(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116127097 A
(43)申请公布日 2023.05.16

(21)申请号 202310136023.4
(22)申请日 2023.02.20
(71)申请人 广东工业大学
地址 510000 广东省广州市东风东路729号
申请人 广州安特激光技术有限公司
(72)发明人 杨祖元 黄永清 李珍妮 谢胜利
(74)专利代理机构 深圳市广诺专利代理事务所(普通合伙) 44611
专利代理师 王允亮
(51)Int.Cl.
G06F 16/36(2019.01)
G06F 16/35(2019.01)
G06F 40/211(2020.01)
G06F 40/30(2020.01)
G06F 40/279(2020.01)
G06N 3/084(2023.01)

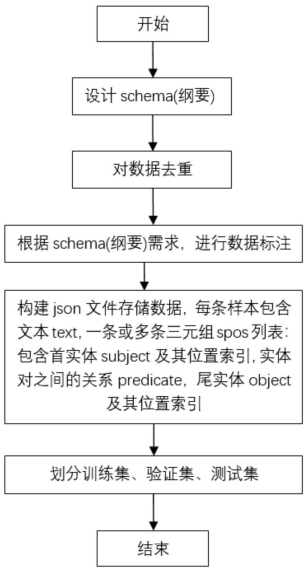
权利要求书2页 说明书7页 附图2页

(54)发明名称

一种结构化文本关系抽取方法、装置、设备

(57)摘要

本申请公开了一种结构化文本关系抽取方法、装置、设备和储存介质，涉及人工智能和自然语言处理技术领域；具体为：步骤1：设计关系抽取数据的schema(纲要)，为了规范结构化数据的表达，关系抽取的每条数据必须满足纲要(schema)预先定义的实体对象及其类型，对数据去重和标注，构建模型的训练集、验证集和测试集；步骤2：构建基于深度学习的关系抽取模型；步骤3：利用训练集数据训练深度学习模型，保存在验证集上效果最好的模型权重；步骤4、利用保存的模型对待测数据抽取关系三元组；步骤5、将抽取的实体关系三元组进行结构化存储。该技术可以从结构化文本信息中抽取知识三元组，从数据中提取高层抽象特征，为知识图谱的构建提供技术支持。



CN 116127097 A

1.一种结构化文本关系抽取方法,其特征在于,包括以下步骤:

S1:数据构建和预处理;

S2:搭建深度学习关系抽取模型;

S3:经过Bert-base模型编码的句向量分别经过实体识别层和关判别层,并获取模型的损失值;

S4:通过反向传播和梯度下降,对模型参数进行更新

S5:从测试集中选取数据,通过训练好的模型进行三元组关系抽取,挖掘数据;

S6:将得到的数据关系三元组存入数据库Mysql中。

2.根据权利要求1所述的一种结构化文本关系抽取方法,其特征在于:

所述对数据集构建和预处理,获得实体以及实体间关系,具体包括以下步骤:

S2.1:针对关系抽取数据,设计好纲要(schema),用于定义需要存储的关系数据具体信息:subject(首实体)类型、predicate(关系)和object(尾实体)类型,三者之间具有相对应的关系,规范结构化数据的表达,关系抽取的每条数据必须满足纲要(schema)预先定义的实体对象及其类型;

S2.2:通过代码定义数据预处理类,对数据进行去重、构建文本数据集。将数据集存储为json文件,样本形式以键-值对存在;

S2.3:不采用传统的CRF做实体识别时对数据的处理方式:即不采用“BIESO”标注实体方式,只需要知道实体对在文本中的位置信息。构建实体类别标签映射ID的字典;

S2.4:统计数据中出现的实体间关系类别,构建实体间关系类别映射ID的字典;

S2.5:数据经过预处理后会划分为训练集、验证集和测试集,分别用于深度学习模型的训练、验证用于保存最优训练模型和对模型进行测试。

3.根据权利要求1所述的一种结构化文本关系抽取方法,其特征在于:

所述使用预训练语言模型如Bert-base构建深度学习关系抽取模型,具体包括以下步骤:

S3.1:采用Bert-base作为模型的编码器。Bert-base具有12层Encoder层,每一层学习文本的不同语义方面的知识,为了充分利用上下文信息以及不同方面的语义知识,结合bert模型的最后四层,对其进行加权平均,用作整条文本的句向量;

S3.2:搭建两个实体识别模块,分别用于识别首实体subject和尾实体object;

S3.3:搭建两个关系匹配模型,分别根据首实体、尾实体的区间位置信息中起止位置信息、结束位置信息进行实体对关系之间的匹配;

S3.4:在关系抽取主任务上,通过添加注意力机制(attention),新增一个下游任务层形成辅助任务,用于后处理抽取的关系三元组个数,构成多任务学习,为模型增加鲁棒性;

S3.5:计算损失函数值,通过反向传播和梯度下降,对模型参数进行更新。

4.根据权利要求3所述的一种结构化文本关系抽取方法,其特征在于:

所述S3.1具体包括以下步骤:

S3.1.1:将训练数据中的每条样本按照字进行划分,如果按词切分可能会导致数据中的实体不在字典中,即OOV(英文全称:out of vocabulary);

S3.2.2:如果当前句子为x,则划分后得到序列表示$x=[x_0,x_1,...,x_{n-1},x_n]$,根据Bert预训练语言模型的收入要求,令$x_0=$[CLS],$x_n=$[SEP],其中[CLS]标志位于句子的首位,而

[SEP]标志位于句子的末尾；

S3.2.3:将得到的文本序列x经过Bert模型，并取最后四层的加权平均h＝concatenate([$layer_9$,$layer_{10}$,$layer_{11}$,$layer_{12}$])，其中$layer_i$表示第i层输出的向量，可得到结合上下文语义信息的word embedding(词嵌入)。

5.根据权利要求3所述的一种结构化文本关系抽取方法，其特征在于：

所述S3.5具体包括以下步骤：

S3.5.1:在S3.2.3中得到包含上下文语义的句向量[h_1,h_2,...,h_n]，通过变换$q_{i,\alpha}=W_{q,\alpha}h_i+b_{q,\alpha}$和$k_{i,\alpha}=W_{k,\alpha}h_i+b_{k,\alpha}$得到向量序列[$q_{1,\alpha}$,$q_{2,\alpha}$,...,$q_{n,\alpha}$]和[$k_{1,\alpha}$,$k_{2,\alpha}$,...,$k_{n,\alpha}$]；

S3.5.2:利用这两个向量序列可以构建一个实体识别的打分函数$f_\alpha(i,j)=q_{i,\alpha}^T k_{j,\alpha}$，表示$q_{i,\alpha}$与$k_{i,\alpha}$的内积。其中[i:j]是文本text中的一段连续子串，可构成一个实体；

S3.5.3:模型的首实体和尾实体层(S3.2所述)通过打分函数可得到两个向量e1和e2，向量维度是[n,seq_len,seq_len]；

S3.5.4:在关系匹配层中，一样可以利用上述的打分函数，在实体识别中是根据实体的位置信息进行打分，而在关系匹配层中，模型是根据首尾实体的开始和结束位置信息进行打分；

S3.5.5:损失函数采用多标签分类的损失函数$loss_1=\log(1+\sum_{(i,j)\in P\alpha}e^{-s_\alpha(i,j)})+\log(1+\sum_{(i,j)\in Q\alpha}e^{-s_\alpha(i,j)})$，其中$P_\alpha$是该样本所以类型为α的实体的首尾集合，$Q_\alpha$则是非实体或者非α类型的实体的首尾集合；

S3.5.6:针对辅助任务的损失函数，可采用常见的交叉熵损失函数计算损失值$loss_2=-\frac{1}{N}\sum_{i=1}^{N}[y_i\ln a+(1-y_i)\ln(1-a)]$。最后模型损失值为两者和loss＝$loss_1$+$loss_2$；

S3.5.7:最后通过反向传播，对模型参数进行更新。

6.一种结构化文本关系抽取方法，其所述模块在于：

数据标注模块:用于对获取到的无监督数据进行手工标注，标注内容需要符合设计好的纲要schema:包括实体及其类型和实体对间的语义关系，实体在文本中起止位置信息索引，标注后的数据用于训练深度学习关系抽取模型；

命名实体识别模块，用于训练实体识别模型，为了在关系抽取过程中提取首实体和尾实体；

关系匹配模型，用于提取实体对之间的关系；

三元组信息存储模块:用于将抽取得到的三元组数据存入到数据库Mysql中。

7.一种结构化文本关系抽取方法，其特征在于，所述设备包括处理器以及存储器；

所述存储器用于存储程序代码，并将所述程序代码传输给所述处理器；

所述处理器用于根据所述程序代码中的指令执行权利要求1-5任一项所述的方法。

8.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质用于存储程序代码，所述程序代码被处理器执行时实现权利要求1-5任一项所述的方法。

一种结构化文本关系抽取方法、装置、设备

技术领域

[0001] 本申请涉及人工智能和自然语言处理技术领域，尤其涉及一种深度学习模型基于实体对的实体关系联合抽取方法、装置、设备和储存介质。

背景技术

[0002] 信息抽取定义为从自然语言文本中提取出指定类型的实体、关系、事件等信息，并形成结构化数据输出的技术。其具体包含的任务有：命名实体识别(NER)、关系抽取(Relation Extraction)和事件抽取(Event Extraction)。关系抽取作为信息抽取技术中重要的一个任务，就是在数据中找出主体与客体之间存在的关系，并将其表示为实体关系三元组，即(首实体，关系，尾实体)，简单叙述，表示为(subject,predicate,object)，缩写为(s,p,o)。提取出的三元组数据可用于知识图谱的构建，进而服务于像信息检索、智能问答等应用。

[0003] 现有的关系抽取主要有两大方案：1、流水线(pipeline)方法：分为两步，首先利用命名实体识别(NER)模型从文本中提取出全部实体，然后将全部可能的实体进行组合，通过多分类模型判断组合实体之间的关系属于哪一类；2、联合抽取方法(joint)：模型利用实体和关系之间的交互信息同时进行实体识别和关系分类任务，采取“一步走”的方式可有效缩短流水线方法中因为任务顺序而带来的误差传递问题。一般联合抽取方法可分为“参数共享的联合模型”和“结构化预测”，但联合抽取方法仍存在以下问题：

[0004] (1)参数共享是实体跟关系共用一个编码器，在解码阶段主体、客体和关系的抽取并不是同步的，而是利用编码层的信息先识别出首实体subject，再利用主体的特征信息识别相应尾实体object，最后根据主体客体的特征识别出相应的关系类型，并没用做到真正的“联合”。

[0005] (2)上述参数共享方法并没有实现真正的实体和关系之间的联合，研究关系抽取的学者们也提出了复杂的联合解码算法，没有将解码方案明确分为几个步骤。但是这种解码方法需要设计出相对复杂的解码过程，并且在三元组重叠问题上效果欠佳。

发明内容

[0006] 针对现有关系抽取技术中存在的问题，本申请提供了一种基于深度学习的文本关系抽取方法、装置、设备，采用了以BERT-base预训练模型为代表的深度学习算法。通过对数据进行构造，基于token-pair(实体对)的方式建模实体和实体对之间的关系，可以在保持一定速度的同时提高关系抽取的精度，又可以有效解决三元组重叠问题。

[0007] 本发明解决技术问题所采用的技术方案如下：

[0008] S1、数据构建和预处理；

[0009] S2、数据经过预处理后会划分为训练集、验证集和测试集，分别用于深度学习模型的训练、验证用于保存最优训练模型和对模型进行测试。

[0010] S3、搭建深度学习关系抽取模型

[0011] S4、同时经过实体识别层和关系判别层，并获取模型的损失值。

[0012] S5、通过反向传播和梯度下降，对模型参数进行更新。

[0013] S6、根据训练好的模型，针对未标注的数据提取三元组信息，挖掘句子中包含的语义信息。

[0014] S7、将得到的结果进行结构化存储。

[0015] 进一步叙述，步骤1中，收集文本数据用于训练模型。针对关系抽取数据，需要设计好纲要(schema)，表明关系三元组的具体类别：subject首实体类型、predicate关系类别和object尾实体类型，三者之间具有相对应的关系。通过编写代码定义数据预处理类，对数据进行去重、构建文本数据集。将数据存储为json文件，样本的形式以键-值对存在，每条样本数据须包含相应的文本text、关系类型spos列表，spos列表包含一条或多条实体关系数据，每条实体关系数据格式为首实体subject、predicate和object，分别表示首实体、关系和尾实体，以及subject(首实体)和object(尾实体)在文本text的位置信息区间索引。

[0016] 所述步骤二中，训练集用于深度学习模型的训练，验证集用于在训练过程中对模型进行验证，保存在验证集上评价指标最高的训练模型参数权重，利用保存验证过程中得分最高的模型参数在训练集上进行测试。

[0017] 所述步骤三中，构建深度学习模型主要包括以下过程：

[0018] 3.1)使用预训练语言模型如Bert-base模型作为编码其搭建关系抽取模型，主要包含如下几部分：1、识别首实体subject网络层和识别尾实体object网络层，这一部分属于实体抽取模块；2、根据首实体subject、尾实体object来判别关系类型；3、在关系抽取主任务上，添加辅助任务，用于后处理抽取的关系三元组个数，构成多任务学习，为模型增加鲁棒性。

[0019] 3.2)将训练数据中的每条样本按照字进行划分，如果按词切分可能会导致数据中的实体不在字典中，即OOV(英文全称：out of vocabulary)。如果当前句子为x，则划分后得到序列表示$x=[x_0,x_1,...,x_{n-1},x_n]$，根据Bert预训练语言模型的收入要求，令$x_0$=[CLS]，$x_n$=[SEP]，其中[CLS]标志位于句子的首位，而[SEP]标志位于句子的末尾。将得到的文本序列经过Bert模型，可得到结合上下文语义信息的word embedding(词嵌入)。

[0020] 3.3)抽取文本中的关系数据时，需要识别出subject首实体、object尾实体，在针对实体识别基础上可以采用token-pair(实体对)的方式，将实体的首尾视为一个整体去判别，在针对实体识别方面，通过两种类型张量tensor，分别用N_1和N_2来构建subject首实体和object尾实体输入，张量tensor维度为[n,seq_len,seq_len]，第一个维度n表示有多少中实体类型，第二个和第三个维度用来表示句子的长度，当实体属于第i类(i<n)，且该实体在文本中的位置索引信息为(s,t)，则$N_j[i,s,t]=1$(j=1or j=2)，当句子长度为l时，则有n×l(l+1)/2种组合，而我们只为文本中出现过的实体构建特征，可减少输入数据的复杂度；同时我们需要根据实体建模两者之间的关系，同理构造两种类型张量R_1和R_2，维度为[r,seq_len,seq_len]，r表示关系类别数目，当subject首实体的位置索引为(s_1,t_1)，object尾实体的位置索引为(s_2,t_2)，两者之间的关系为第k类(k<r)，则$R_1[k,s_1,s_2]=1$，$R_2[k,t_1,t_2]=1$，两者R_1和R_2表示根据首尾实体对的位置信息对predicate(关系)的匹配。

[0021] 3.4)将3.2中得到的序列x输入到bert-base模型中，有12层的encoder层，不同编码层学习到不同的语义信息，取最后四层输出的向量进行加权平均，得到包含上下文语义

的句向量[h_1,h_2,...,h_n],通过变换$q_{i,\alpha}=W_{q,\alpha}h_i+b_{q,\alpha}$和$k_{i,\alpha}=W_{k,\alpha}h_i+b_{k,\alpha}$得到向量序列[$q_{1,\alpha}$,$q_{2,\alpha}$,...,$q_{n,\alpha}$]和[$k_{1,\alpha}$,$k_{2,\alpha}$,...,$k_{n,\alpha}$],利用这两个向量序列可以构建一个实体识别的打分函数$f_\alpha(i,j)=q_{i,\alpha}^T k_{j,\alpha}$,表示$q_{i,\alpha}$与$k_{i,\alpha}$的内积,其中[i:j]是文本text中的一段连续子串,可构成一个实体。模型的首实体和尾实体层通过打分函数可得到两个向量e1和e2,向量维度是[n,seq_len,seq_len]。而在关系匹配层中,一样可以利用上述的打分函数,在3.3中构建输入关系匹配特征输入中,利用两个张量R1和R2建模实体及实体间关系。

[0022] 3.5)将3.4中向量e1跟e2分别引入全连接层dense输出实体对的向量表示e1～和e2～并将两者进行向量拼接得到向量表示e,再跟经过3.4中bert输出的句向量h计算attention(注意力)得分α_i=Attention(h_i,e),最后按照公式$S=\sum_{i=1}^{n}\alpha_i h_i$计算加权之后的句向量S,这样就得到融入实体信息的增强句向量,用于多任务中的实体关系个数预测。

[0023] 上述步骤四中,包括以下步骤:

[0024] 4.1)在长度为l的文本中,一共会有l(l+1)/2个不同的连续子序列,也即会出现l(l+1)/2个实体,则每个实体有两种选择:0或1,因为每条文本中的三元组个数不能缺点,所以变成了在l(l+1)/2类的多标签分类问题,于是损失函数需要用于多标签分类的损失函数$loss=\log(1+\sum_{(i,j)\in P\alpha}e^{-s_\alpha(i,j)})+\log(1+\sum_{(i,j)\in Q\alpha}e^{-s_\alpha(i,j)})$,其中$P_\alpha$是该样本所以类型为α的实体的首尾集合,$Q_\alpha$则是非实体或者非α类型的实体的首尾集合。同理在对关系进行匹配时,我们采用的思想也是类似于实体识别方式,只是将实体类型更换为关系类型,实体的位置索引更换为subject首实体、object尾实体的位置索引,故在关系匹配任务上也采用上述损失函数。

[0025] 4.2)辅助任务三元组个数判断任务中,是多类别分类的一种,针对这一任务,可以常用的交叉熵损失函数计算损失值loss。

[0026] 步骤五:通过反向传播和梯度下降,对模型参数进行更新。

[0027] 步骤六:通过训练好的模型,对未标注数据预测,提取新的关系三元组,具体为:

[0028] 6.1)在3.4中提及打分函数,训练过程中,通过对输入数据建模,让标注的主实体、客实体以及两者之间的关系得分大于0;而在预测过程中,只需要列出所有可能的实体,然后让打分函数验证主实体得分大于0,客实体得分大于0,然后基于提取出的主客实体匹配关系得分大于0,满足上述条件的三元组才是我们需要的最终输出。

[0029] 步骤7,将模型输出结果按照对应关系输入MySQL数据库中,进行数据的存储。方便落地应用于知识图谱构建,智能问答中。

附图说明

[0030] 为了更清楚地说明本申请实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要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申请的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动性的前提下,还可以根据这些附图获得其它的附图。

[0031] 图1为本申请实施例提供的对数据集进行预处理的流程图;

[0032] 图2为本申请实施例提供的构建文本关系抽取模型的流程图。

具体实施方式

[0033] 为了使本技术领域的人员更好地理解本申请方案，下面将结合本申请实施例中的附图，对本申请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申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，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，都属于本申请保护的范围。

[0034] 为了便于理解，首先对本申请实施例提供了一种基于深度学习模型的文本关系抽取方法进行详细介绍：

[0035] 参照图1和图2，一种基于深度学习模型的结构化文本关系抽取的方法、装置和设备包括以下步骤：

[0036] 步骤1、首先需要收集文本数据，构建文本数据集以及预处理。

[0037] 根据设计好的schema(纲要)，这是为了规范结构化数据的表达，关系抽取的每条数据必须满足纲要(schema)预先定义的实体对象及其类型。通过编写代码定义数据预处理类，对数据进行去重、构建文本数据集。将数据存储为json文件，样本的形式以键-值对存在，每条样本数据须包含相应的文本text、关系类型spos列表，spos列表包含一条或多条实体关系数据，每条实体关系数据格式为subject、predicate和object，分别表示头实体、关系和尾实体，以及subject首实体和object尾实体在文本text的位置信息区间索引。

[0038] 步骤2、对数据进行划分，得到训练集、验证集和测试集。

[0039] 步骤3、搭建深度学习关系抽取模型。

[0040] 3.1使用预训练语言模型如Bert-base模型作为编码其搭建关系抽取模型，主要包含如下几部分：1、识别首实体subject网络层和识别尾实体object网络层，这一部分属于实体抽取模块；2、根据subject首实体、object尾实体来判别关系类型；3、在关系抽取主任务上，添加辅助任务，用于后处理抽取的关系三元组个数，构成多任务学习，为模型增加鲁棒性。

[0041] 3.2将训练数据中的每条样本按照字进行划分，如果按词切分可能会导致数据中的实体不在字典中，即OOV(英文全称:out of vocabulary)。如果当前句子为x，则划分后得到序列表示$x=[x_0,x_1,...,x_{n-1},x_n]$，根据bert预训练语言模型的收入要求，令$x_0=$[CLS]，$x_n=$[SEP]，其中[CLS]标志位于句子的首位，而[SEP]标志位于句子的末尾。将得到的文本序列经过bert模型，可得到结合上下文语义信息的word embedding(词嵌入)。

[0042] 3.3抽取文本中的关系数据时，需要识别出subject首实体、object尾实体，在针对实体识别基础上可以采用token-pair(实体对)的方式，将实体的首尾视为一个整体去判别，在针对实体识别方面，通过两种类型张量tensor，分别用N_1和N_2来构建subject首实体和object尾实体输入，张量tensor维度为[n,seq_len,seq_len]，第一个维度n表示有多少中实体类型，第二个和第三个维度用来表示句子的长度，当实体属于第i类($i<n$)，且该实体在文本中的位置索引信息为(s,t)，则$N_j[i,s,t]=1$($j=1$or $j=2$)，当句子长度为l时，则有$n\times l(l+1)/2$种组合，而我们只为文本中出现过的实体构建特征，可减少输入数据的复杂度；同时我们需要根据实体建模两者之间的关系，同理构造两种类型张量R_1和R_2，维度为[r,seq_len,seq_len]，r表示关系类别数目，当subject首实体的位置索引为(s_1,t_1)，object尾实体的位置索引为(s_2,t_2)，两者之间的关系为第k类($k<r$)，则$R_1[k,s_1,s_2]=1$，$R_2[k,t_1,t_2]=1$，两者R_1和R_2表示根据首尾实体对的位置信息对predicate(关系)的匹配。

[0043] 3.4将3.2中得到的序列x输入到bert-base模型中，有12层的encoder层，不同编码层学习到不同的语义信息，取最后四层输出的向量进行加权平均，得到包含上下文语义的句向量$[h_1,h_2,...,h_n]$，通过变换$q_{i,\alpha}=W_{q,\alpha}h_i+b_{q,\alpha}$和$k_{i,\alpha}=W_{k,\alpha}h_i+b_{k,\alpha}$得到向量序列$[q_{1,\alpha},q_{2,\alpha},...,q_{n,\alpha}]$和$[k_{1,\alpha},k_{2,\alpha},...,k_{n,\alpha}]$，利用这两个向量序列可以构建一个实体识别的打分函数$f_\alpha(i,j)=q_{i,\alpha}^T k_{j,\alpha}$，表示$q_{i,\alpha}$与$k_{i,\alpha}$的内积，其中[i:j]是文本text中的一段连续子串，可构成一个实体。模型的首实体和尾实体层通过打分函数可得到两个向量e1和e2，向量维度是[n,seq_len,seq_len]。而在关系匹配层中，一样可以利用上述的打分函数，在3.3中构建输入关系匹配特征输入中，利用两个张量R1和R2建模实体及实体间关系。

[0044] 3.5将3.4中e1跟e2分别引入全连接层输出实体对的向量表示e1～和e2～并将两者进行向量拼接得到向量表示e，再跟经过bert输出的句向量h计算attention(注意力)得分$\alpha_i=\text{Attention}(h_i,e)$，最后按照公式$S=\sum_{i=1}^{n}\alpha_i h_i$计算加权之后的句向量S，这样就得到融入实体信息的增强句向量，用于多任务中的实体关系个数预测。

[0045] 步骤4、经过实体识别层和关系判别层，并获取模型的损失值。

[0046] 在长度为l的文本中，一共会有l(l+1)/2个不同的连续子序列，也即会出现l(l+1)/2个实体，则每个实体有两种选择:0或1，因为每条文本中的三元组个数不能缺点，所以变成了在l(l+1)/2类的多标签分类问题，于是损失函数需要用于多标签分类的损失函数$loss=\log(1+\sum_{(i,j)\in P\alpha}e^{-s_\alpha(i,j)})+\log(1+\sum_{(i,j)\in Q\alpha}e^{-s_\alpha(i,j)})$，其中$P_\alpha$是该样本所以类型为α的实体的首尾集合，$Q_\alpha$则是非实体或者非α类型的实体的首尾集合。同理在对关系进行匹配时，我们采用的思想也是类似于实体识别方式，只是将实体类型更换为关系类型，实体的位置索引更换为subject首实体、object尾实体的位置索引，故在关系匹配任务上也采用上述损失函数。

[0047] 辅助任务三元组个数判断任务中，是多类别分类的一种，针对这一任务，可以常用的交叉熵损失函数计算损失值loss。

[0048] 步骤5、通过反向传播和梯度下降，对模型参数进行更新

[0049] 步骤6、通过训练好的模型，对未标注数据预测，提取新的关系三元组，具体为：

[0050] 在3.4中提及打分函数，训练过程中，通过对输入数据建模，让标注的主实体、客实体以及两者之间的关系得分大于0;而在预测过程中，只需要列出所有可能的实体，然后让打分函数验证主实体得分大于0，客实体得分大于0，然后基于提取出的主客实体匹配关系得分大于0，满足上述条件的三元组才是我们需要的最终输出。

[0051] 步骤7、步骤6输出的结果按照相应的格式存入到MySQL数据库中，进行数据的存储，方便落地应用于知识图谱构建，智能问答等任务中。

[0052] 本发明实施提供了一种结构化文本关系抽取任务模型训练方法，通过采用联合抽取方法，基于token-pair(实体对)的方式建模实体和实体对之间的关系，可以在保持一定速度的同时提高关系抽取的精度，又可以有效解决三元组重叠问题。

[0053] 本申请实施例还提供了一种文本关系抽取方法设备，设备包括处理器以及存储器；

[0054] 存储器用于存储程序代码，并将程序代码传输给处理器；

[0055] 处理器用于根据程序代码中的指令执行前述方法实施例中的光斑质量判别方法。

[0056] 本申请实施例还提供了一种计算机可读存储介质，计算机可读存储介质用于存储

程序代码,程序代码被处理器执行时实现前述方法实施例中的光斑质量判别方法。

[0057] 所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的装置和单元的具体工作过程,可以参考前述方法实施例中的对应过程,在此不再赘述。

[0058] 本申请的说明书及上述附图中的术语“第一”、“第二”、“第三”、“第四”等(如果存在)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,以便这里描述的本申请的实施例例如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。

[0059] 应当理解,在本申请中,“至少一个(项)”是指一个或者多个,“多个”是指两个或两个以上。“和/或”,用于描述关联对象的关联关系,表示可以存在三种关系,例如,“A和/或B”可以表示:只存在A,只存在B以及同时存在A和B三种情况,其中A,B可以是单数或者复数。字符“/”一般表示前后关联对象是一种“或”的关系。“以下至少一项(个)”或其类似表达,是指这些项中的任意组合,包括单项(个)或复数项(个)的任意组合。例如,a,b或c中的至少一项(个),可以表示:a,b,c,“a和b”,“a和c”,“b和c”,或“a和b和c”,其中a,b,c可以是单个,也可以是多个。

[0060] 在本申请所提供的几个实施例中,应该理解到,所揭露的装置和方法,可以通过其它的方式实现。例如,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,装置或单元的间接耦合或通信连接,可以是电性,机械或其它的形式。

[0061] 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。

[0062] 另外,在本申请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。

[0063] 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,本申请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,包括若干指令用以通过一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本申请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、移动硬盘、只读存储器(英文全称:Read-Only Memory,英文缩写:ROM)、随机存取存储器(英文全称:Random Access Memory,英文缩写:RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0064] 以上所述,以上实施例仅用以说明本申请的技术方案,而非对其限制;尽管参照前

述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

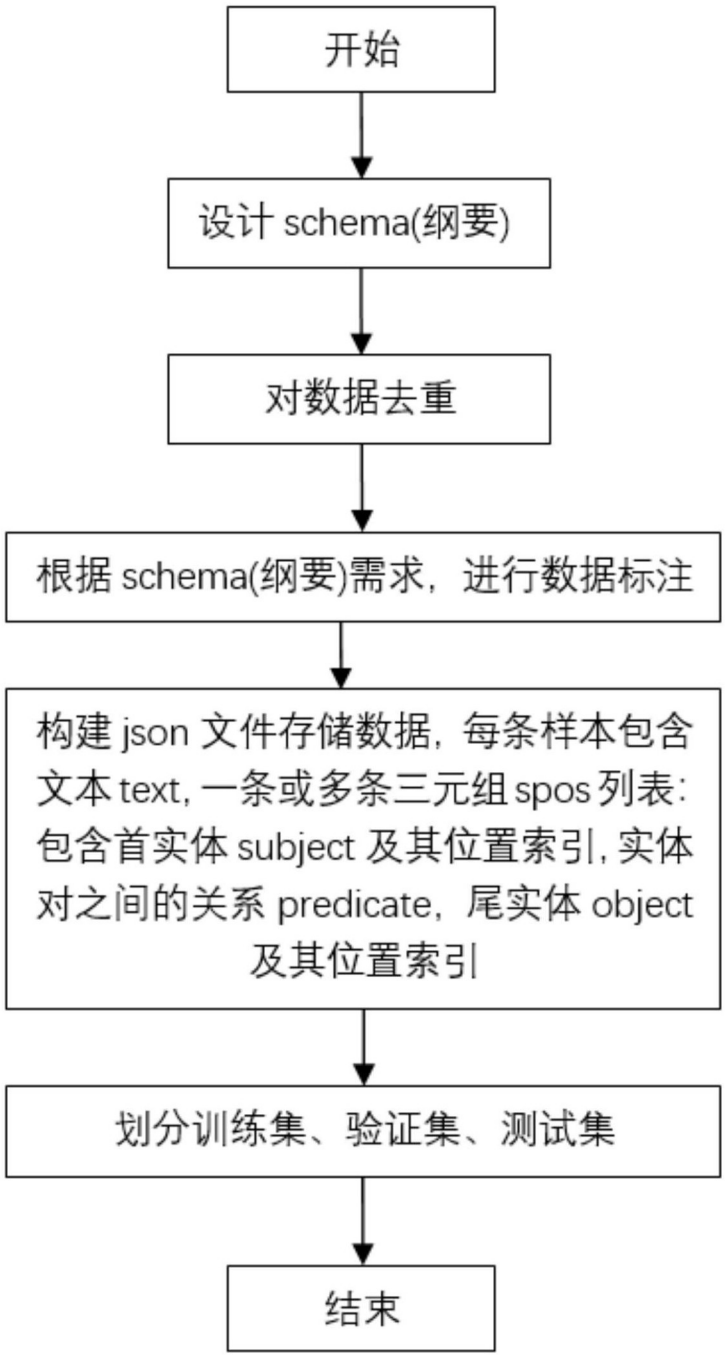

开始
设计 schema(纲要)
对数据去重
根据 schema(纲要)需求，进行数据标注
构建 json 文件存储数据，每条样本包含文本 text，一条或多条三元组 spos 列表：包含首实体 subject 及其位置索引，实体对之间的关系 predicate，尾实体 object 及其位置索引
划分训练集、验证集、测试集
结束

图1

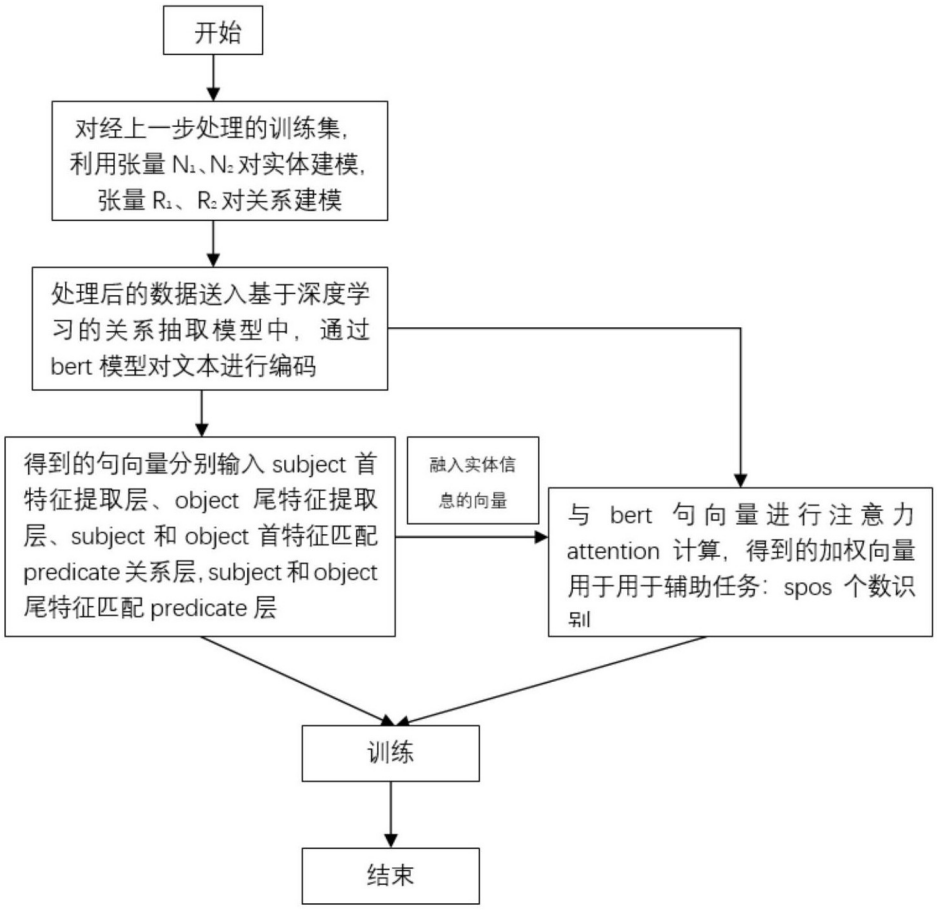

开始
对经上一步处理的训练集，利用张量 N_1、N_2 对实体建模，张量 R_1、R_2 对关系建模
处理后的数据送入基于深度学习的关系抽取模型中，通过 bert 模型对文本进行编码
得到的句向量分别输入 subject 首特征提取层、object 尾特征提取层、subject 和 object 首特征匹配 predicate 关系层，subject 和 object 尾特征匹配 predicate 层
融入实体信息的向量
与 bert 句向量进行注意力 attention 计算，得到的加权向量用于用于辅助任务：spos 个数识别
训练
结束

图2