(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 110851577 A
(43)申请公布日 2020.02.28

(21)申请号 201911044753.1

(22)申请日 2019.10.30

(71)申请人 国网江苏省电力有限公司电力科学研究院
地址 210029 江苏省南京市鼓楼区凤凰西街243号
申请人 国家电网有限公司
江苏省电力试验研究院有限公司

(72)发明人 吴宁 何维民 邹云峰 赵洪莹

(74)专利代理机构 南京纵横知识产权代理有限公司 32224
代理人 张赏

(51)Int.Cl.
G06F 16/332(2019.01)
G06F 16/36(2019.01)
G06F 16/35(2019.01)
G06Q 50/06(2012.01)

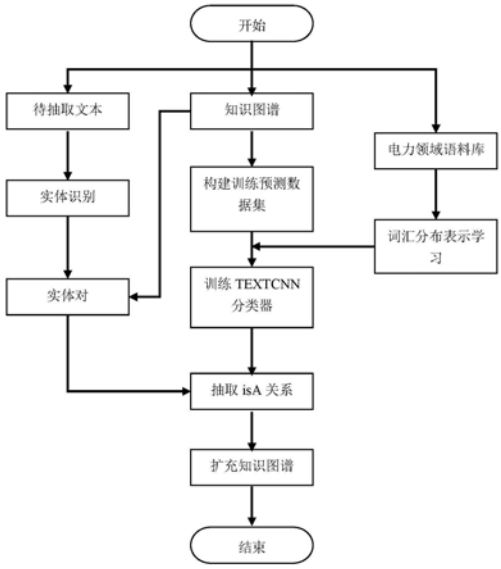
权利要求书2页 说明书5页 附图1页

(54)发明名称

一种电力领域的知识图谱扩充方法及装置

(57)摘要

本发明公开了一种电力领域的知识图谱扩充方法及装置，通过对电力营销领域内部知识的场景预设，经由人工审核标注后，利用深度卷积神经网络训练学习，进而完成对于意图场景的训练，最终实现了对电力领域知识图谱进行扩充。本方法针对知识图谱提供更多场景内知识补充，提高电力营销知识图谱的深度，同时降低人工维护图谱带来的相关成本。



CN 110851577 A

1.一种电力领域的知识图谱扩充方法,其特征在于,包括:

获取候选实体;

查询所述候选实体对应的词向量;

计算所述候选实体中两两实体的词向量的差值得到向量偏差;

采用训练好的TEXTCNN分类器对所得到的向量偏差进行分类,得到两个实体是否存在isA关系;

抽取分类为isA关系的实体对扩充知识图谱。

2.根据权利要求1所述的一种电力领域的知识图谱扩充方法,其特征在于,所述获取候选实体,包括:

通过电力领域命名实体识别模块对待抽取文本进行实体识别获取候选实体,

或者从知识图谱中选取实体作为候选实体。

3.根据权利要求1所述的一种电力领域的知识图谱扩充方法,其特征在于,所述查询所述候选实体对应的词向量,包括:

对知识图谱中的实体进行词汇分布表示学习,得到每个实体对应的词向量;

从词汇分布表示学习结果中查询候选实体对应的词向量。

4.根据权利要求3所述的一种电力领域的知识图谱扩充方法,其特征在于,所述词汇分布表示学习采用Skip-gram模型。

5.根据权利要求1所述的一种电力领域的知识图谱扩充方法,其特征在于,所述TEXTCNN分类器训练过程如下:

抽取知识图谱中的isA关系和notisA关系作为数据集;所述数据集中的每条记录由两个实体以及这两个实体是否存在isA关系构成;

把数据集分成训练数据集D和测试数据集T;所述训练数据集D用于训练TEXTCNN分类器,所述测试数据集T用于评估TEXTCNN分类器;

将所述训练数据集D中每条记录的两个实体的向量偏差作为输入特征,输入TEXTCNN分类器,输出两个实体是否存在isA关系,用Y代表构成isA关系,用N代表不构成isA关系。

6.根据权利要求5所述的一种电力领域的知识图谱扩充方法,其特征在于,所述抽取知识图谱中的isA关系和notisA关系后,进行标注纠正该数据集。

7.根据权利要求1所述的一种电力领域的知识图谱扩充方法,其特征在于,所述抽取分类为isA关系的实体对扩充知识图谱,包括:

根据本体公理“isA(X,Y), isA(Y,Z) →isA(X,Z)”,推出知识图谱中更多的isA关系,其中,isA(X,Y)表示实体X和实体Y构成isA关系。

8.一种电力领域的知识图谱扩充装置,其特征在于,包括:

采集模块,用于获取候选实体;

查询模块,用于查询所述候选实体对应的词向量;、

计算模块,用于计算所述候选实体中两两实体的词向量的差值得到向量偏差;

分类模块,用于采用TEXTCNN分类器对所得到的向量偏差进行分类,得到两个实体是否存在isA关系;

以及扩充模块,用于抽取分类为isA关系的实体对扩充知识图谱。

9.根据权利要求8所述的一种电力领域的知识图谱扩充装置,其特征在于,所述采集模

块,具体用于,

通过电力领域命名实体识别模块对待抽取文本进行实体识别获取候选实体,

或者从知识图谱中选取实体作为候选实体。

10.根据权利要求8所述的一种电力领域的知识图谱扩充装置,其特征在于,所述分类模块,具体还用于,

抽取知识图谱中的isA关系和notisA关系作为数据集;所述数据集中的每条记录由两个实体以及这两个实体是否存在isA关系构成;

把数据集分成训练数据集D和测试数据集T;所述训练数据集D用于训练TEXTCNN分类器,所述测试数据集T用于评估TEXTCNN分类器;

将所述训练数据集D中每条记录的两个实体的向量偏差作为输入特征,输入TEXTCNN分类器,输出两个实体是否存在isA关系,用Y代表构成isA关系,用N代表不构成isA关系。

11.根据权利要求8所述的一种电力领域的知识图谱扩充装置,其特征在于,所述扩充模块,具体用于,根据本体公理“isA(X,Y), isA(Y,Z) →isA(X,Z)”,推出知识图谱中更多的isA关系,其中,isA(X,Y)表示实体X和实体Y构成isA关系。

一种电力领域的知识图谱扩充方法及装置

技术领域

[0001] 本发明属于电力系统客服技术领域,特别涉及一种电力领域的知识图谱扩充方法及装置。

背景技术

[0002] 目前的电力行业的客服问答系统中,存在着推理类问答回答准确率不高,提高推理类问题是提升电力客服智能问答性能的关键,因此,需要找出那些隐藏或隐含在知识图谱中的知识来充实知识图谱,以满足回答推理类问题中对知识的需求。这些知识主要包括领域实体间的上下位、部分与整体、等价等关系下隐藏的知识,例如“交电费的途径之一是支付宝”隐藏地表达了“交电费途径之一是电子缴费渠道”,以及隐含在关系之间或实体之间的规律性的关联知识。因此,需要扩充电力客服领域知识图谱中隐藏和隐含的知识,以应对推理类问题的问答是电力客服智能问答研究的关键和难点之一。目前构建图谱中,缺失了上下位关系、部分与整体、等价等用于推理的知识。

发明内容

[0003] 本发明的目的在于提供一种电力领域的知识图谱扩充方法及装置,通过找出那些隐藏或隐含在知识图谱中的知识来充实知识图谱,以满足回答推理类问题中对知识的需求。

[0004] 为达到上述目的,本发明采用的技术方案如下:

[0005] 本发明实施例提供一种电力领域的知识图谱扩充方法,包括:

[0006] 获取候选实体;

[0007] 查询所述候选实体对应的词向量;

[0008] 计算所述候选实体中两两实体的词向量的差值得到向量偏差;

[0009] 采用训练好的TEXTCNN分类器对所得到的向量偏差进行分类,得到两个实体是否存在isA关系;

[0010] 抽取分类为isA关系的实体对扩充知识图谱。

[0011] 进一步的,所述获取候选实体,包括:

[0012] 通过电力领域命名实体识别模块对待抽取文本进行实体识别获取候选实体,

[0013] 或者从知识图谱中选取实体作为候选实体。

[0014] 进一步的,所述查询所述候选实体对应的词向量,包括:

[0015] 对知识图谱中的实体进行词汇分布表示学习,得到每个实体对应的词向量;

[0016] 从词汇分布表示学习结果中查询候选实体对应的词向量。

[0017] 进一步的,所述词汇分布表示学习采用Skip-gram模型。

[0018] 进一步的,所述TEXTCNN分类器训练过程如下:

[0019] 抽取知识图谱中的isA关系和notisA关系作为数据集;所述数据集中的每条记录由两个实体以及这两个实体是否存在isA关系构成;

[0020] 把数据集分成训练数据集D和测试数据集T;所述训练数据集D用于训练TEXTCNN分类器,所述测试数据集T用于评估TEXTCNN分类器;

[0021] 将所述训练数据集D中每条记录的两个实体的向量偏差作为输入特征,输入TEXTCNN分类器,输出两个实体是否存在isA关系,用Y代表构成isA关系,用N代表不构成isA关系。

[0022] 进一步的,所述抽取知识图谱中的isA关系和notisA关系后,进行标注纠正该数据集。

[0023] 进一步的,所述抽取分类为isA关系的实体对扩充知识图谱,包括:

[0024] 根据本体公理“isA(X,Y),isA(Y,Z)→isA(X,Z)”,推出知识图谱中更多的isA关系,其中,isA(X,Y)表示实体X和实体Y构成isA关系。

[0025] 本发明实施例还提供一种电力领域的知识图谱扩充装置,包括:

[0026] 采集模块,用于获取候选实体;

[0027] 查询模块,用于查询所述候选实体对应的词向量;、

[0028] 计算模块,用于计算所述候选实体中两两实体的词向量的差值得到向量偏差;

[0029] 分类模块,用于采用TEXTCNN分类器对所得到的向量偏差进行分类,得到两个实体是否存在isA关系;

[0030] 以及扩充模块,用于抽取分类为isA关系的实体对扩充知识图谱。

[0031] 进一步的,所述采集模块,具体用于,

[0032] 通过电力领域命名实体识别模块对待抽取文本进行实体识别获取候选实体,

[0033] 或者从知识图谱中选取实体作为候选实体。

[0034] 进一步的,所述分类模块,具体还用于,

[0035] 抽取知识图谱中的isA关系和notisA关系作为数据集;所述数据集中的每条记录由两个实体以及这两个实体是否存在isA关系构成;

[0036] 把数据集分成训练数据集D和测试数据集T;所述训练数据集D用于训练TEXTCNN分类器,所述测试数据集T用于评估TEXTCNN分类器;

[0037] 将所述训练数据集D中每条记录的两个实体的向量偏差作为输入特征,输入TEXTCNN分类器,输出两个实体是否存在isA关系,用Y代表构成isA关系,用N代表不构成isA关系。

[0038] 进一步的,所述扩充模块,具体用于,根据本体公理“isA(X,Y),isA(Y,Z)→isA(X,Z)”,推出知识图谱中更多的isA关系,其中,isA(X,Y)表示实体X和实体Y构成isA关系。

[0039] 本发明通过对电力领域内部知识的场景预设,经由人工审核标注后,利用Skip-gram模型训练学习,进而完成对于意图场景的分类训练,最终实现对电力领域知识图谱的扩充,本发明能够针对知识图谱提供更多场景内知识补充,提高电力营销知识图谱的深度,同时降低人工维护图谱带来的相关成本。

附图说明

[0040] 图1为本发明所提出的知识图谱扩充方法的整体流程图。

具体实施方式

[0041] 下面对本发明作进一步描述。以下实施例仅用于更加清楚地说明本发明的技术方案，而不能以此来限制本发明的保护范围。

[0042] 本发明提供一种电力领域的知识图谱扩充方法，具体包括：

[0043] 步骤1：利用现有的电力知识图谱构建训练数据集D和预测数据集T。

[0044] 针对电力知识场景预定义一些意图场景(例如隶属，存在，具备等特定关系类)准备数据，并针对数据进行标注，切分训练数据集与预测数据集。

[0045] IsA关系是知识图谱中的最核心的一种关系，它定义了某个概念B是概念A的一种，比如：“开增值税普通发票”是一种(isA)“开发票”业务，“开增值税专用发票”也是一种(isA)“开发票”业务。notisA关系与isA关系相反，它表示某个概念B不是A的一种，比如电子发票不是一种(notisA)电度表。

[0046] 在本步骤中，抽取知识图谱中的isA关系和notisA关系作为数据集，然后人工标注纠正该数据集。标注后的数据集中的每条记录由两个实体以及这两个实体是否存在isA关系构成，比如：(开增值税普通发票，开发票，Y)，Y代表可以构成isA关系。

[0047] 然后把数据集分成两部分：训练数据集D和测试数据集T。训练数据集D主要用来训练判断isA关系的isA关系分类器，而测试数据集T用来评估isA关系分类器性能。

[0048] 步骤2：训练isA关系分类器

[0049] 本步骤中所要训练的神经网络模型的输入是多对实体，输出是每一对实体能否构成isA关系。该模型分为两部分：词汇分布表示学习和TEXTCNN分类器。

[0050] 采用步骤1中的训练数据集D训练TEXTCNN分类器，测试数据集T评估TEXTCNN分类器。

[0051] 词汇分布表示学习：将知识图谱中的实体映射为向量。词汇分布表示学习使用Skip-gram模型，输入为电力领域的语料库，输出为数值型的词向量。电力领域的语料库的内容为一段文本，输出是这段文本中所有单词的词向量，比如这段文本里有三个单词，那么就会输出着三个单词的词向量。词向量就是一组0-1之间的小数，举个例子：电费的词向量可以是(0.1,0.2,0.33,0.5,0.6)，说白了一个词由一组数字表示，这一组数字称之为词向量。

[0052] 相对于基于神经网络的模型，其训练代价小，效率高。对于本发明中的电力领域文本，需要通过实验设置上下文窗口和参数α的值来改变词向量的分布，并调整至合理位置，避免过拟合。至此，词向量隐含了词语的语义关系。然后对词向量两两作差得到向量偏差，向量偏差就刻画了词语间的语义关系。

[0053] TEXTCNN分类器：使用TEXTCNN作为划分偏差向量的分类器，将向量偏差作为输入特征，输出是两个实体能否构成isA关系。

[0054] 表1 TEXTCNN参数列表

[0055]

参数名称	参数值
Batch size	64
word embedding size	200
kernel size	128
Filter Window	2,3,4,5

[0056] 步骤3:抽取isA关系

[0057] 通过电力领域命名实体识别模块对待抽取文本进行实体识别或者从知识图谱中选取实体来获取候选实体,然后使用步骤2训练的isA关系分类器对这些候选实体判断能否构成isA关系。具体实施步骤如下:

[0058] 从词汇分布表示学习结果中,查询实体对应的词向量;

[0059] 计算两两实体之间的词向量差值得到向量偏差;

[0060] 使用TEXTCNN分类器对isA和notisA根据向量偏差进行二分类;

[0061] 分类结果为isA的向量偏差对应的两个实体之间的关系即为抽取目标。

[0062] 步骤4:根据抽取得到的isA关系扩充知识图谱

[0063] 即结合本体公理“isA(X,Y),isA(Y,Z)→isA(X,Z)”,推出知识图谱中更多的isA关系。比如通过步骤3抽取到isA关系”单相电能表”是”交流电能表”的一种,又已知关系”交流电能表”是一种”电能表”,因此可以推理得到”单相电能表”是一种”电能表”,可以将推理得到的关系存入知识图谱中从而丰富了知识图谱内容。

[0064] 本发明实施例还提供一种电力领域的知识图谱扩充装置,包括:

[0065] 采集模块,用于获取候选实体;

[0066] 查询模块,用于查询所述候选实体对应的词向量;、

[0067] 计算模块,用于计算所述候选实体中两两实体的词向量的差值得到向量偏差;

[0068] 分类模块,用于采用TEXTCNN分类器对所得到的向量偏差进行分类,得到两个实体是否存在isA关系;

[0069] 以及扩充模块,用于抽取分类为isA关系的实体对扩充知识图谱。

[0070] 进一步的,所述采集模块,具体用于,

[0071] 通过电力领域命名实体识别模块对待抽取文本进行实体识别获取候选实体,

[0072] 或者从知识图谱中选取实体作为候选实体。

[0073] 进一步的,所述分类模块,具体还用于,

[0074] 抽取知识图谱中的isA关系和notisA关系作为数据集;所述数据集中的每条记录由两个实体以及这两个实体是否存在isA关系构成;

[0075] 把数据集分成训练数据集D和测试数据集T;所述训练数据集D用于训练TEXTCNN分类器,所述测试数据集T用于评估TEXTCNN分类器;

[0076] 将所述训练数据集D中每条记录的两个实体的向量偏差作为输入特征,输入TEXTCNN分类器,输出两个实体是否存在isA关系,用Y代表构成isA关系,用N代表不构成isA关系。

[0077] 进一步的,所述扩充模块,具体用于,根据本体公理“isA(X,Y),isA(Y,Z)→isA(X,Z)”,推出知识图谱中更多的isA关系,其中,isA(X,Y)表示实体X和实体Y构成isA关系。

[0078] 值得指出的是,该装置实施例是与上述方法实施例对应的,上述方法实施例的实现方式均适用于该装置实施例中,并能达到相同或相似的技术效果,故不在此赘述。

[0079] 本领域内的技术人员应明白,本申请的实施例可提供为方法、系统、或计算机程序产品。因此,本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产

品的形式。

[0080] 本申请是参照根据本申请实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0081] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0082] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上,使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理,从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0083] 最后应当说明的是:以上实施例仅用以说明本发明的技术方案而非对其限制,尽管参照上述实施例对本发明进行了详细的说明,所属领域的普通技术人员应当理解:依然可以对本发明的具体实施方式进行修改或者等同替换,而未脱离本发明精神和范围的任何修改或者等同替换,其均应涵盖在本发明的权利要求保护范围之内。

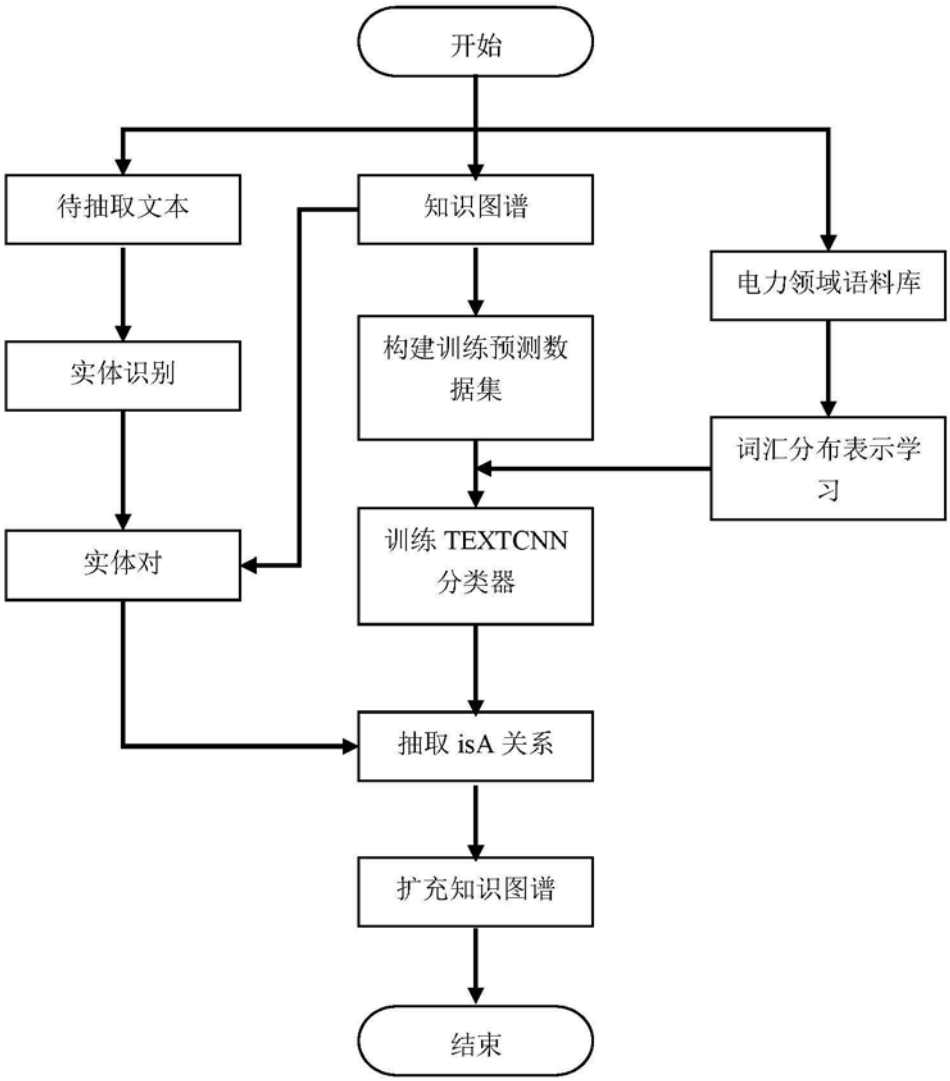

开始
待抽取文本
知识图谱
电力领域语料库
实体识别
构建训练预测数据集
词汇分布表示学习
实体对
训练 TEXTCNN 分类器
抽取 isA 关系
扩充知识图谱
结束

图1