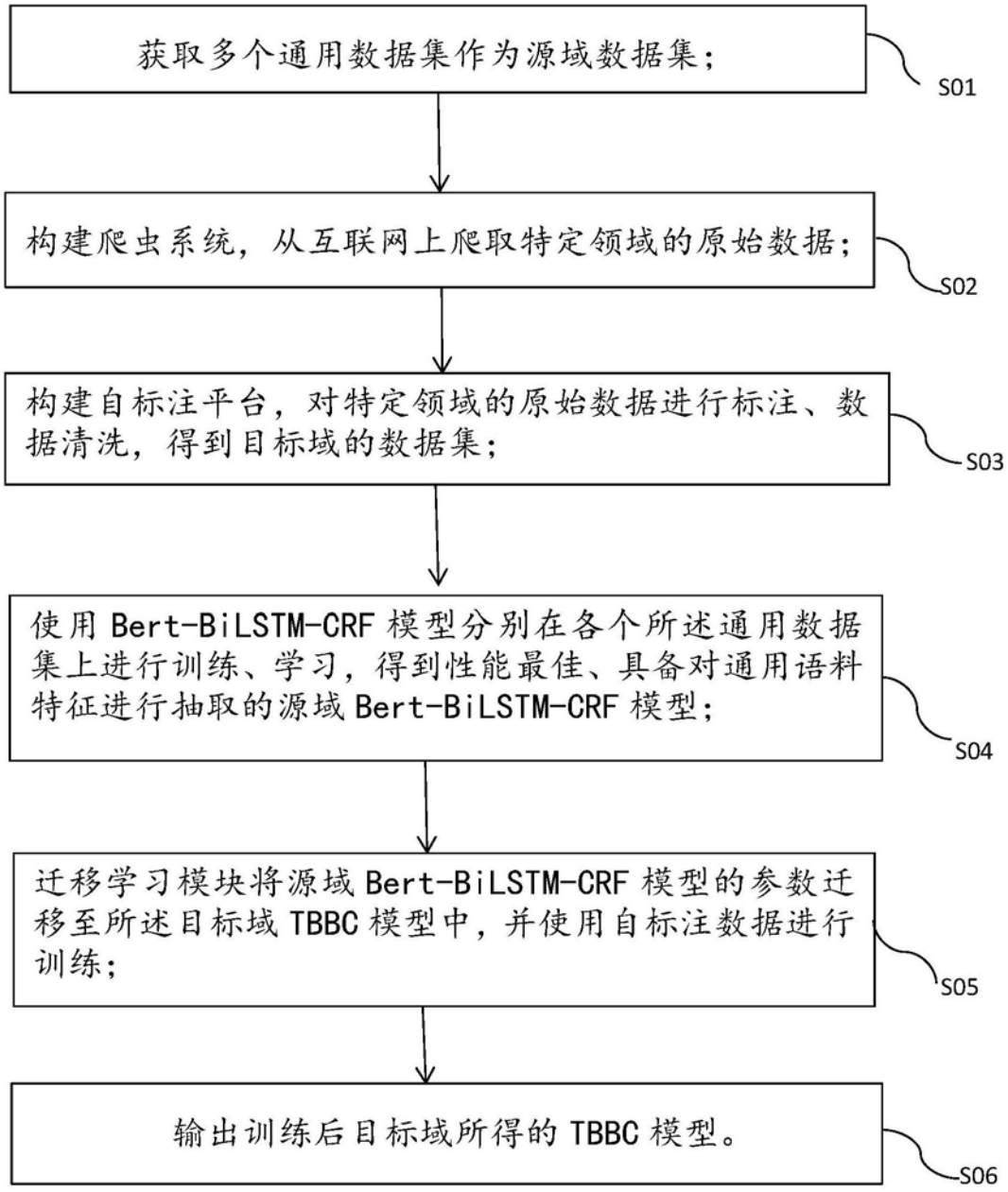

获取多个通用数据集作为源域数据集；
S01
构建爬虫系统，从互联网上爬取特定领域的原始数据；
S02
构建自标注平台，对特定领域的原始数据进行标注、数据清洗，得到目标域的数据集；
S03
使用 Bert-BiLSTM-CRF 模型分别在各个所述通用数据集上进行训练、学习，得到性能最佳、具备对通用语料特征进行抽取的源域 Bert-BiLSTM-CRF 模型；
S04
迁移学习模块将源域 Bert-BiLSTM-CRF 模型的参数迁移至所述目标域 TBBC 模型中，并使用自标注数据进行训练；
S05
输出训练后目标域所得的 TBBC 模型。
S06

图2

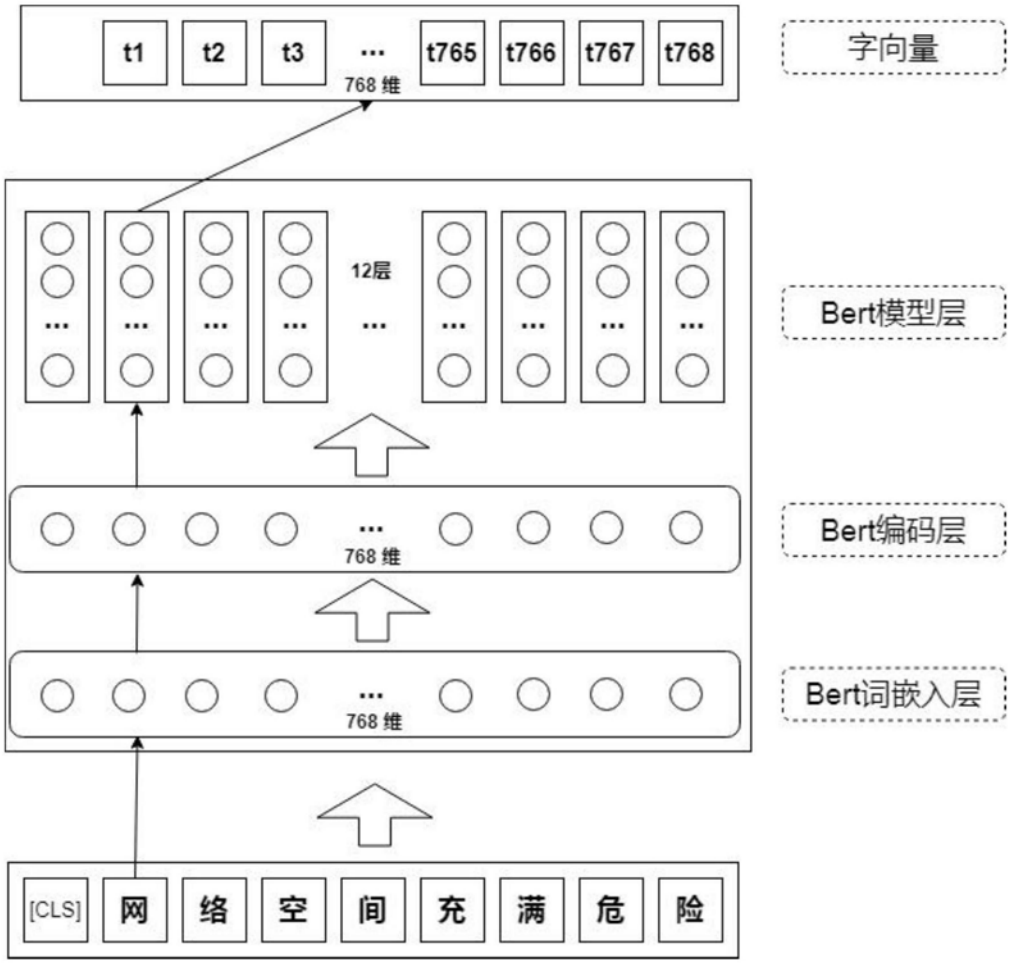

t1
t2
t3
…
768 维
t765
t766
t767
t768
字向量
12层
Bert模型层
Bert编码层
Bert词嵌入层
[CLS]
网
络
空
间
充
满
危
险

图3

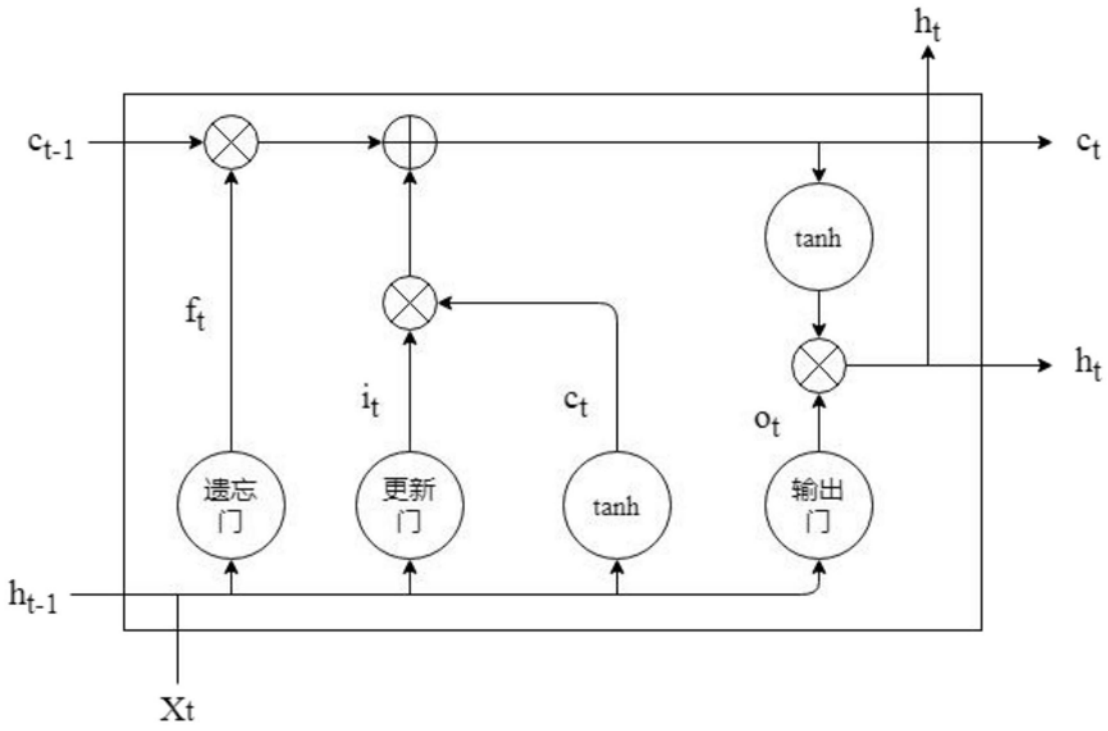

h_t
c_{t-1}
c_t
tanh
f_t
i_t
c_t
o_t
h_t
遗忘门
更新门
tanh
输出门
h_{t-1}
Xt

图4

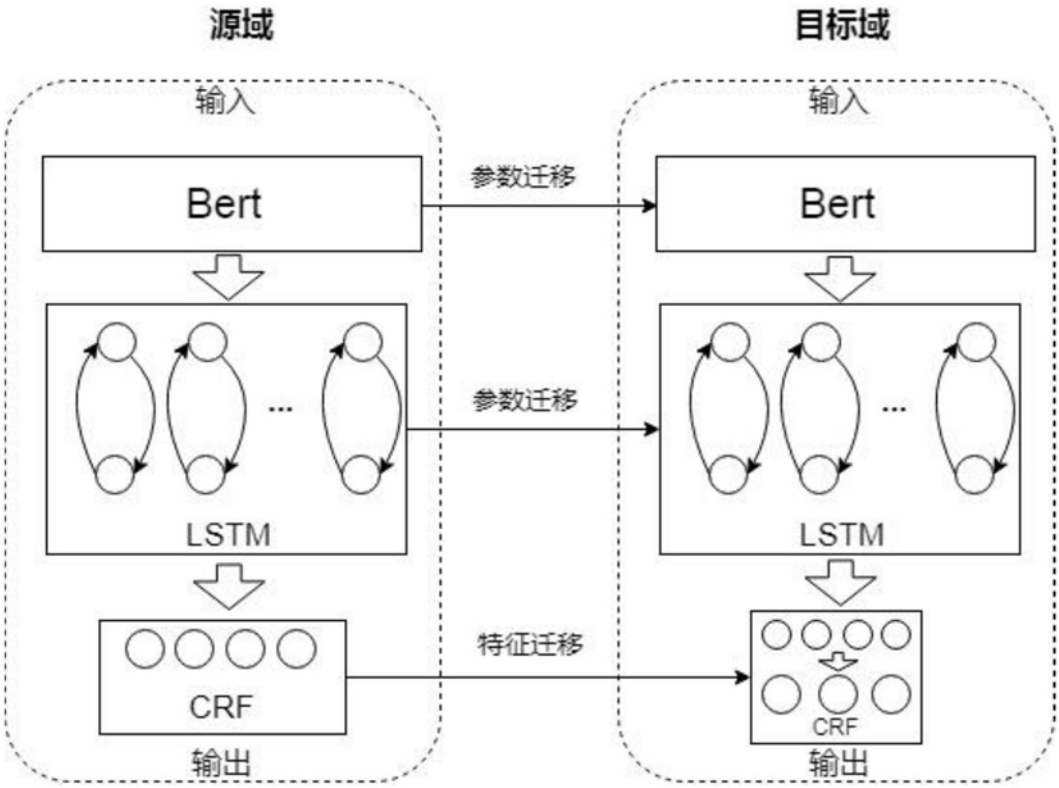

源域
目标域
输入
输入
Bert
参数迁移
Bert
...
参数迁移
...
LSTM
LSTM
CRF
特征迁移
CRF
输出
输出

图5

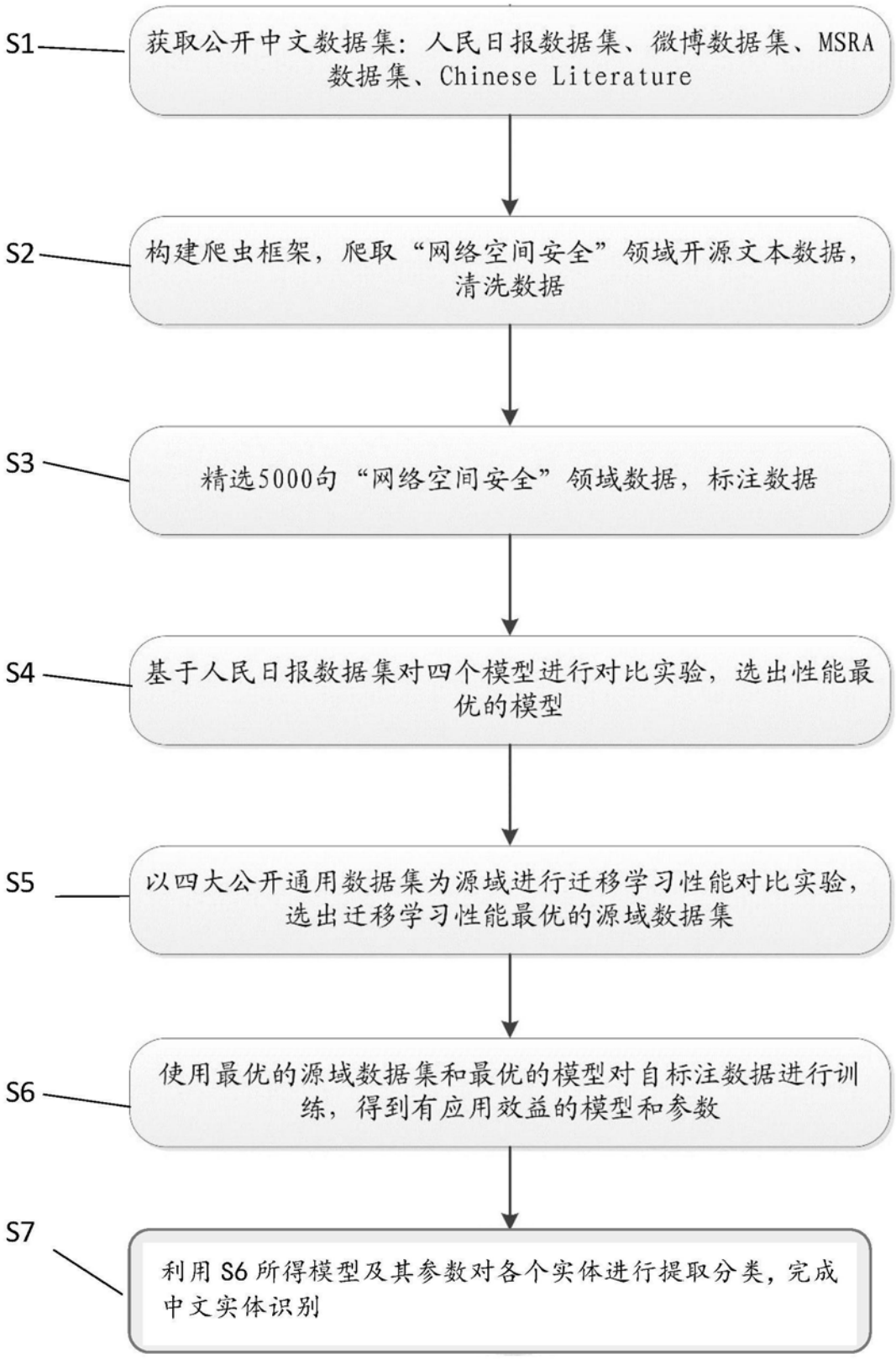
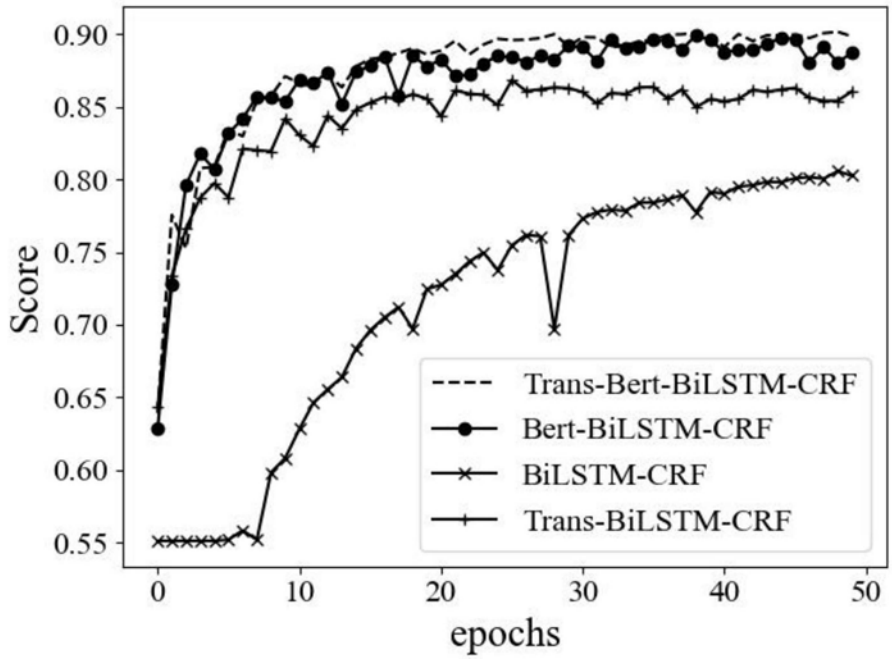
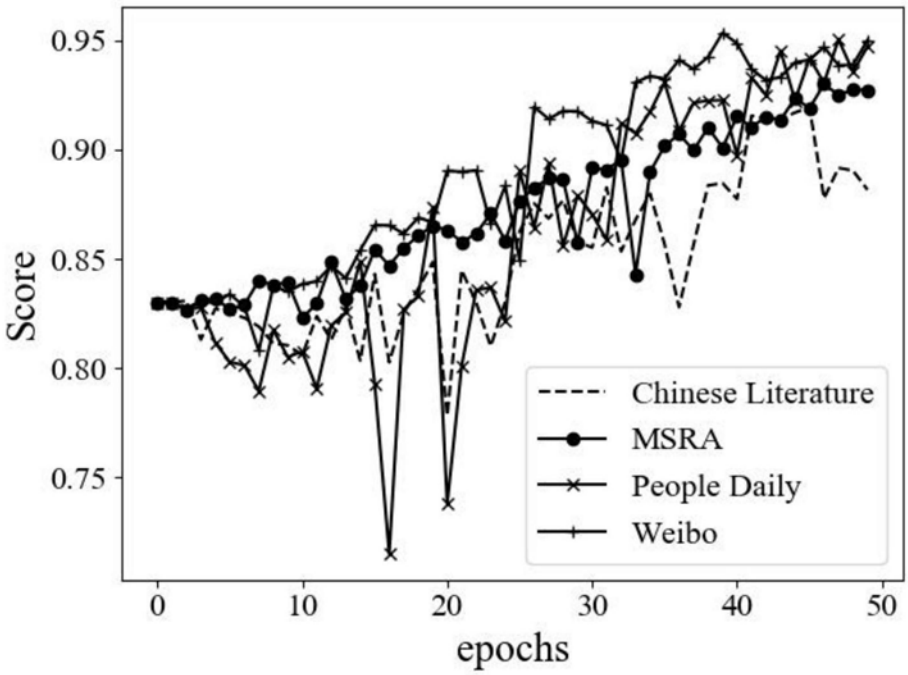

S1
获取公开中文数据集：人民日报数据集、微博数据集、MSRA数据集、Chinese Literature
S2
构建爬虫框架，爬取“网络空间安全”领域开源文本数据，清洗数据
S3
精选5000句“网络空间安全”领域数据，标注数据
S4
基于人民日报数据集对四个模型进行对比实验，选出性能最优的模型
S5
以四大公开通用数据集为源域进行迁移学习性能对比实验，选出迁移学习性能最优的源域数据集
S6
使用最优的源域数据集和最优的模型对自标注数据进行训练，得到有应用效益的模型和参数
S7
利用S6所得模型及其参数对各个实体进行提取分类，完成中文实体识别

图6


0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
Score
0
10
20
30
40
50
epochs
Trans-Bert-BiLSTM-CRF
Bert-BiLSTM-CRF
BiLSTM-CRF
Trans-BiLSTM-CRF

（a）


0.95
0.90
0.85
0.80
0.75
Score
0
10
20
30
40
50
epochs
Chinese Literature
MSRA
People Daily
Weibo

（b）

图7