(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112836482 B
(45)授权公告日 2024.02.23

(21)申请号 202110181755.6

(22)申请日 2021.02.09

(65)同一申请的已公布的文献号
申请公布号 CN 112836482 A

(43)申请公布日 2021.05.25

(73)专利权人 浙江工商大学
地址 310000 浙江省杭州市下沙高教园区学正街18号

(72)发明人 李玉娥 董黎刚 蒋献 吴梦莹 诸葛斌

(74)专利代理机构 杭州五洲普华专利代理事务所(特殊普通合伙) 33260
专利代理师 朱林军

(51)Int.Cl.
G06F 40/186(2020.01)
G06F 40/30(2020.01)
G06F 16/36(2019.01)
G06F 16/35(2019.01)

(56)对比文件
CN 108897857 A,2018.11.27
CN 109408812 A,2019.03.01
CN 109582789 A,2019.04.05
CN 110688491 A,2020.01.14
CN 111767409 A,2020.10.13
US 2018329884 A1,2018.11.15
US 2020073933 A1,2020.03.05
US 9984062 B1,2018.05.29
WO 2020232861 A1,2020.11.26
董黎刚 等.面向SDN的动态网络策略部署与实现.电信科学.2016,第32卷(第10期),137-138,148-149.
Yuanli.Towards Chinese clinical named entity recognition by dynamic embedding using domain-specific knowledge.Journal of Biomedical Informatics.2020,第106卷1-9.
陈文实. 基于编码解码器与深度主题特征抽取的多标签文本分类.南京师大学报.2019,第42卷(第4期),61-68.
杨丹浩;吴岳辛;范春晓.一种基于注意力机制的中文短文本关键词提取模型.计算机科学.2020,(第01期),199-204.
冯读娟;杨璐;严建峰.基于双编码器结构的文本自动摘要研究.计算机工程.2020,(第06期),66-70.

审查员 杨楚莹

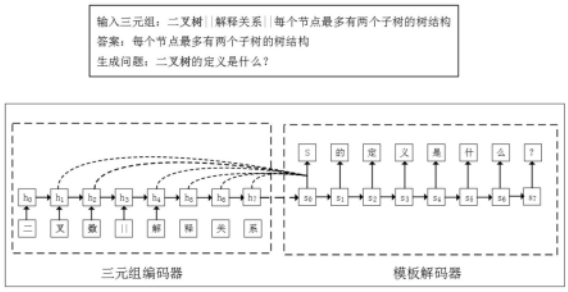
权利要求书2页 说明书7页 附图5页

(54)发明名称

一种基于模板的序列生成模型生成问题的方法及装置

(57)摘要

本发明公开了一种基于模板的序列生成模型生成问题的方法及系统,构建文本抽取模型对用于生成问题的文本进行分类,获得预测文本;构建文本识别模型,文本识别模型将文本转换成向量表示,基于向量表示和该模型中设置的字词级和语句级注意力机制,获得字向量和词向量对应的输入序列,基于所述输入序列进行分类得到预测文本的预测关系标签;构建序列生成模型,编码单元接收问题并映射成多元向量输入给模板解码单元,对此时解码单元输出的问题模板,根据所述预测根据所述预测关系标签对所述问题模板进行替换。



CN 112836482 B

1.一种基于模板的序列生成模型生成问题的方法,其特征在于,

构建文本抽取模型:对用于生成问题的文本输入到联合模型中,利用所述联合模型对所述文本进行分类,得到预测文本;

构建文本关系识别模型:根据所述预测文本内容获取文本语义特征向量,利用带有关系标签的训练集训练所述文本关系识别模型,其中根据所述文本关系识别模型中设置的字词级和语句级的注意力机制对所述文本语义特征向量进行训练,可获得字向量和词向量对应的输入序列,根据所述对应的输入序列对所述字向量和词向量进行分类,获取预测文本的预测关系标签;

构建序列生成模型,所述序列生成模型包括编码单元和模板解码单元,输入问题给所述编码单元,并根据所述问题语义映射成对应的多元组向量,并将所述多元组向量依次输入到模板解码单元中,且由模板解码单元输出为问题模板,根据所述预测关系标签对所述问题模板进行替换;

其中,构建所述文本抽取模型步骤包括:

所述联合模型为双向长短记忆网络模型和条件随机场模型的联合结构,所述文本抽取模型包括字词向量表示、语句特征提取和语句级的序列标注三层结构,

其中对所述文本数据进行序列标注得到训练集文本,

读取所述训练集文本作为双向长短记忆网络模型的输入进行无监督训练,使得双向长短记忆网络模型初始化所述训练集文本的权值以及构建特征空间;

基于所述特征空间和文本的权值,利用条件随机场模型对所述训练集文本进行有监督学习;

使用归一化函数获得所述训练集文本中各个字词的分类概率;

利用得到的分类概率进行分类得到所述预测文本。

2.根据权利要求1所述的一种基于模板的序列生成模型生成问题的方法,其特征在于,所述文本关系识别模型包括输入表示层、字词级层、语句级层以及实体关系分类层,其中,

输入层用于将输入的文字转换成向量的表示,用于获取文本语义特征向量;

字词级层学习用于学习所述文本上下文的内容信息,得到每个字词对文本语义信息的重要程度;

语句级层学习用于根据上下文语句,给每个输出字词分配不同的权重,获取字词对语句信息的重要程度;

实体关系分类层对语句信息的重要程度进行归一化处理,得到向量的关系标签,从而对实体间的关系进行分类。

3.如权利要求1所述的一种基于模板的序列生成模型生成问题的方法,其特征在于,构建文本关系识别模型步骤包括:

将所述预测文本内容输入到预训练的Word2vec模型中将文字转换成低维稠密的向量表示,其中所述向量表示为文本语义特征向量,将所述文本语义特征向量输入至构建文本关系识别模型的字词级层中,获得所述文本语义特征向量包含的字义信息、词义信息和上下文信息。

4.如权利要求1所述的一种基于模板的序列生成模型生成问题的方法,其特征在于,构建文本关系识别模型步骤还包括:

获得文本语义特征向量后,将所述文本语义特征向量输入至构建文本关系识别模型的语句级学习层中,获得所述文本语义特征向量每个字词的权值,根据加权平均值法得到每个字词的attention值;

对所述attention值进行归一化处理,得到所述文本语义特征向量的预测关系标签,根据所述预测关系标签对所述预测文本进行分类得到语句实体。

5.如权利要求1所述的一种基于模板的序列生成模型生成问题的方法,其特征在于,所述预测关系标签采用HowNet定义的16种标签和5种自定义标签。

6.如权利要求1所述的一种基于模板的序列生成模型生成问题的方法,其特征在于,将所述输入问题映射成对应的多元组向量时的内容至少包括主题实体、实体关系、实体,其中所述输入问题是根据所述主题实体和所述实体关系提出,且能由所述实体回答。

7.一种基于模板的序列生成模型生成问题的装置,其特征在于,包括:

文本抽取模型模块:用于对用于生成问题的文本内容输入到基于模板序列的联合模型中,利用所述联合模型对所述文本进行分类,得到预测文本;

文本关系识别模型模块:用于根据所述预测文本内容获取文本语义特征向量,利用带有关系标签的训练集训练所述文本关系识别模型,其中根据所述文本关系识别模型中设置的字词级和语句级的注意力机制对所述文本语义特征向量进行训练,可获得字向量和词向量对应的输入序列,根据所述对应的输入序列对所述字向量和词向量进行分类,获取预测文本的预测关系标签;

序列生成模型模块:用于所述序列生成模型包括编码单元和模板解码单元,输入问题给所述编码单元,并根据所述问题语义映射成对应的多元组向量,并将所述多元组向量依次输入到模板解码单元中,且由模板解码单元的输出为问题模板,根据所述预测关系标签对所述问题模板进行替换;

其中,构建所述文本抽取模型步骤包括:

所述联合模型为双向长短记忆网络模型和条件随机场模型的联合结构,所述文本抽取模型包括字词向量表示、语句特征提取和语句级的序列标注三层结构,

其中对所述文本数据进行序列标注得到训练集文本,

读取所述训练集文本作为双向长短记忆网络模型的输入进行无监督训练,使得双向长短记忆网络模型初始化所述训练集文本的权值以及构建特征空间;

基于所述特征空间和文本的权值,利用条件随机场模型对所述训练集文本进行有监督学习;

使用归一化函数获得所述训练集文本中各个字词的分类概率;

利用得到的分类概率进行分类得到所述预测文本。

8.一种计算机设备,其特征在于,包括存储器和处理器,所述存储器存储有计算机程序,所述计算机程序被所述处理器执行时,使得所述处理器执行权利要求1-6中任一项所述方法的步骤。

9.一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存储计算机程序,所述计算机程序包括程序命令,所述程序指令被处理器执行时,执行如权利要求1-6任一项所述的方法。

一种基于模板的序列生成模型生成问题的方法及装置

技术领域

[0001] 本发明涉及人工智能自然语言处理技术领域,尤其涉及一种基于模板的序列生成模型生成问题的方法及装置。

背景技术

[0002] 目前,在自然语言处理领域中对中文问题生成的研究很大一部分都是基于模板或者基于规则的模型构建的知识图谱,由于模板和规则的局限性,所生成的问题种类单一且语言缺乏灵活性。

[0003] 基于规则的方法通常需要耗费大量的人力和时间,并且生成的问题通常存在语句不通顺,与文章内容不匹配的问题。基于模板的方法生成的问题比较死板,类型单一,缺乏语言的多样性,而且模板的好坏直接决生成问题的质量。单纯基于序列模型的问题生成方法存在主题实体识别不明确等问题,同样影响生成问题的质量。

发明内容

[0004] 鉴于现有技术存在的上述问题,本发明设计了一种基于模板的序列生成模型生成问题的方法及装置,把基于模板和基于序列生成模型的方法结合在一起,在知识图谱中的序列生成模型中生成相关问题时,提高了生成问题的质量。

[0005] 为实现上述目的,本发明提出了一种基于模板的序列生成模型生成问题的方法,包括:

[0006] 构建文本抽取模型:对用于生成问题的文本内容输入到联合模型中,利用所述联合模型对所述文本进行分类,得到预测文本;

[0007] 构建文本关系识别模型:根据所述预测文本内容获取文本语义特征向量,利用带有关系标签的训练集训练所述文本关系识别模型,其中根据所述文本关系识别模型中设置的字词级和语句级的注意力机制对所述文本语义特征向量进行训练,可获得字向量和词向量对应的输入序列,根据所述对应的输入序列对所述字向量和词向量进行分类,获取预测文本的预测关系标签;

[0008] 构建序列生成模型,所述序列生成模型包括编码单元和模板解码单元,输入问题给所述编码单元,并根据所述问题语义映射成对应的多元组向量,并将所述多元组向量依次输入到模板解码单元中,且由模板解码单元的输出为问题模板,根据所述预测关系标签对所述问题模板进行替换。

[0009] 可选的,所述文本关系识别模型包括输入表示层、字词级层、语句级层以及实体关系分类层,其中将文本语义信息对应的字向量和词向量序列,其中,

[0010] 输入层用于将输入的文字转换成向量的表示,用于获取文本语义特征向量;

[0011] 字词级层学习用于学习所述文本上下文的内容信息,得到每个字词对文本语义信息的重要程度;

[0012] 语句级层学习用于根据上下文语句,给每个输出字词分配不同的权重,获取字词

对语句信息的重要程度；

[0013] 实体关系分类层对语句信息的重要程度进行归一化处理，得到向量的关系标签，从而对实体间的关系进行分类。

[0014] 可选的，构建所述文本抽取模型步骤包括：

[0015] 所述联合模型为双向长短记忆网络模型和条件随机场模型的联合结构，所述文本抽取模型包括字词向量表示、语句特征提取和语句级的序列标注三层结构，

[0016] 其中对所述文本数据进行序列标注得到训练集文本，

[0017] 读取所述训练集文本作为双向长短记忆网络模型的输入进行无监督训练，使得双向长短记忆网络模型初始化所述训练集文本的权值以及构建特征空间；

[0018] 基于所述特征空间和文本的权值，利用条件随机场模型对所述训练集文本进行有监督学习；

[0019] 使用归一化函数获得所述训练集文本中各个字词的分类概率；

[0020] 利用得到的分类概率进行分类得到所述预测文本。

[0021] 可选的，建文本关系识别模型步骤包括：

[0022] 将所述预测文本内容输入到预训练的词袋模型中将文字转换成低维稠密的向量表示，其中所述向量表示为文本语义特征向量，将所述文本语义特征向量输入至构建文本关系识别模型的字词级学习层中，获得所述文本语义特征向量包含的字义信息、词义信息和上下文信息。

[0023] 可选的，构建文本关系识别模型步骤还包括：

[0024] 获得文本语义特征向量后，将所述文本语义特征向量输入至构建文本关系识别模型的语句级学习层中，获得所述文本语义特征向量每个字词的权值，根据加权平均值法得到每个字词的attention值；

[0025] 对所述attention值进行归一化处理，基于所述归一化处理得到所述文本语义特征向量的预测关系标签，根据所述预测关系标签对所述预测文本进行分类得到语句实体。

[0026] 可选的，所述预测关系标签采用HowNet定义的16种标签和5种自定义标签。

[0027] 可选的，所述输入问题映射成对应的多元组向量内容包括问题主题、问题关系、问题，其中所述接收问题是根据所述主题实体和所述实体关系提出，且能由实体回答。

[0028] 本发明实施例提供一种基于模板序列模型生成问题的装置，包括：

[0029] 文本抽取模型模块：用于对用于生成问题的文本内容输入到联合模型中，利用所述联合模型对所述文本进行分类，得到预测文本；

[0030] 文本关系识别模型模块：用于根据所述预测文本内容获取文本语义特征向量，利用带有关系标签的训练集训练所述文本关系识别模型，其中根据所述文本关系识别模型中设置的字词级和语句级的注意力机制对所述文本语义特征向量进行训练，可获得字向量和词向量对应的输入序列，根据所述对应的输入序列对所述字向量和词向量进行分类，获取预测文本的预测关系标签；

[0031] 序列生成模型模块：用于所述序列生成模型包括编码单元和模板解码单元，输入问题给所述编码单元，并根据所述问题语义映射成对应的多元组向量，并将所述多元组向量依次输入到模板解码单元中，且由模板解码单元的输出为问题模板，根据所述预测关系标签对所述问题模板进行替换。

[0032] 本发明实施例提供一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，所述计算机程序被所述处理器执行时，使得所述处理器执行上述中任一项所述方法的步骤

[0033] 与现有技术相比较，本发明提供的一种基于模板与序列相结合模型生成问题的方法，具备以下有益效果：

[0034] 构建文本抽取模型，根据文本抽取模型中设置的双向长短记忆网络模型和条件随机场模型对用于生成问题的文本进行分类，得到预测文本。之后又对预测文本进行关系识别，加入了字词级和字句级的文本关系模型使得该预测文本的分类更为准确，进而获取的预测关系标签更符合文本语义，不存在语句不通顺、与文本内容不匹配的问题。输入问题至序列生成模型中的编码单元，再由生成序列生成模型中的模板解码单元输出为问题模板，可根据文本识别模型中得到的预测关系标签对序列生成模型中的问题模板进行替换，即将输出模板中的主题实体可被任意标签替换，此时生成的问题种类比较灵活、语言类型丰富。

[0035] 应当理解，前面的一般描述和以下详细描述都仅是示例性和说明性的，而不是用于限制本公开。

[0036] 本申请文件提供本公开中描述的技术的各种实现或示例的概述，并不是所公开技术的全部范围或所有特征的全面公开。

附图说明

[0037] 图1是本发明实施例中的实体关系标签示意图；

[0038] 图2是本发明实施例中训练集和测试集合所使用的格式示意图；

[0039] 图3是本发明实施例中基于知识图谱生成的的部分实例示意图；

[0040] 图4是本发明实施中基于知识图谱的问题生成实验结果示意图；

[0041] 图5是本发明实施例中实体抽取功能的整体框架示意图；

[0042] 图6是本发明实施例中实体抽取模型结构的整体结构示意图；

[0043] 图7是本发明实施例中实体关系识别功能框架示意图；

[0044] 图8是本发明实施例中实体关系识别模型结构示意图；

[0045] 图9是本发明实施中基于序列生成模型的结果示意图。

具体实施方式

[0046] 为了使得本公开实施例的目的、技术方案和优点更加清楚，下面将结合本公开实施例的附图，对本公开实施例的技术方案进行清楚、完整地描述。显然，所描述的实施例是本公开的一部分实施例，而不是全部的实施例。基于所描述的本公开的实施例，本领域普通技术人员在无需创造性劳动的前提下所获得的所有其他实施例，都属于本公开保护的范围。

[0047] 除非另外定义，本公开实施例使用的技术术语或者科学术语应当为本公开所属领域内具有一般技能的人士所理解的通常意义。本公开实施例中使用的“包括”或者“包含”等类似的词语意指出现该词前面的元件或者物件涵盖出现在该词后面列举的元件或者物件及其等同，而不排除其他元件或者物件。

[0048] 为了保持本公开实施例的以下说明清楚且简明，本公开实施例省略了已知功能和

已知部件的详细说明。

[0049] 本发明实施例提供了一种基于模板的序列生成模型生成问题的方法,具体包括:

[0050] 构建文本抽取模型:对用于生成问题的文本内容输入到联合模型中,利用联合模型对文本进行分类,得到预测文本。文本抽取模型包括字词向量表示、语句特征提取和语句级的序列标注三层结构。

[0051] 联合模型为双向长短记忆网络模型(Bi-LSTMs)和条件随机场模型的联合结构,基于双向长短记忆网络抽取语句实体和实体上下文信息,使用条件随机场对实体进行分类。使用BIO标注策略对数据集进行序列标注,基于双向长短记忆网络(Bi-LSTMs),将语句中的各个字向量经过一层Bi-LSTMs网络,拼接Bi-LSTMs网络正向和反向的隐含状态序列,使得模型学习到各个词向量的上下文信息,使用条件随机场对Bi-LSTMs网络输出的隐含状态序列进行有监督学习,最后使用归一化函数得到语句中的各个词语的分类概率。

[0052] 在本发明实施例中,参考图5和图6,实体抽取功能的整体框架图分为了实体抽取的训练部分和实体抽取的测试部分将文本集数据进行划分,实体抽取模型结构图中的条件随机场CRF层,“B-PER”、“O”、“B-LOC”是标注策略,将文本数据的80%作为训练集,剩下的20%作为测试集。

[0053] 在本发明实施例中,参考图2,为训练集和测试集所使用的格式,包括问题、对应于问题的三元组、问题的答案,使得最后的文本抽取模型具有典型的文本特征。

[0054] 具体的,给定一个文本,进行数据集构造—序列标注,因为文本是连续的句子,对其进行分词得到每个字和词语,这样可在基于对文本中出现的每个字和出现的词语对文本进行结构化表示。对得到的字、词进行序列标注得到训练集文本,此时不对训练集文本进行训练,直接进行建模,即将其作为双向长短记忆网络的输入进行无监督训练,即执行Bi-LSTMs无监督学习。

[0055] 利用双向长短记忆网络模型初始化训练集文本的权重以及构建特征向量,其中,因为一般的数学模型接收只接收数值输入,所以还要将文本转换成向量表示,向量表示包括为字向量、词向量,即应当理解为一个文本所代表的语义里面用向量表示时包含有数值1和0。训练集文本的权重值应该理解为对用于生成问题的文本内容找到一个关键词,能代表该词对该文本的重要程度,也侧面的反映了该词预测文本主题的能力。此时用空间向量模型,将文本的集合看成空间向量中的多个向量,每个词对应有坐标轴,坐标轴上有对应词的权重值,构建成所述的特征空间。

[0056] 基于特征空间和文本的权重,利用条件随机场模型对训练集文本进行有监督学习,即执行图5中CRF有监督学习。通过带有标签的训练集文本进行训练,得到一个最优模型,利用这个模型将训练集映射为相应的输出,对输出进行判断实现分类,其中训练模型最优选择的是条件随机场模型,基于条件随机场模型对训练集文本进行分词,而其实对于中文分词的模型工具有很多,在这里不再赘述。

[0057] 基于上述步骤,使用归一化函数获得训练集文本中各个字词的分类概率;利用得到的分类概率进行分类得到预测文本。

[0058] 构建文本关系识别模型:根据预测文本内容获取文本语义特征向量,利用带有关系标签的训练集训练所述文本关系识别模型,其中根据文本关系识别模型中设置的字词级和语句级的注意力机制对所述文本语义特征向量进行训练,可获得字向量和词向量对应的

输入序列,根据对应的输入序列对文本进行分类,也可理解为对文本对应的字向量和词向量进行分类。

[0059] 在本发明实施例中,参考图7和图8,实体关系识别模型分为了实体关系识别的训练部分和实体关系识别的测试部分,将文本集数据的80%作为训练集,剩下的20%作为测试集。

[0060] 参考图2,为本发明实施例提供的一种训练集和测试集所用的格式,根据输入的问题设置的三元组内容包括主题实体、实体关系、实体。在其他实施例中,根据提出的问题映射出的多元组包括但不限于二元组、四元组等。

[0061] 具体的,对实体抽取模型输出的预测文本对数据集进行构造,参考图8,该实体关系识别模型中添加了字词级AM的BI-GIU和语句级AM的BI-GIU,除此该模型还包括基础的输入表示层以及实体关系分类层;

[0062] 实体分类关系层中具体包括有隐含层,该隐含层的目的是获取更清晰更好的分类效果。

[0063] 输入层用于将输入的文字转换成向量的表示,用于获取文本语义特征向量;

[0064] 字词级层学习用于学习文本上下文的内容信息,得到每个字词对文本语义信息的重要程度;

[0065] 语句级层学习用于根据上下文语句,给每个输出字词分配不同的权重,获取字词对语句信息的重要程度。

[0066] 具体的,利用Tensorflow框架构建双向门控制单元层与AM的联合结构,利用带有关系标签的训练数据集进行训练,以构建实体关系识别模型。

[0067] 带有关系标签的训练数据集格式参考上述抽取模型的图2提到的格式。

[0068] 构建文本关系识别模型步骤包括:

[0069] 将预测文本输入到预训练的Word2vec模型中将连续文字转换成低维稠密的向量表示,其中向量表示为文本语义特征向量,将文本语义特征向量输入至构建文本关系识别模型的字词级学习层中进行学习,获得所述文本语义特征向量包含的字义信息、词义信息和上下文信息;

[0070] 用空间向量模型,将文本的集合看成空间向量中的多个向量,每个词对应有坐标轴,坐标轴上有对应词的权重值,构建成所述的特征空间,此时,Bi-GRUs结构实现了权值的初始化以及特征空间的构造。

[0071] 构建文本关系识别模型步骤还包括:

[0072] 获得文本语义特征向量时,要理解文本语义,不止需要词向量,也需要获得一些句子级的向量。来判断该句子向量的下一句是当前的句子还是噪声,即通过在文本关系识别模型中设置语句级AM的Bi-GRU层,它的作用是根据上下文语句,给每个输出字词分配不同的权重,获取词语对语句信息的重要程度,通过加权平均值得到每个词语的attention值。

[0073] 对attention值进行归一化处理,得训练集文本中各个字词的分类概率,基于分类概率得到文本语义特征向量的预测关系标签,关系标签参考图1,根据预测关系标签对预测文本进行分类得到主题实体,其中通过调整参数来将预测关系标签与真实标签间的误差最小。

[0074] 以下面文本为例:输入文本:“懂爱的女人通常输得很惨,爱情本来就是残忍的,胜

者为王。感情可以转账,婚姻可以随时冻结,爱情善贾而沽”,通过文本识别模型最后将其分类为“女性”,即该段文本获得的语句实体是“女性”。输入文本为“数据之间的逻辑关系?”,通过文本识别模型后将其分类为“数据”,即该段的语句实体为“数据”。

[0075] 参考图1,预测关系标签采用HowNet定义的16种标签和5种自定义标签,采用的训练集格式为文本1、文本2、关系标签、文本对共同出现的语句。

[0076] 构建序列生成模型,请参考图9,序列生成模型包括编码单元和模板解码单元,编码单元接收问题,并映射成对应的多元组向量,并将所述多元组向量依次输入到模板解码单元中,且由模板解码单元的输出为问题模板,根据所述预测关系标签对所述问题模板进行替换。

[0077] 在本实施例中,所述编码单元为三元组编码器,基于三元组编码器,输入一个三元组F=(主题实体T,实体关系R,实体O),表示为提出一个和主题实体T且具有关系R的问题,并且能由实体O回答,三元组的实例如图2所示。首先使用三元组编码器将三元组并将词语映射到一个实值的向量空间,然后将把向量依次输入到模板解码单元中,模板解码单元的输出为主题实体被任意标签替换的问题模板。

[0078] 利用模板解码单元,将标签替换成特定的主题实体T,从而将问题模板转换成完整问题。

[0079] 请参考图3,实际问题为“逻辑结构的概念是什么”,基于本发明的方法,最后生成的问题是“S的定义是什么?”,语义预测精准,S应该理解为是文本主题,也就是进行分类得到的语句实体,且S具有预测关系标签的数据,即S可以对该数据进行堆排序,形成完全二叉树,可理解为S内容包括但不限于为逻辑结构,保证了生成问题的完整性。

[0080] 为了对比模型效果,本实施案列使用相同的问题生成数据集和实验环境,设计了基于知识图谱的问题生成实验,对比了三种问题生成方法产生问题的正确性。将本案列的使用的模型与基于模板和基于序列生成模型进行对比,使用BLEU测度,METEOR测度,ROUGE测度三个评价标准衡量这三种方法生成问题的优劣,对比实验结果如下图4的表格所示,性能最佳。

[0081] 本发明还提出一种基于模板的序列生成模型生成问题的装置,包括:

[0082] 文本抽取模型模块:用于对用于生成问题的文本内容输入到联合模型中,利用联合模型对所述文本进行分类,得到预测文本;

[0083] 文本关系识别模型模块:用于根据预测文本内容获取文本语义特征向量,利用带有关系标签的训练集训练所述文本关系识别模型,其中根据文本关系识别模型中设置的字词级和语句级的注意力机制对文本语义特征向量进行训练,可获得字向量和词向量对应的输入序列,根据对应的输入序列对所述字向量和词向量进行分类;

[0084] 序列生成模型模块:用于序列生成模型包括编码单元和模板解码单元,编码单元接收问题,并映射成对应的多元组向量,并将多元组向量依次输入到模板解码单元中,且由模板解码单元的输出为问题模板,根据预测关系标签对所述问题模板进行替换。

[0085] 在一个实施例中,提供了一种计算机设备,包括存储器和处理器,存储器存储有计算机程序,计算机程序被处理器执行时,使得处理器执行上述基于模板序列模型生成问题的方法及对应系统的步骤。

[0086] 在一个实施例中,提供了一种计算机可读存储介质,存储有计算机程序,计算机程

序被处理器执行时,使得处理器执行上述一种基于模板序列模型生成问题的方法及对应系统的步骤。

[0087] 上述实施例是对本发明的说明,不是对本发明的限定,任何对本发明简单变换后的方案均属于本发明的保护范围。

关系标签	解释	示例
上下位关系	两个实体是包含和被包含的关系	“国家”和“中国”
同义	两个词语是同义词	无
反义	两个词语是反义词	无
对义	两个词语属于同级概念，不能跨越不同的级别，且在逻辑上相关	无
部件-整体	用于描述两个实体之间的模块构成关系	“手”和“身体”
属性-宿主	用于描述实体之间的是否为属性关系	“年龄”和“人”
材料-成品	用于描述实体之间的组成关系	“面粉”和“馒头”
施事/经验者/关系主体-事件	用于描述两个实体之间的某个动作和动作发出者关系	“医生”和“手术”
受事/内容/领属物等-事件	用于描述两个实体之间的被动关系	“员工”和“被雇佣”
工具-事件关系	用于描述实体之间的事件和实体的关系	“键盘”和“打字”
场所-事件关系	用于描述两个实体之间的位置和时间关系	“公园”和“跑步”
时间-事件关系	用于描述时间和事件的关系时间和时间的关系	“过年”和“发红包”
值-属性关系	用于描述两个属性之间的键和值的关系	“蓝色”和“颜色”
实体-值关系	用于描述两个实体之间统一关系	“首都”和“北京”
事件-角色关系	用于描述场地和实体的关系	“逛街”和“消费者”
相关关系	用于描述两个实体之间的从属关系	“田地”中的“田”和“地”
手段-目的关系	实体1中的操作是为了实现实体2	“学习数据结构是为了程序的模块化”
概念-性质	实体2是概念实体1所具有的区别于其他概念实体的根本属性	“三角形的内角和是180度”
表象-实质	实体1中的现象反映出了实体2	无
对象-方法	实体2是实体1的存储方式	无
事件-对象	实体1是对实体2的操作	“允许删除的一端称为队头”

图1

问题	逻辑结构的概念是什么？
三元组	（逻辑结构，解释关系，数据之间的逻辑关系）
答案	数据之间的逻辑关系

图2

1	三元组：（逻辑结构||解释关系||数据之间的逻辑关系） 答案：数据之间的逻辑关系。 问题：逻辑结构的概念是什么？
2	三元组：（空格串||解释关系||包含空格字符的串） 答案：包含空格字符的串。 问题：什么是空格串？
3	三元组：（数据结构||上下位关系||逻辑结构和物理结构） 答案：逻辑结构和物理结构。 问题：数据结构包括哪些？
4	三元组：（线性结构||解释关系||有序数据元素的集合） 答案：有序数据元素的集合。 问题:线性结构的定义是什么？
5	三元组：（树形结构||解释关系||数据元素之间存在着“一对多”的树形系） 答案：数据元素之间存在着“一对多”的树形关系。 问题：树形结构指什么？
6	三元组：（线性表查找方法||上下位关系||顺序查找、二分查找、分块查找） 答案：顺序查找、二分查找、分块查找。 问题：线性表查找方法有哪些？
7	三元组：（图的遍历方式||上下位关系||深度优先遍历、广度优先遍历） 答案：深度优先遍历、广度优先遍历。 问题：图的遍历方式包括哪些？
8	三元组：（分块查找||同义关系||索引顺序查找） 答案：索引顺序查找。 问题：分块查找的又称为什么？
9	三元组：（数据结构||解释关系||数据之间的组成结构） 答案：数据之间的组成结构。 问题：数据结构是指什么？
10	三元组：（二叉查找树||同义关系||二叉排序树） 答案：二叉排序树。 问题：二叉查找树的别名是什么？

图3

模型	BLEU	METEOR	ROUGE
基于模板	77.23	36.51	37.65
序列生成模型	76.74	35.92	38.71
基于模板的序列生成模型	79.68	37.24	43.85

图4

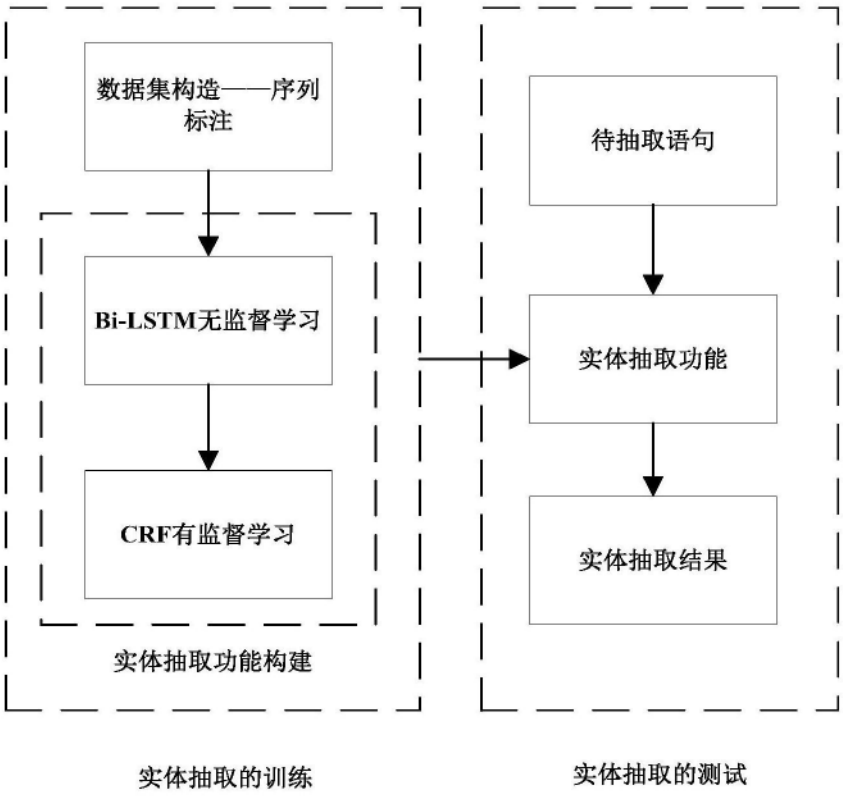


图5

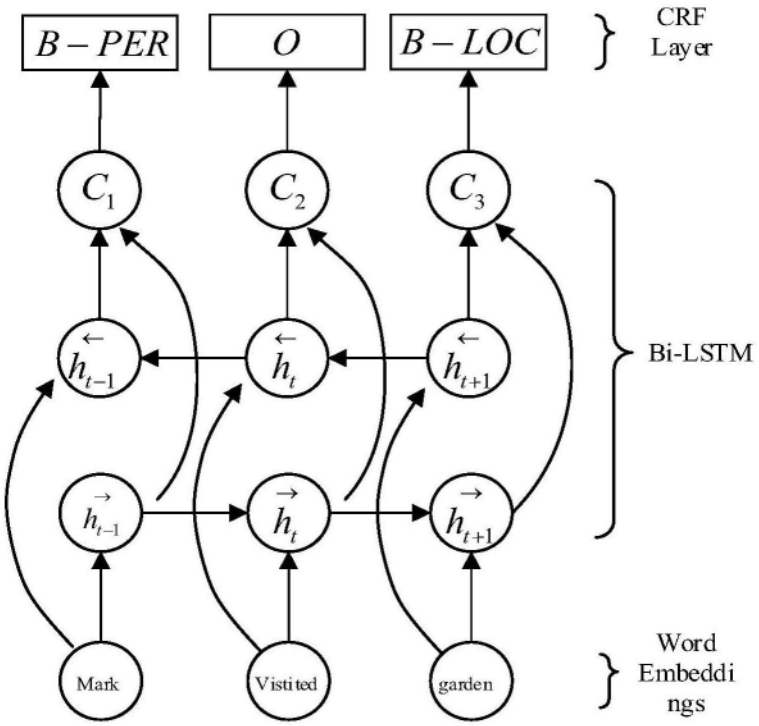

B – PER
O
B – LOC
CRF Layer
C1
C2
C3
Bi-LSTM
Mark
Vistited
garden
Word Embeddings

图6

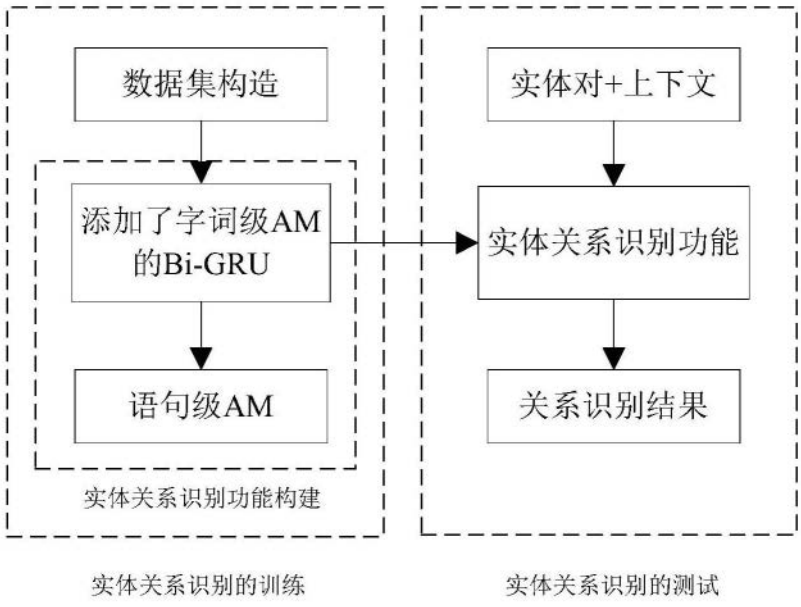

数据集构造
添加了字词级AM的Bi-GRU
语句级AM
实体关系识别功能构建
实体对+上下文
实体关系识别功能
关系识别结果
实体关系识别的训练
实体关系识别的测试

图7

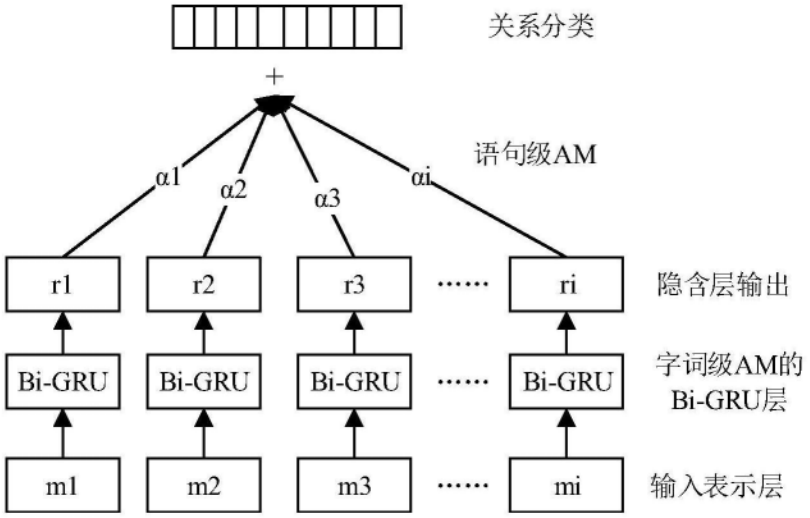


图8

输入三元组：二叉树||解释关系||每个节点最多有两个子树的树结构
答案：每个节点最多有两个子树的树结构
生成问题：二叉树的定义是什么？

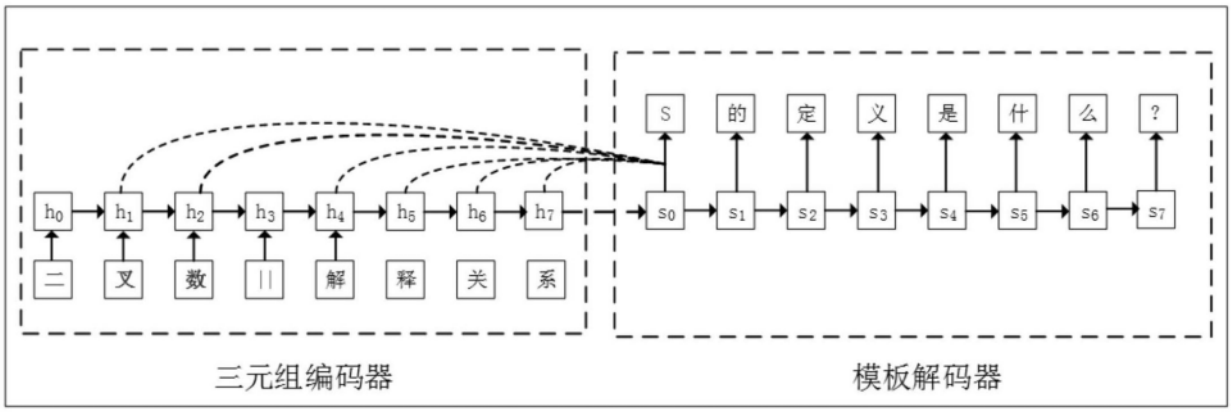


图9