(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111444721 B
(45)授权公告日 2022.09.23

(21)申请号 202010460134.7
(22)申请日 2020.05.27
(65)同一申请的已公布的文献号
申请公布号 CN 111444721 A
(43)申请公布日 2020.07.24
(73)专利权人 南京大学
地址 210046 江苏省南京市栖霞区仙林大道163号
(72)发明人 俞扬 詹德川 周志华 李龙宇
(74)专利代理机构 南京乐羽知行专利代理事务所(普通合伙) 32326
专利代理师 李玉平
(51)Int.Cl.
G06F 40/295(2020.01)
G06F 40/30(2020.01)
G06F 40/211(2020.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)
(56)对比文件
CN 109858018 A,2019.06.07
CN 111160026 A,2020.05.15
CN 110427627 A,2019.11.08
CN 110543639 A,2019.12.06
CN 110674639 A,2020.01.10
CN 111160026 A,2020.05.15
CN 109918644 A,2019.06.21
CN 105138575 A,2015.12.09
CN 110348008 A,2019.10.18
CN 107862039 A,2018.03.30
US 2020012657 A1,2020.01.09
CN 110083831 A,2019.08.02
Frank Cao.BERT:Bidirectional Encoder Representations from Transformers (基于转换器的双向编码表征).《知乎》.2018,
光彩照人.BERT(Bidirectional Encoder Representations from Transformers)理解.《博客园》.2018,
Jacob Devlin 等.BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.《arXiv》.2019,
NLP学习笔记.彻底理解 Google BERT 模型.《百度》.2019,
审查员 刘佳

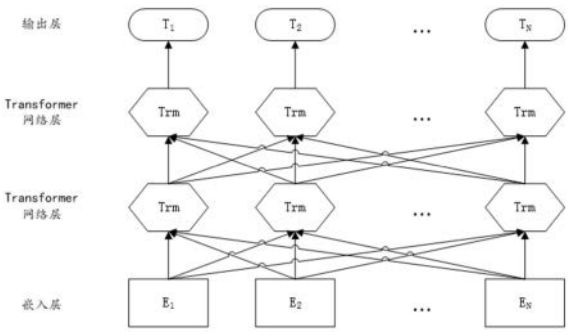
权利要求书3页 说明书5页 附图2页

(54)发明名称

一种基于预训练语言模型的中文文本关键信息抽取方法

(57)摘要

本发明公开了一种基于预训练语言模型的中文文本关键信息抽取方法,具体步骤如下:(1)将待抽取的关键信息进行分类,易于归纳组成规则的信息类别,使用正则匹配的方法抽取。(2)对命名实体使用序列标注模型抽取。(3)序列标注模型采用对预训练语言模型微调的方法进行构建,首先使用大规模无标记文本语料学习得到预训练语言模型,并在预训练阶段引入词边界特征。(4)将使用规则匹配的数据内容替换为其对应的规则模板标签,以完成规则匹配与深度网络的融合。(5)根据有标记的训练数据,在预训练语言模型上进行微调,将其迁移到命名实体的序列标注任务上。本发明可以有效提取文本上下文语义特征,并且在复杂信息类别的场景下有效地识别各个信息种类。



CN 111444721 B

1.一种基于预训练语言模型的中文文本关键信息抽取方法,其特征在于,包括如下步骤:

步骤(1),对基于规则匹配方法进行识别的信息类别,编写相应的规则模板,并为每一个类别设置对应的标签名;

步骤(2),基于任务文本环境,收集无标记的文本语料;

步骤(3),对步骤(2)中收集的无标记的文本语料使用规则模板进行抽取,将数字串和字符串使用规则模板抽取出来,之后将文本语料中匹配的数字串、字符串在原文中的位置替换为其对应的类别标签;

步骤(4),基于步骤(3)处理后的无标记文本语料,基于Transformer网络结构构建预训练语言模型,使用遮掩语言模型任务在收集到的文本语料上进行预训练;并在预训练语言模型网络的输入阶段,通过将文本分词的嵌入表示结合到输入中,在预训练语言模型中引入分词特征;

步骤(5),基于任务文本环境收集文本语料数据集,构建命名实体识别数据集,采用BIO标注格式对该文本语料数据集中的命名实体类别进行标注,得到命名实体识别数据集;

步骤(6),使用规则模板匹配,对步骤(5)中带标记的命名实体识别数据集使用规则模板匹配数字串、字符串,并将匹配的数字串在原文中的位置替换为其对应的类别标签;

步骤(7),针对步骤(4)中得到的预训练语言模型,使用步骤(5)标注的命名实体识别数据集对其进行微调;

步骤(8),使用微调后的预训练语言模型对待预测文本数据进行识别抽取;

将训练数据中的每条训练语句,通过字符表将语句转化为对应字符编号的序列,并使用随机初始化的字嵌入对语句中的每个字符进行表示,对每个字符使用嵌入向量进行表示;同时,还对训练数据中的每个语句添加位置嵌入,对语句中的每个字符计算位置嵌入;并且,针对训练数据中的每条中文语句进行分词,对文本中的每个字符构造分词嵌入;最终,将这三种嵌入相加,相加后作为预训练语言模型的输入;中文分词共有4种特征:BIES,分别表示词的起始字符B;词的中间字符I;词的结尾字符E;和独字词S;使用Transformer来训练得到输入语句的语义特征;

遮掩语言模型为:随机遮掩住句子中的一部分字,然后通过该部分字的上下文表征进行预测被遮掩位置上的字;预训练文本语料中的每条文本数据,会有15%的字会被随机选中;在被选中的字中,有80%会被遮掩,即将需遮掩字替换为一个特殊标记[MASK];有10%会被随机替换为一个任意字符;剩余10%不进行任何操作;完成文本语料的遮掩之后,得到处理完成的预训练语言模型的训练数据;

对于处理好的预训练语言模型的训练数据,首先基于训练数据中的词频建立字符表,便于对训练数据进行处理,并按字符表的顺序对字符表里的字符编号;同时,字符表中也包含规则匹配类别的标签。

2.根据权利要求1所述的基于预训练语言模型的中文文本关键信息抽取方法,其特征在于,所述步骤(7)中的微调为:在已有的预训练语言模型参数上添加参数,然后基于命名实体识别数据集,使用小学习率对所有的参数进行训练,从而将预训练语言模型迁移到命名实体识别任务上去。

3.根据权利要求1所述的基于预训练语言模型的中文文本关键信息抽取方法,其特征

在于,所述步骤(5)中收集的数据集,数据集规模为几千至几万条,并对其中的命名实体进行标注,该数据集是用作命名实体识别的训练数据。

4.根据权利要求1所述的基于预训练语言模型的中文文本关键信息抽取方法,其特征在于,所述基于Transformer网络结构构建预训练语言模型,即构建基于Transformer网络结构的双向深度语言模型,整个网络由12个连续相同的网络层组成,每一层中有2个子层,分别是:多头注意力层和前馈网络层,这两个子层之间都有残差连接与层归一化操作;

多头注意力是一种注意力机制,其计算形式如下:

$MultiHead(Q,K,V)=Concat(head_1,...head_h)W^O$

其中,$head_i=Attention(QW_i^Q,KW_i^K,VW_i^V)$

其中,W^Q,W^K,W^V是参数映射矩阵,h是注意力头数,将注意力分为h个注意力头,能够分别抽取不同子区域的特征;W^O同样是参数映射矩阵,Concat函数在各个注意力头完成注意力计算后,将所有注意力头拼接到一起;

对于Attention的计算如下式:

$$Attention(Q,K,V)=softmax(\frac{QK^T}{\sqrt{d_k}})V$$

其中,Q,K,V均是输入字向量矩阵,d_k是输入向量的维度;通过上式注意力机制的计算,即可得到Q关于在V上的注意力机制,即应该重点关注的V中的区域。

5.根据权利要求1所述的基于预训练语言模型的中文文本关键信息抽取方法,其特征在于,

位置嵌入是基于正弦函数计算得出:

$$PE_{(pos,2i)}=\sin(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)}=cos(pos/10000^{2i/d_{model}})$$

其中,pos是输入的位置,i则表示输入向量的维度,d_{model}是模型输入的维度。

6.根据权利要求5所述的基于预训练语言模型的中文文本关键信息抽取方法,其特征在于,完成预训练语言模型之后,即可开始迁移到命名实体识别任务上:

(3.1)收集标注命名实体标记数据集,对于命名实体类型,其中人名标记为PER、地名标记为LOC、机构名标记为ORG;对于实体边界,采用BIO标注格式,其中B表示实体的起始字符,I表示实体的非起始字符,O表示在实体外,意为非命名实体;两类标记进行组合;

(3.2)使用规则模板,将标记数据中规则匹配到的内容替换为其对应的类别标签;

(3.3)在预训练语言模型的网络最上层再添加一层全连接网络,并使用小学习率在标记数据上对预训练语言模型进行微调;在已有的预训练语言模型参数上添加一层全连接网络,然后基于命名实体的标记数据,对所有的参数进行小学习率的训练,从而将模型迁移到命名实体识别任务上去;

(3.4)由全连接网络产生输入文本中每个字符对应的命名实体类别标签,输出每个实体的类别标记;

至此完成了命名实体识别任务的训练,保存微调后的模型,即可对待预测文本数据进行识别抽取:

(4.1)执行预测时同样需要先使用规则模板进行匹配,对命中规则的内容使用对应的

类别标签进行替换,并保存规则匹配的结果;

(4.2)之后携同上下文文本,一起投入微调后的预训练语言模型中去,即可得出文本各个字符对应的BIO标记;将BIO标记进行格式转换,即可得到最终的命名实体名。

一种基于预训练语言模型的中文文本关键信息抽取方法

技术领域

[0001] 本发明涉及一种基于预训练语言模型的中文文本关键信息抽取方法,属于自然语言处理识别技术领域。

背景技术

[0002] 文本关键信息抽取指根据具体业务的需求,对文本中指定的关键数据类型进行识别和抽取。主要包括对命名实体(Named Entity)的识别和一些特定类型的数字串、字符串的识别。命名实体的识别问题可以使用基于深度学习的序列标注模型较好的解决,但无法同时解决其它数字串、字符串的识别需求。因为,数字串无法携带有效的语义信息,且多种数字串之间会相互干扰。

[0003] 现有效果较为出色的中文命名实体识别方法大都基于循环神经网络(Recurrent Neural Network,RNN)的字标注模型,这种方法首先需要利用自然语言语料无监督地训练得到汉字的分布式表示,得到每个汉字在语义空间中的向量表示。然后,使用循环神经网络(RNN)或其变种长短时记忆网络(LSTM、GRU)等,对文本序列进行建模,抽取字与字、词与词之间的语义及语法特征表示。最后,对循环神经网络得到的特征提取结果,使用条件随机场(CRF)对序列中隐状态之间的转移规则做进一步约束,强化特征转移。训练得到一个基于字符的命名实体识别深度学习模型。

[0004] 但是,基于传统的word2vec方法学得的词表示,将每个单词的上下文信息限制在一个较小的固定大小的窗口内,无法学得全文单词之间的长距离依赖关系,只能将语义关系建模在一个小范围的上下文内。并且,传统的词嵌入(word embedding)方法,将每个词保存为一条静态向量,即每个词只能占用一个语义空间中的位置,此时不同上下文的信息都会被编码到同一个参数空间中,导致传统词嵌入无法解决多义词问题。因为多义词的现象在实际应用场景中十分常见,必须动态地根据上下文环境的变化,给出不同的语义表示。例如:一名叫做武汉市的男子给儿子取名为武昌。显然在这里的语境中,“武汉市”与“武昌”不再是地名实体,而变为了人名实体。

[0005] 而且,基于字符的中文命名实体识别模型,完全摒弃了中文词边界特征,命名实体的边界一般也是词边界。完全基于字符的模型,丧失了中文词边界内蕴含的语义信息。不利于中文文本的语义表示,进而影响命名实体识别的准确率。

发明内容

[0006] 发明目的:针对传统方法中无法解决一词多义以及词边界信息缺失等问题,本发明提出了一种基于预训练语言模型的关键信息抽取方法。基于现有序列标注方法进行的改进优化,更好地获得对中文文本的语义表示,以此强化深度学习模型的表达能力,进而更好地为中文命名实体识别任务服务。本发明深度融合规则匹配与深度模型,可以有效提取文本上下文语义特征,并且在复杂信息类别的场景下有效地识别各个信息种类,取得很好的识别效果。在内部数据集上的F1值超过传统基于BiLSTM-CRF的命名实体识别方法2个多百

分点。

[0007] 技术方案:一种基于预训练语言模型的中文文本关键信息抽取方法,包括如下步骤:

[0008] 步骤(1):将待抽取的中文文本关键信息分类进行识别,对命名实体类别使用深度学习模型进行识别;对可以归纳组成规则的信息类别(如数字串和字符串),使用正则匹配的方法识别。对基于规则匹配方法进行识别的信息类别,归纳出其内部组成结构,编写相应的规则模板,并为每一个类别设置对应的标签名;

[0009] 步骤(2):基于任务文本环境,收集大规模无标记的文本语料;

[0010] 步骤(3):对步骤(2)中收集的无标记的文本语料使用规则模板进行抽取,将数字串和字符串等内容先使用规则模板抽取出来,之后并将文本语料中匹配的数字串、字符串在原文中的位置替换为其对应的类别标签;

[0011] 步骤(4):基于步骤(3)处理后的无标记文本语料,基于Transformer网络结构构建预训练语言模型,使用遮掩语言模型任务在收集到的文本语料上进行预训练。并在预训练语言模型网络的输入阶段,通过将文本分词的嵌入表示结合到输入中,在预训练语言模型中引入分词特征;

[0012] 步骤(5):基于任务文本环境收集文本语料数据集,构建命名实体识别数据集,采用BIO标注格式对该文本语料数据集中的命名实体类别进行标注,得到命名实体识别数据集;

[0013] 步骤(6):类似于步骤(3)使用规则模板匹配,对步骤(5)中带标记的命名实体识别数据集使用规则模板匹配数字串、字符串,并将匹配的数字串在原文中的位置替换为其对应的类别标签;

[0014] 步骤(7):针对步骤(4)中得到的预训练语言模型,使用步骤(5)标注的命名实体识别数据集对其进行微调。微调即意为:在已有的预训练语言模型参数上添加参数(如:添加一层全连接网络),然后基于命名实体识别数据集,使用小学习率对所有的参数(包括预训练语言模型参数)进行训练,从而将预训练语言模型迁移到命名实体识别任务上去;

[0015] 进一步的,所述步骤(2)和步骤(5)中,都收集相关数据集,但功用完全不同。步骤(2)中收集的数据集规模较大,可以达到百万甚至千万条数据的规模,但这些数据无需标记,主要用于语言模型能够从大规模语言文本中抽取到深层次的文本语义特征,由此构建预训练语言模型;而步骤(5)中收集的数据集,无需太大规模几千至几万条即可,并需要对其中的命名实体进行标注,该数据集是用作命名实体识别的训练数据,预训练语言模型在该数据集上进行微调,即可迁移到命名实体识别任务上去。

[0016] 有益效果:与现有技术相比,本发明提供的基于预训练语言模型的中文文本关键信息抽取方法,具有如下优点:

[0017] (1)规则匹配与深度网络两个模块进行深度融合,可以将规则模板的特征传递给预训练语言模型,使之产生语义更加丰富的上下文表示,辅助对于命名实体更好地识别。

[0018] (2)深度双向预训练语言模型,能够无监督地从文本语料中抽取出深层语义特征。且根据中文命名实体识别任务,对预训练过程加以改进,引入词边界特征作为输入的一个维度,丰富了上下文语义特征。

[0019] (3)使用基于Transformer的网络结构进行遮掩语言模型的训练,可以对全文进行

注意力操作，不再仅仅依赖于固定大小窗口的上下文信息，而可以学习到文本关于全局上下文的表征。预训练语言模型得到的是基于上下文的动态词表征，即对于同一个词的不同上下文环境，模型会给出其在参数空间中的不同表示，很好地解决了一词多义的问题。

[0020] 不同于循环神经网络，基于Transformer的网络结构完全基于张量的前向操作，可以完美地契合GPU的并行计算。

附图说明

[0021] 图1是Transformer的网络结构图；

[0022] 图2是本发明采用的预训练语言模型结构示意图；

[0023] 图3是基于预训练语言模型的文本关键信息抽取关键步骤工作流程图。

具体实施方式

[0024] 下面结合具体实施例，进一步阐明本发明，应理解这些实施例仅用于说明本发明而不用于限制本发明的范围，在阅读了本发明之后，本领域技术人员对本发明的各种等价形式的修改均落于本申请所附权利要求所限定的范围。

[0025] 本发明主要针对复杂场景下的文本关键信息抽取，呈现了一种基于预训练语言模型的方法。该方法将待抽取的信息类别分为两个模块：一是使用规则匹配模块；二是基于深度学习模型的命名实体识别模块。该方法能够深度融合正则匹配特征与深度语言模型语义特征，由此带来识别准确性的提升。如图3所示，是整个基于预训练语言模型的关键步骤工作流程图，其中预训练语言模型的结构如图2所示，预训练语言模型所采用的特征抽取网络Transformer的网络结构如图1所示。预训练语言模型完成后，即可对带标记的序列标注训练数据进行处理。先使用规则去匹配待抽取的文本，然后将其类别标签作为特征引入到预训练语言模型中，并在预训练语言模型上引入中文分词特征，最后，通过序列标注任务对预训练语言模型进行微调，将模型迁移到序列标注任务上。

[0026] 本发明的基于预训练语言模型的文本关键信息抽取方法，其具体步骤如下：

[0027] (1)针对基于规则匹配的类别，归纳其内部组成规则，编写相对应的正则表达式。并给每个信息类别一个特殊的标签，例如：给邮箱标记为<EMAIL>等。

[0028] (2)构建预训练语言模型。

[0029] (2.1)基于任务文本环境收集数百万条无标记的文本语料，用作预训练语言模型的训练数据。

[0030] (2.2)针对(2.1)收集的无标记的文本语料，使用(1)中编写的正则表达式去匹配该语料，对匹配正则表达式的文本部分使用其对应的类别标签替换(例如：使用标签<EMAIL>替换文本中匹配到的邮箱内容)。

[0031] (2.3)基于遮掩语言模型任务进行预训练，需要对(2.2)处理后的文本语料做遮掩处理。所谓遮掩语言模型即为：随机遮掩住句子中的一部分字，然后通过该部分字的上下文表征进行预测被遮掩位置上的字。预训练文本语料中的每条文本数据，会有15％的字会被随机选中。在被选中的字中，有80％会被遮掩，即将需遮掩字替换为一个特殊标记[MASK]；有10％会被随机替换为一个任意字符；剩余10％不进行任何操作。完成文本语料的遮掩之后，得到处理完成的预训练语言模型的训练数据。

[0032] (2.4)对于(2.3)中处理好的预训练语言模型的训练数据,首先基于训练数据中的词频建立字符表,便于对训练数据进行处理,并按字符表的顺序对字符表里的字符编号。同时,字符表中也包含规则匹配类别的标签。

[0033] (2.5)构建基于Transformer网络结构(如图1)的双向深度语言模型,整个网络由12个连续相同的网络层组成,每一层中有2个子层,分别是:多头注意力层和前馈网络层,这两个子层之间都有残差连接与层归一化操作。模型结构如图2所示。

[0034] 多头注意力是一种注意力机制,其计算形式如下:

[0035] $\mathrm{MultiHead}(Q,K,V)=\mathrm{Concat}(\mathrm{head}_1,\cdots\mathrm{head}_h)W^O$

[0036] 其中,$\mathrm{head}_i=\mathrm{Attention}(QW_i^Q,KW_i^K,VW_i^V)$

[0037] 其中,W^Q,W^K,W^V是参数映射矩阵,h是注意力头数(h取值为8),将注意力分为h个注意力头,能够分别抽取不同子区域的特征;W^O同样是参数映射矩阵,Concat函数在各个注意力头完成注意力计算后,将所有注意力头拼接到一起。

[0038] 对于Attention的计算如下式:

[0039] $$Attention(Q,K,V)=softmax(\frac{QK^T}{\sqrt{d_k}})V$$

[0040] 其中,Q,K,V均是输入字向量矩阵,d_k是输入向量的维度。通过上式注意力机制的计算,即可得到Q关于在V上的注意力机制,即应该重点关注的V中的区域。

[0041] (2.6)将(2.3)处理得到的训练数据中的每条训练语句,通过字符表将语句转化为对应字符编号的序列,并使用随机初始化的字嵌入对语句中的每个字符进行表示,对每个字符使用768维的嵌入向量进行表示;同时,还对(2.3)处理得到的训练数据中的每个语句添加位置嵌入,对语句中的每个字符计算位置嵌入;并且,针对(2.3)处理得到的训练数据中的每条中文语句进行分词,对文本中的每个字符构造分词嵌入。最终,将这三种嵌入相加,相加后作为预训练语言模型的输入。中文分词共有4种特征:BIES,分别表示词的起始字符B;词的中间字符I;词的结尾字符E;和独字词S。如图2所示,使用Transformer来训练得到输入语句的语义特征。

[0042] 其中,位置嵌入是基于正弦函数计算得出:

[0043] $$PE_{(pos,2i)}=\sin(pos/10000^{2i/d_{model}})$$

[0044] $$PE_{(pos,2i+1)}=cos(pos/10000^{2i/d_{model}})$$

[0045] 其中,pos是输入的位置,i则表示输入向量的维度,d_{model}是模型输入的维度。

[0046] (2.7)预训练完成后,保存预训练语言模型参数,以待后续微调。

[0047] (3)完成预训练语言模型之后,即可开始迁移到命名实体识别任务上,首先要获取带标记的命名实体数据集。

[0048] (3.1)收集标注命名实体标记数据集,对于命名实体类型,其中人名标记为PER、地名标记为LOC、机构名标记为ORG。对于实体边界,采用BIO标注格式,其中B表示实体的起始字符,I表示实体的非起始字符,O表示在实体外,意为非命名实体。两类标记进行组合,例如,句子“张三想去北京工作”:

[0049]

张	三	想	去	北	京	工	作
B-PER	I-PER	O	O	B-LOC	I-LOC	O	O

[0050] (3.2)使用(1)中编写的规则,将标记数据中规则(正则表达式)匹配到的内容替换为其对应的类别标签。

[0051] (3.3)在预训练语言模型的网络最上层再添加一层全连接网络,并使用小学习率在标记数据上对预训练语言模型进行微调;在已有的预训练语言模型参数上添加少量参数(添加一层全连接网络),然后基于命名实体的标记数据,对所有的参数(包括预训练语言模型参数)进行小学习率的训练,从而将模型迁移到命名实体识别任务上去。

[0052] (3.4)由全连接网络产生输入文本中每个字符对应的命名实体类别标签,输出每个实体的类别标记。

[0053] (4)至此完成了命名实体识别任务的训练,保存微调后的模型,即可对待预测文本数据进行识别抽取。

[0054] (4.1)执行预测时同样需要先使用规则模板(正则表达式)进行匹配,对命中规则(匹配正则表达式)的内容使用对应的类别标签进行替换,并保存规则匹配的结果。

[0055] (4.2)之后携同上下文文本,一起投入微调后的预训练语言模型中去,即可得出文本各个字符对应的BIO标记。将BIO标记进行格式转换,即可得到最终的命名实体名。

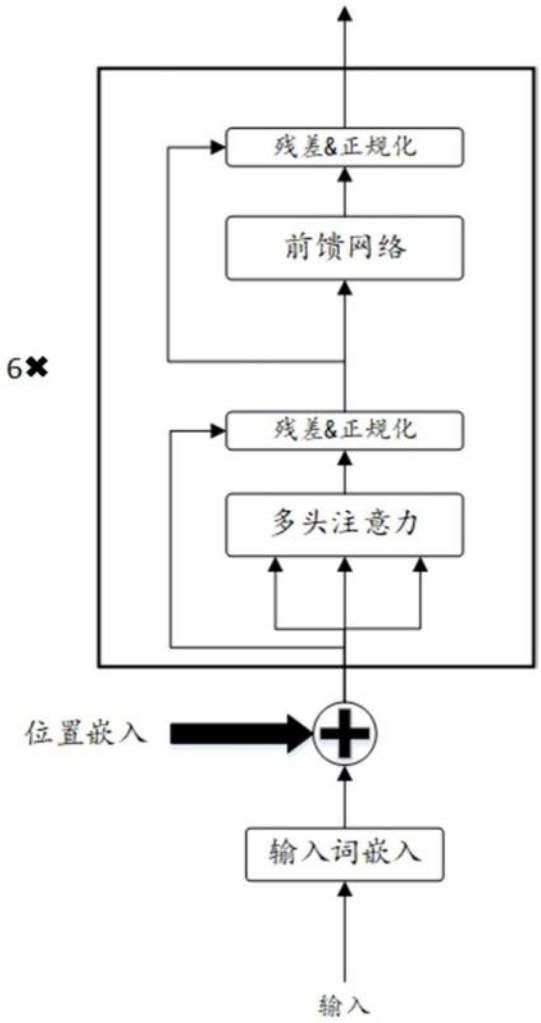


图1

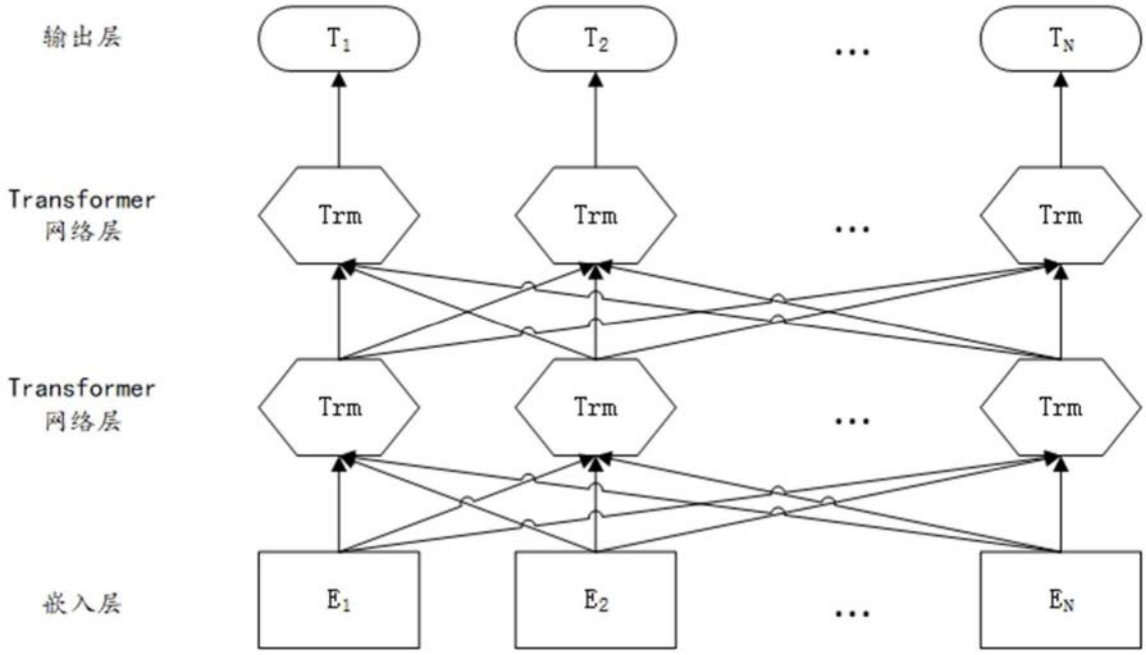

输出层
T1
T2
...
TN
Transformer
网络层
Trm
Trm
...
Trm
Transformer
网络层
Trm
Trm
...
Trm
嵌入层
E1
E2
...
EN

图2

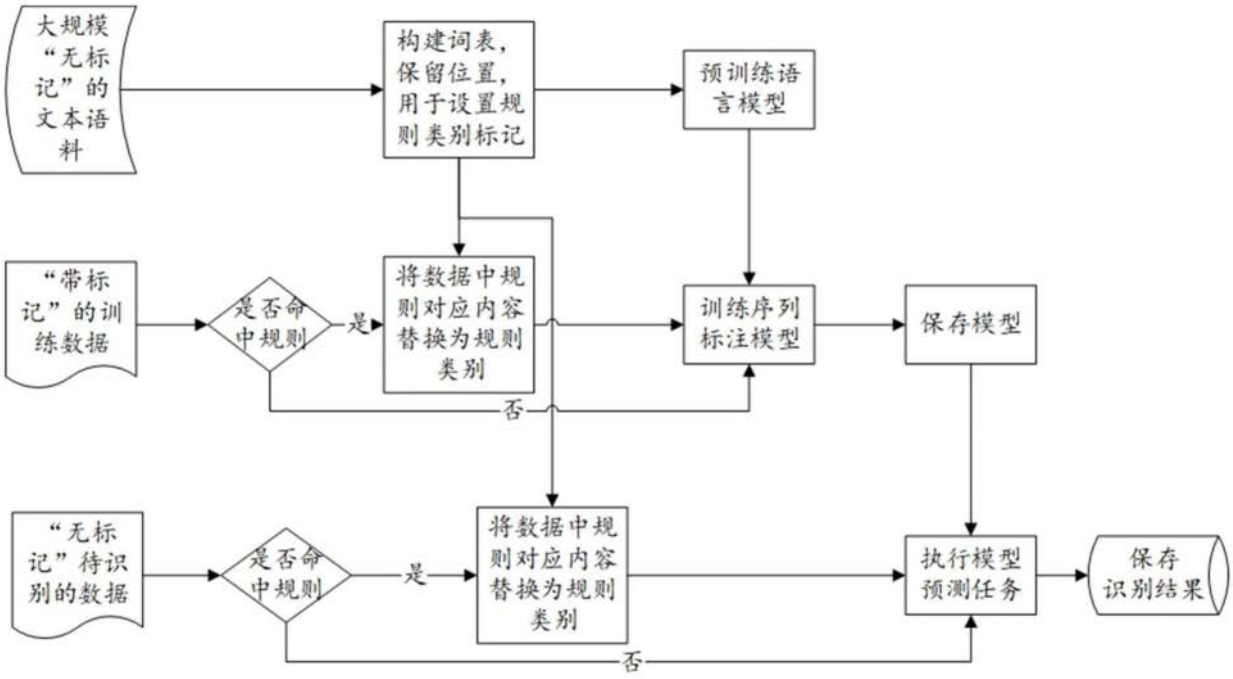

大规模“无标记”的文本语料
构建词表，保留位置，用于设置规则类别标记
预训练语言模型
“带标记”的训练数据
是否命中规则
是
将数据中规则对应内容替换为规则类别
训练序列标注模型
保存模型
否
“无标记”待识别的数据
是否命中规则
是
将数据中规则对应内容替换为规则类别
执行模型预测任务
保存识别结果
否

图3