(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 111522965 A
(43)申请公布日 2020.08.11

(21)申请号 202010323470.7

(22)申请日 2020.04.22

(71)申请人 重庆邮电大学
地址 400065 重庆市南岸区南山街道崇文路2号

(72)发明人 韩雨亭 邓蔚 王瑛琦 王国胤 周政

(74)专利代理机构 重庆辉腾律师事务所 50215
代理人 王海军

(51)Int.Cl.
G06F 16/36(2019.01)
G06F 40/295(2020.01)
G06K 9/62(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)
G06F 16/35(2019.01)

权利要求书2页 说明书8页 附图3页

(54)发明名称

一种基于迁移学习的实体关系抽取的问答方法及系统

(57)摘要

本发明涉及自然语言处理技术领域,具体涉及基于迁移学习的实体关系抽取的问答方法,关系分类结果的获得包括:获取源域和目标域文本数据集,预处理;将预处理后的数据输入skip-gram模型训练,得到源域和目标域文本数据的词向量,获取源域和目标域文本数据的位置向量,将位置向量与词向量级联,得到源域目标域文本数据的联合特征向量;将源域文本数据的联合特征向量输入BiLSTM网络预训练,得到预训练过程中的网络参数和源域文本数据的上下文信息、语义特征;将目标域文本数据的联合特征向量输入BiLSTM_CNN融合模型重训练,得到目标域文本数据的高维特征向量并送入分类器,输出关系分类结果。本发明可以提高问答准确率。

CN 111522965 A

1.一种基于迁移学习的实体关系抽取的问答方法,将关系分类结果链接至知识图谱中,根据知识图谱的关系页面实时查询输入实体词之间的关系信息,输出答案,其特征在于,关系分类结果的获得包括:

S1、获取源域文本数据集和目标域文本数据集,所述源域文本数据集和所述目标域文本数据集中包括至少一个句子,每一个句子中至少包括一个实体,对所述源域文本数据集和所述目标域文本数据集中每个句子中的每个实体进行识别和标注;

S2、将预处理后的源域文本数据集和目标域文本数据集输入skip-gram模型进行训练,分别得到源域文本数据的词向量和目标域文本数据的词向量;

S3、分别获取源域文本数据的位置向量和目标域文本数据的位置向量,将源域文本数据的词向量和位置向量进行拼接,得到源域文本数据的联合特征向量;将目标域文本数据的词向量和位置向量进行拼接,得到目标域文本数据的联合特征向量;

S4、将源域文本数据的联合特征向量输入BiLSTM网络进行预训练,得到源域文本数据的上下文信息、源域文本数据的语义特征和预训练过程中的网络参数;

S5、将目标域文本数据的联合特征向量输入BiLSTM_CNN融合模型中,根据预训练过程中的网络参数对目标域文本数据进行重训练,得到目标域文本数据的高维特征向量;

S6、将目标域文本数据的高维特征向量送入分类器,得到关系分类结果。

2.根据权利要求1所述的一种基于迁移学习的实体关系抽取的问答方法,其特征在于,所述skip-gram模型是word2vec工具中的一个模型,用于词向量的训练,将预处理后的源域文本数据和目标域文本数据分别输入到skip-gram模型中,训练时设定词向量维度为100维度,训练结束后,分别得到源域文本数据的词向量映射表和目标域文本数据的词向量映射表,词向量映射表中包含了词与向量的映射关系,根据词向量映射表得到每个词对应的词向量。

3.根据权利要求1所述的一种基于迁移学习的实体关系抽取的问答方法,其特征在于,源域文本数据的位置向量和目标域文本数据的位置向量的获取包括:以实体词在句子中的位置为原点,一个单词相对于一个实体词来说,从左至右表示矢量正向,从右至左表示矢量负向,即一个单词在实体词右侧用正数表示,单词在实体左侧用负数表示,所述实体为一个句子中的名词。

4.根据权利要求1所述的一种基于迁移学习的实体关系抽取的问答方法,其特征在于,源域文本数据的联合特征向量输入BiLSTM网络进行预训练包括:计算三个门单元:输入门i,输出门o和遗忘门f;通过三个门单元计算记忆单元;然后通过记忆单元计算并输出源域文本数据的上下文信息、源域文本数据的语义特征,保留预训练过程的网络参数。

5.根据权利要求4所述的一种基于迁移学习的实体关系抽取的问答方法,其特征在于,所述三个门单元的计算方式包括:

$i_t=\sigma(W_i v_t+U_i h_{t-1}+b_i)$

$f_t=\sigma(W_f v_t+U_f h_{t-1}+b_f)$

$o_t=\sigma(W_o v_t+U_o h_{t-1}+b_o)$

其中,i_t表示t时刻LSTM单元中的输入门的值,f_t表示t时刻LSTM单元中的遗忘门,o_t表示t时刻LSTM单元中的遗忘门,σ表示sigmod激活函数,Wi,Wf,Wo分别为LSTM网络中输入门、遗忘门、输出门的权值矩阵,v_t表示当前输入,U_i、U_f、U_o分别表示LSTM网络中输入门、遗忘

门、输出门的h_{t-1}对应的权重，h_{t-1}表示上一时刻的隐层状态向量，t表示当前时刻，b_i、b_f、b_o代表LSTM网络中输入门、遗忘门、输出门的偏置向量。

6.根据权利要求1所述的一种基于迁移学习的实体关系抽取的问答方法，其特征在于，利用BiLSTM_CNN融合模型对目标域文本数据进行重训练的过程包括：将目标域文本数据的联合特征向量输入BiLSMT网络结构，根据BiLSMT 网络结构，计算三个门单元；根据三个门单元计算t时刻记忆单元；根据t时刻记忆单元计算t时刻的隐层状态向量，得到经过BiLSTM网络提取的时序性特征；将经过BiLSTM网络提取的时序性特征输入CNN网络中进行训练，通过卷积池化的操作进行局部特征提取，最终得到目标域文本数据的高维特征向量。

7.根据权利要求1所述的一种基于迁移学习的实体关系抽取的问答方法，其特征在于，所述分类器选用softmax分类器，分类器最终输出结果的计算公式为：

$$p(y|S)=\mathrm{softmax}(W_cC+b_c)$$

其中，$p(y|S)$表示样本分布中的最大概率值，y表示正确分类样本，S表示样本数量，W_c表示卷积层输出的权重，C表示卷积层输出，b_c表示卷积层输出的偏置向量。

8.一种基于迁移学习的实体关系抽取的问答系统，其特征在于，所述系统包括：数据预处理模块、源域文本数据预训练模块、权重迁移模块、特征提取模块和分类模块，

所述数据预处理模块用于对文本信息进行数据预处理工作，包括数据清洗、实体查找、实体标注；

所述源域文本数据预训练模块用于预训练源域文本数据，并保留预训练源域文本数据中的网络参数；

所述权重迁移模块用于迁移预训练源域文本数据过程中的网络参数；

所述特征提取模块用于提取目标域文本数据的高维特征；

所述分类模块用于获得关系分类结果。

9.根据权利要求8所述一种基于迁移学习的实体关系抽取的问答系统，其特征在于，分类模块采用以下方式实现：

S1、获取源域文本数据集和目标域文本数据集，所述源域文本数据集和所述目标域文本数据集中包括至少一个句子，每一个句子中至少包括一个实体，对所述源域文本数据集和所述目标域文本数据集中每个句子中的每个实体进行识别和标注；

S2、将预处理后的源域文本数据集和目标域文本数据集输入skip-gram模型进行训练，分别得到源域文本数据的词向量和目标域文本数据的词向量；

S3、分别获取源域文本数据的位置向量和目标域文本数据的位置向量，将源域文本数据的词向量和位置向量进行拼接，得到源域文本数据的联合特征向量；将目标域文本数据的词向量和位置向量进行拼接，得到目标域文本数据的联合特征向量；

S4、将源域文本数据的联合特征向量输入BiLSTM网络进行预训练，得到源域文本数据的上下文信息、源域文本数据的语义特征和预训练过程中的网络参数；

S5、将目标域文本数据的联合特征向量输入BiLSTM_CNN融合模型中，根据预训练过程中的网络参数对目标域文本数据进行重训练，得到目标域文本数据的高维特征向量；

S6、将目标域文本数据的高维特征向量送入分类器，得到关系分类结果。

一种基于迁移学习的实体关系抽取的问答方法及系统

技术领域

[0001] 本发明涉及信息技术领域中的自然语言处理技术领域，具体涉及一种基于迁移学习的实体关系抽取的问答方法及系统。

背景技术

[0002] 在互联网技术的不断发展和推动下，网络数据内容及碎片化信息正在呈现爆发式增长的态势。知识图谱作为人工智能技术的的重要分支，利用其强大的语义处理能力和开放互联能力将信息和知识有序，有机地进行组织，构建大规模语义网络，为互联网时代的知识获取和信息处理提供了便捷。关系抽取作为知识图谱构建的子任务，从细粒度的无结构化文本信息中挖掘句子的语义关系信息，形成结构化知识，并将其结果服务于构建知识图谱及本体知识库，为知识获取和其他智能应用提供帮助，因此关系抽取任务在基于知识图谱的问答和搜索的应用场景中具有重要意义。目前关系抽取任务分为有监督关系抽取，半监督关系抽取和无监督关系抽取。

[0003] 有监督关系抽取方法中，基于规则和模板匹配的方法需要通过人工和机器学习总结出规则和模板，费时费力;基于特征向量的方法无法充分利用上下文结构信息。远程监督关系抽取方法，通过知识库自动获取训练数据从而完成数据标注任务。但由于自动标注过程引入大量噪声文本，需额外解决数据噪声问题。由于先实体识别后关系抽取的流水线方法会造成错误传播，同时产生了冗余信息。现阶段无监督的关系抽取方法抽取效果也没有达到理想结果。

[0004] 中国专利CN107832400A提出了一种基于位置LSTM和CNN联合模型进行进行关系抽取的方法，通过借助联合模型解决了关系抽取模型特征提取不充分的问题，故而可以提高关系抽取准确率。该专利结合了两种模型提取特征的优势并进行组合，从而完成关系抽取任务。

[0005] 但在关系抽取任务中，先实体识别后关系抽取的流水线方法会导致错误传播的问题，即实体识别的准确率直接影响关系抽取的效果。中国专利CN110781683A提出了一种实体关系联合抽取方法，利用联合抽取模型提高三元组抽取的准确率。该方法很好地避免了流水线方法错误传播的问题，同时提高了关系抽取效率。

[0006] 现有技术的关系抽取方法在领域样本数量较少的情况下，关系抽取准确率会大大降低，只能通过人工构建数据集，或通过远程监督标注数据的方法扩充数据样本，然而人工标注和构建的过程费时费力，消耗大量人力成本，远程监督标注的数据会产生大量噪声，从而降低关系分类结果的准确性，极大降低线上输入问题的答案准确性。

发明内容

[0007] 为了解决上述现有技术在目标领域样本数量较少条件下无法在训练模型中得到理想的学习效果，从而导致的关系抽取结果不准确的问题，本发明提出了一种基于BiLSTM_CNN融合网络和迁移学习的关系抽取方法，该方法首先利用数据量较大且与目标域文本数

据相似度较高的源域文本数据进行预训练，借助迁移学习方法将预训练得到的参数进行重训练，通过这种权重迁移的方式帮助目标域少样本数据完成关系抽取任务，提升关系抽取的效率和准确度。

[0008] 一种基于迁移学习的实体关系抽取的问答方法，将关系分类结果链接至知识图谱中，根据知识图谱的关系页面实时查询输入实体词之间的关系信息，输出答案，关系分类结果的获得包括以下步骤：

[0009] S1、获取源域文本数据集和目标域文本数据集，所述源域文本数据集和所述目标域文本数据集中包括至少一个句子，每一个句子中至少包括一个实体，对所述源域文本数据集和所述目标域文本数据集中每个句子中的每个实体进行识别和标注；

[0010] S2、将预处理后的源域文本数据集和目标域文本数据集输入skip-gram模型进行训练，分别得到源域文本数据的词向量和目标域文本数据的词向量；

[0011] S3、分别获取源域文本数据的位置向量和目标域文本数据的位置向量，将源域文本数据的词向量和位置向量进行拼接，得到源域文本数据的联合特征向量；将目标域文本数据的词向量和位置向量进行拼接，得到目标域文本数据的联合特征向量；

[0012] S4、将源域文本数据的联合特征向量输入BiLSTM网络进行预训练，得到源域文本数据的上下文信息、源域文本数据的语义特征和预训练过程中的网络参数；

[0013] S5、将目标域文本数据的联合特征向量输入BiLSTM_CNN融合模型中，根据预训练过程中的网络参数对目标域文本数据进行重训练，得到目标域文本数据的高维特征向量；

[0014] S6、将目标域文本数据的高维特征向量送入分类器，输出关系分类结果。

[0015] 进一步的，所述skip-gram模型是word2vec工具中的一个模型，用于词向量的训练，将预处理后的源域文本数据和目标域文本数据分别输入到skip-gram模型中，训练时设定词向量维度为100维度，训练结束后，分别得到源域文本数据的词向量映射表和目标域文本数据的词向量映射表，词向量映射表中包含了词与向量的映射关系，根据词向量映射表得到每个词对应的词向量。

[0016] 进一步的，源域文本数据的位置向量和目标域文本数据的位置向量的获取包括：以实体词在句子中的位置为原点，一个单词相对于一个实体词来说，从左至右表示矢量正向，从右至左表示矢量负向，即一个单词在实体词右侧用正数表示，单词在实体左侧用负数表示，所述实体为一个句子中的名词。

[0017] 进一步的，源域文本数据的联合特征向量输入BiLSTM网络进行预训练包括：计算三个门单元：输入门i，输出门o和遗忘门f；通过三个门单元计算记忆单元；然后通过记忆单元计算并输出源域文本数据的上下文信息、源域文本数据的语义特征，保留预训练过程的网络参数。

[0018] 进一步的，所述三个门单元的计算方式包括：

[0019] $i_t=\sigma(W_i v_t+U_i h_{t-1}+b_i)$

[0020] $f_t=\sigma(W_f v_t+U_f h_{t-1}+b_f)$

[0021] $o_t=\sigma(W_o v_t+U_o h_{t-1}+b_o)$

[0022] 其中，i_t表示t时刻LSTM单元中的输入门的值，f_t表示t时刻LSTM单元中的遗忘门，o_t表示t时刻LSTM单元中的遗忘门，σ表示sigmod激活函数，Wi，Wf，Wo分别为LSTM网络中输入门、遗忘门、输出门的权值矩阵，v_t表示当前输入，U_i、U_f、U_o分别表示LSTM网络中输入门、

遗忘门、输出门的h_{t-1}对应的权重，h_{t-1}表示上一时刻的隐层状态向量，t表示当前时刻，b_i、b_f、b_o代表LSTM网络中输入门、遗忘门、输出门的偏置向量。

[0023] 进一步的，利用BiLSTM_CNN融合模型对目标域文本数据进行重训练的过程包括：将目标域文本数据的联合特征向量输入BiLSMT网络结构，根据BiLSMT网络结构，计算三个门单元；根据三个门单元计算t时刻记忆单元；根据t时刻记忆单元计算t时刻的隐层状态向量，得到经过BiLSTM网络提取的时序性特征；将经过BiLSTM网络提取的时序性特征输入CNN网络中进行训练，通过卷积池化的操作进行局部特征提取，最终得到目标域文本数据的高维特征向量。

[0024] 进一步的，所述分类器选用softmax分类器，分类器的计算公式包括：

[0025] $p(y|S)=\mathrm{softmax}(W_cC+b_c)$

[0026] 其中，$p(y|S)$表示样本分布中的最大概率值，y表示正确分类样本，S表示样本数量，W_c表示卷积层输出的权重，C表示卷积层输出，b_c表示卷积层输出的偏置向量。

[0027] 一种基于迁移学习的实体关系抽取的问答系统，所述系统包括：数据预处理模块、源域文本数据预训练模块、权重迁移模块、特征提取模块和分类模块，所述数据预处理模块用于对文本信息进行数据预处理工作，包括数据清洗；所述源域文本数据预训练模块用于预训练源域文本数据，保留网络参数；所述权重迁移模块用于迁移预训练源域文本数据过程中的网络参数；所述特征提取模块用于提取目标域文本数据的高维特征；所述分类模块用于获得关系分类结果。

[0028] 进一步的，分类模块采用以下方式实现：

[0029] S1、获取源域文本数据集和目标域文本数据集，所述源域文本数据集和所述目标域文本数据集中包括至少一个句子，每一个句子中至少包括一个实体，对所述源域文本数据集和所述目标域文本数据集中每个句子中的每个实体进行识别和标注；

[0030] S2、将预处理后的源域文本数据集和目标域文本数据集输入skip-gram模型进行训练，分别得到源域文本数据的词向量和目标域文本数据的词向量；

[0031] S3、分别获取源域文本数据的位置向量和目标域文本数据的位置向量，将源域文本数据的词向量和位置向量进行拼接，得到源域文本数据的联合特征向量；将目标域文本数据的词向量和位置向量进行拼接，得到目标域文本数据的联合特征向量；

[0032] S4、将源域文本数据的联合特征向量输入BiLSTM网络进行预训练，得到源域文本数据的上下文信息、源域文本数据的语义特征和预训练过程中的网络参数；

[0033] S5、将目标域文本数据的联合特征向量输入BiLSTM_CNN融合模型中，根据预训练过程中的网络参数对目标域文本数据进行重训练，得到目标域文本数据的高维特征向量；

[0034] S6、将目标域文本数据的高维特征向量送入分类器，得到关系分类结果。

[0035] 本发明的有益效果：

[0036] 1.本发明在融合模型的基础上，利用迁移学习方法将源域文本数据预训练过程的网络参数迁移至目标域文本数据训练模型中进行重训练，所述源域文本数据数据量大且与目标领域数据共享某些语义信息，相似度较高，可以利用源域文本数据训练的结果帮助目标域文本数据达到更加理想的训练结果，从而提高目标域文本数据关系抽取的准确率和效率。即借助迁移的外部知识，帮助完成目标域少样本领域关系抽取任务，提高关系抽取准确率和效率，提高在线问答的准确率。

[0037] 2.本发明在利用迁移学习方法时,不同于一般情况预训练和重训练过程在相同模型中完成,为了提高关系抽取效率和模型鲁棒性,预训练过程采用BiLSTM模型结构进行训练,重训练过程采用BiLSTM_CNN的融合模型进行训练。

附图说明

[0038] 下面结合附图和具体实施方式对本发明做进一步详细的说明。

[0039] 图1为本发明实施例提供的一种基于迁移学习的实体关系抽取的问答方法流程图;

[0040] 图2为本发明实施例提供的BiLSTM模型结构示意图;

[0041] 图3为本发明实施例提供的一种基于迁移学习的实体关系抽取的问答方法示意图;

[0042] 图4是本发明实施例提供的一种基于BiLSTM_CNN融合网络和迁移学习的关系抽取的问答系统示意图。

具体实施方式

[0043] 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,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,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,都属于本发明保护的范围。

[0044] 如图1所示为本发明一种基于迁移学习的实体关系抽取的问答方法的流程图,该方法可以解决现有技术中在目标领域样本数量较少条件下关系抽取结果不准确导致的输入问题的答案准确性较低的问题,该方法包括但不限于以下步骤:

[0045] 获取源域文本数据集和目标域文本数据集,进行数据预处理。所述目标域文本数据集中的文本数据量较少,所述源域文本数据集中的文本数据量较多。所述源域文本数据集和所述目标域文本数据集中包括至少一个句子,每一个句子中至少包括一个实体。所述源域文本数据集从公开数据集中获取,本专利可选取New York Times关系抽取数据集作为源域文本数据集,所述源域文本数据集中共包含有18,252种关系类型和522,611条句子。所述目标域文本数据可选取地理领域数据,地理领域样本数据比较稀缺,从wikipedia和互联网网页信息中通过爬虫的方式获取。源域文本数据的数据量远大于目标域文本数据的数据量,源域文本数据和目标域文本数据中中均为无结构化文本信息,但源域文本数据和目标域文本数据中结构化文本的关系类型不同,源域文本数据中涉及18,252种关系类型的数据,针对目标域数据涉及9个类别的关系。由于目标域文本数据的数据量较少,通过目标数据直接进行训练进行关系抽取的效果不可理想,而源域文本数据的数据量远大于目标域文本数据的数据量,且源域文本数据和目标域文本数据共享部分语义信息,与目标域文本数据具有较大的相似性,因而可以借助数据量大且相似较高的源域文本数据帮助完成目标域文本数据关系抽取任务。

[0046] 由于公开数据集中的原始数据不符合模型输入要求,因此先对源域文本数据和目标域文本数据进行预处理,所述预处理主要包括:针对源域文本数据,进行数据清洗,去掉无意义字符和格式。针对目标域文本数据,将从wikipedia和网页中通过爬虫技术爬取的文

本去掉与地理概念描述无关的文本信息。然后将所得段落内容切分成句,对得到的每一条句子通过stanfordCoreNLP工具包完成实体识别及标注。

[0047] 方便完成实体的关系抽取任务。所述实体为一个语句中的名词,例如,语句为:“Steve Jobs was the co-founder of the Apple Inc.”,该语句中有两个实体:Steve Jobs、AppleInc。

[0048] 可选的,所述源域文本数据来源于公共数据库中的New York Times关系抽取数据集,所述目标域文本数据来源于利用爬虫技术爬取的公开互联网数据。

[0049] 将预处理后的源域文本数据集和目标域文本数据集输入word2vec工具中的skip-gram模型进行训练,将文本数据转化为数学数据,分别得到源域文本数据的词向量和目标域文本数据的词向量。所述源域文本数据的词向量包括源域文本数据集中每一个词相对应的词向量,所述目标域文本数据的词向量包括目标域文本数据集中每一个词相对应的词向量。将预处理后的源域文本数据和目标域文本数据分别输入word2vec工具中进行词向量的训练,训练时选用word2vec工具中的skip-gram模型,设定词向量维度为100维度,训练结束后,分别得到源域文本数据的词向量映射表和目标域文本数据的词向量映射表,词向量映射表中包含了词与向量的映射关系。根据词向量映射表得到每个词对应的词向量。例如,在句子“The white cat is eating the fish.”中,经过词向量映射,得到“cat”对应的向量表示为[0.712,-0.05,0.152,.....],通过这样的表示,将文本信息转化为计算机可理解的数值信息。

[0050] 分别获取源域文本数据的位置向量和目标域文本数据的位置向量,将源域文本数据的词向量和位置向量进行拼接,得到源域文本数据的联合特征向量;将目标域文本数据的词向量和位置向量进行拼接,得到目标域文本数据的联合特征向量。根据每个句子中实体的位置,计算句子中每个词相对于两个实体的位置向量,一个句子中一个词相对于两个实体的位置向量确定具体包括:以实体词在句子中的位置为原点,一个单词相对于同一句子中的一个实体词来说,从左至右表示矢量正向,从右至左表示矢量负向,即一个单词在所述实体词右侧用正数表示,单词在所述实体左侧用负数表示;一个单词相对于同一句子中的另一个实体词来说,从左至右表示矢量正向,从右至左表示矢量负向。每个词与一个句子中每个实体的相对位置组成该词的位置向量,每个词的位置向量维数由一个句子中的实体词的个数决定,一个句子中若有n个实体词,则每个词的位置向量维数为n维。例如,在句子“Mental[illness]e_1 is one of the biggest causes of personal[unhappiness]e_2 in our society”中,单词“causes”到头实体e_1“illness”和尾实体e_2“unhappiness”的矢量距离分别为6和-3,即单词“causes”的位置向量为(6,-3)。获取每一个词的位置向量后,将每一个词的词向量和该词的位置向量级联,得到每个词的联合特征向量。

[0051] 将源域文本数据的联合特征向量$S=\{v_1,v_2,\cdots,v_t\}$输入BiLSTM网络进行预训练,得到源域文本数据的上下文信息、源域文本数据的语义特征和预训练过程中的网络参数。BiLSTM网络结构如图2所示。源域文本数据预训练过程如图3a)pre-train所示。

[0052] 所述预训练包括如下过程:根据LSTM网络结构,计算三个门单元:输入门i,输出门o和遗忘门f,通过三个门单元计算记忆单元,然后通过记忆单元计算并输出源域文本数据对应的高维特征向量,所述源域文本数据对应的高维特征向量中包括源域文本数据的上下文信息、源域文本数据的语义特征,在预训练过程中保留预训练过程的网络参数,以便后续

利用源域文本数据预训练的网络参数，在目标域文本数据中进行重训练。

[0053] 进一步的，所述三个门单元(输入门i，输出门o和遗忘门f)的计算方式包括：设当前时刻为t，上一时刻的隐层状态向量为h_{t-1}，当前输入为v_t，初始隐层状态向量h_0为0，利用公式(1)计算t时刻LSTM单元中的输入门的值i_t，利用公式(2)计算t时刻LSTM单元中的遗忘门f_t，利用公式(3)计算t时刻LSTM单元中的遗忘门o_t，其计算公式如下：

[0054] $$i_t=\sigma(W_i v_t+U_i h_{t-1}+b_i) \quad (1)$$

[0055] $$f_t=\sigma(W_f v_t+U_f h_{t-1}+b_f) \quad (2)$$

[0056] $$o_t=\sigma(W_o v_t+U_o h_{t-1}+b_o) \quad (3)$$

[0057] 其中，σ表示sigmod激活函数，Wi，Wf，Wo分别为LSTM网络中输入门、遗忘门、输出门的权值矩阵，U_i、U_f、U_o分别表示LSTM网络中输入门、遗忘门、输出门的h_{t-1}对应的权重，b_i，b_f，b_o代表LSTM网络中输入门、遗忘门、输出门的偏置向量。

[0058] 当前时刻的特征向量g_t依赖于前一时刻隐层状态向量h_{t-1}和输入v_t，其计算公式如下：

[0059] $$g_t=\tanh(W_g v_t+U_g h_{t-1}+b_g) \quad (4)$$

[0060] 其中，W_g为L当前时刻特征向量对应的权值矩阵，U_g表示特征向量前一时刻h_{t-1}对应的权重，b_g为求特征向量时当前时刻对应的偏置向量，tanh表示双曲正切函数用作激活函数。

[0061] 设前一时刻记忆单元为c_{t-1}，已知当前时刻特征为g_t，初始记忆单元c_0为0。然后计算t时刻记忆单元c_t公式为：

[0062] $$c_t=i_t\odot g_t+f_t\odot c_{t-1} \quad (5)$$

[0063] 最终得到隐层状态向量h_t。

[0064] $$h_t=o_t\odot\tanh(c_t) \quad (6)$$

[0065] 其中，i_t表示LSTM网络结构输入们结果，f_t表示遗忘门结果，c_{t-1}表示t-1时刻记忆单元，o_t表示t时刻LSTM单元中的遗忘门的计算结果。

[0066] 对于输入序列，为了使LSTM结构包含t时刻前后的信息，采用双向LSTM网络结构分别得到正反向的序列最终将两个序列相加得到BiLSTM网络的输出H。最终保留源域文本数据经过预训练的网络参数。

[0067] 将目标域文本数据的联合特征向量输入BiLSTM_CNN融合模型中，根据预训练过程中的网络参数对目标域文本数据进行重训练，得到目标域文本数据的高维特征向量。重训练过程如图3b)Fine-tuning所示。

[0068] 为了避免BiLSTM_CNN融合模型中参数随机初始化，造成的训练时间长，效率低的问题，本发明采用了迁移学习方法中的网络参数迁移，即将预训练过程保留的网络参数全部更新至目标域文本数据对应的BiLSTM_CNN融合模型中。源域文本数据预训练过程产生的参数类型包括词向量表示对应的参数和位置向量表示对应的参数，本发明将预训练过程的所有参数进行迁移。经过网络参数迁移后，目标领域数据得到相似的源域文本数据对应的语义特征。利用预训练过程的网络参数在BiLSTM_CNN融合模型重训练目标数据，同时借助预训练过程的网络参数所对应的语义特征，帮助目标域文本数据完成关系抽取任务，有利于提高训练的效率，增强训练的效果，可以更加精确高效的进行目标域文本数据中的关系抽取。

[0069] 进一步的，利用BiLSTM_CNN融合模型对目标域文本数据进行重训练的过程包括：在BiLSTM网络中进行训练提取时序性特征。训练过程如下：

[0070] 首先将目标域文本数据的联合特征向量输入BiLSMT网络结构，根据BiLSMT网络结构，计算三个门单元：输入门i，输出门o和遗忘门f，计算过程与预训练中LSTM网络结构的计算方式一致。

[0071] 然后计算t时刻记忆单元c_t，计算公式与预训练中BiLSTM网络结构的计算方式一致。

[0072] 最后根据上述结果计算出t时刻的隐层状态向量，计算公式同预训练过程BiLSTM网络结构的计算方式一致，得到经过BiLSTM网络提取的时序性特征。

[0073] 将经过BiLSTM网络提取的时序性特征输入CNN网络中进行训练，通过卷积池化的操作进行局部特征提取，最终得到目标域文本数据的高维特征向量。

[0074] 具体过程如下：

[0075] $$\overleftrightarrow{h} = \overrightarrow{h} + \overleftarrow{h} \qquad (7)$$

[0076] 其中，$\overleftrightarrow{h}$表示双向隐层状态向量，$\overrightarrow{h}$表示正向隐层状态向量，$\overleftarrow{h}$表示负向隐层状态向量。

[0077] 从BiLSTM层可以得到$\mathbf{h} = \{\overrightarrow{h_1}, \overrightarrow{h_2}, \ldots, \overrightarrow{h_n}\}$，其中，h表示隐层状态向量集合，$\overrightarrow{h_1}$表示经过BiLSTM网络得到的第一个隐层状态向量，$\overrightarrow{h_n}$表示第n个隐层状态向量，n表示隐层的数量，为了更好的提取局部特征，借助CNN卷积的方式获取更高级别的特征，即高维特征向量，设卷积核大小w为3，且$c \in R^{s+w-1}$，卷积公式如下：

[0078] $c_j = f(wh_{j-w+1:j}+b)$ (8)

[0079] 其中，c表示卷积结果，R^{s+w-1}表示矩阵，s表示句子长度，w表示窗口大小；c_j表示卷积层的输出，f表示卷积函数，w表示需要学习的权重，$h_{j-w+1:j}$表示卷积长度从j-w+1到j，j表示卷积长度，b表示偏置向量。

[0080] 最终目标域文本数据经过融合模型训练后输出高维特征向量。

[0081] 将目标域文本数据的高维特征向量输入分类器，分类器输出关系分类结果。优选的，所述分类器为softmax分类器，分类器的计算公式最终得到关系分类结果如下：

[0082] $p(y|S) = \mathrm{softmax}(W_cC+b_c)$ (9)

[0083] 其中，$p(y|S)$表示样本分布中的最大概率值，y表示正确分类样本，S表示样本数量，W_c表示卷积层输出的权重，C表示卷积层输出，b_c表示卷积层输出的偏置向量。

[0084] 最后，将上述关系分类结果链接至知识图谱中，用户输入实体词，根据知识图谱的关系页面实时查询输入实体词之间的关系信息，获得问题的答案。

[0085] 本发明将关系分类结果链接至知识图谱中为本领域常用技术手段，可以采用现有技术实现，不再详细。

[0086] 在一个实施例中，本发明还提供了一种基于迁移学习的实体关系抽取的问答系统，如图4所示，所述系统包括：数据预处理模块，源域文本数据预训练模块，权重迁移模块，特征提取模块和分类模块。

[0087] 所述数据预处理模块用于对文本信息进行数据预处理工作，包括数据清洗、实体

查找、实体标注等；

[0088] 所述源域文本数据预训练模块用于预训练源域文本数据，保留网络参数，以便后续进行权重迁移；

[0089] 所述权重迁移模块用于迁移预训练过程的模型参数，通过外部知识的迁移帮助提高目标数据分类准确率；

[0090] 所述特征提取模块用于提取目标域文本数据的高维特征；

[0091] 所述分类模块用于获得关系分类结果。

[0092] 上述系统各个模块的实现方式可以采用上述方法的实施例。

[0093] 具体的，在预处理过程中对源域文本数据和目标域文本数据均进行数据清洗等工作，然后将完成预处理的源域文本数据送入预训练模块，并保留云训练源于数据过程中的网络参数。将预训练过程产生的网络参数送入权重迁移模块，将参数迁移至目标域对应的模型中。通过权重迁移模块中迁移的参数信息，在特征提取模块中完成对目标域文本数据的特征提取，然后将输出的特征送入分类器中输出关系分类结果。

[0094] 所述基于融合网络和迁移学习的关系抽取系统的创新性在于，将迁移学习完整过程划分为预训练模块，权重迁移模块和特征提取模块。不同于一般的权重迁移方式中源域文本数据和目标域文本数据共享相同的网络结构，所述模块将源域文本数据预训练过程和目标域文本数据特征提取过程分别在不同的网络结构中完成，源域文本数据预训练网络结构使用了部分特征提取模块中的网络结构信息，用以完成参数迁移。这样做的好处在于，预训练过程保留尽可能粗粒度的特征信息，在参数迁移之后再针对目标域文本数据重训练提取细粒度的高维特征，从而针对目标域文本数据提取更高质量的特征信息。

[0095] 尽管已经示出和描述了本发明的实施例，对于本领域的普通技术人员而言，可以理解在不脱离本发明的原理和精神的情况下可以对这些实施例进行多种变化、修改、替换和变型，本发明的范围由所附权利要求及其等同物限定。


获取源域数据集和目标域数据集，进行数据预处理
将预处理后的源域数据集和目标域数据集输入skip-gram模型进行训练，分别得到源域数据的词向量和目标域数据的词向量
分别获取源域数据的位置向量和目标域数据的位置向量，将源域数据的词向量和位置向量进行拼接，得到源域数据的联合特征向量；将目标域数据的词向量和位置向量进行拼接，得到目标域数据的联合特征向量
将源域数据的联合特征向量输入BiLSTM网络进行预训练，得到源域数据的上下文信息、源域数据的语义特征和预训练过程中的网络参数
将目标域数据的联合特征向量输入BiLSTM_CNN融合模型中，根据预训练过程中的网络参数对目标域数据进行重训练，得到目标域数据的高维特征向量
将目标域数据的高维特征向量送入分类器，得到关系分类结果

图1

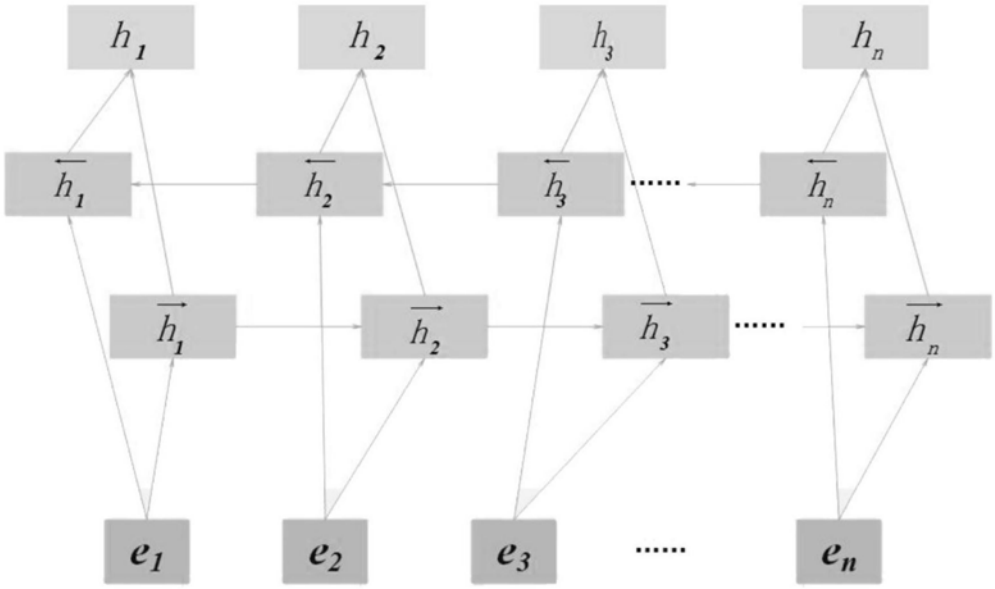
h_1
h_2
h_3
h_n
$\overleftarrow{h_1}$
$\overleftarrow{h_2}$
$\overleftarrow{h_3}$
$\overleftarrow{h_n}$
$\overrightarrow{h_1}$
$\overrightarrow{h_2}$
$\overrightarrow{h_3}$
$\overrightarrow{h_n}$
e_1
e_2
e_3
e_n

图2

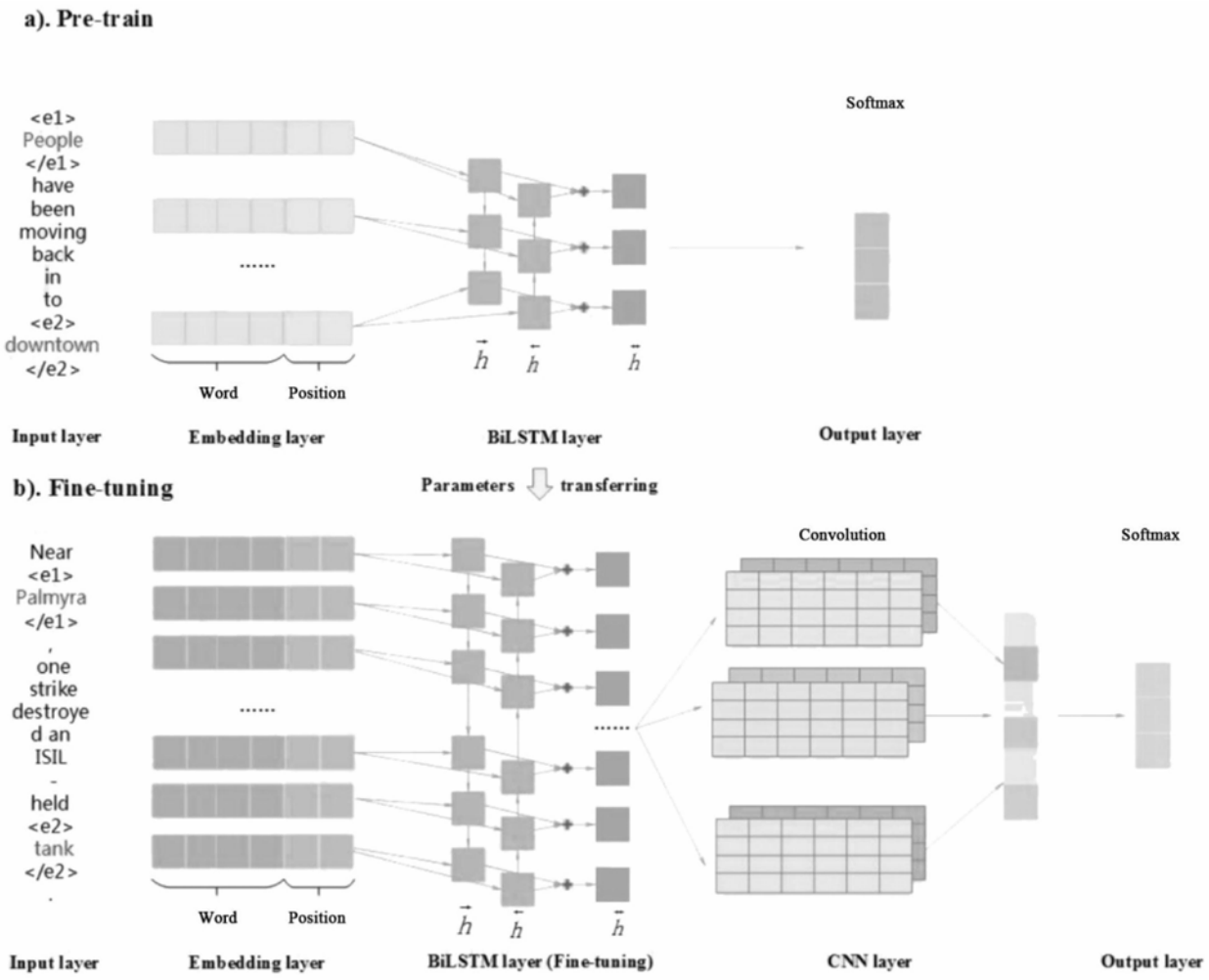
a). Pre-train
Softmax
<e1>
People
</e1>
have
been
moving
back
in
to
<e2>
downtown
</e2>
Word
Position
Input layer
Embedding layer
BiLSTM layer
Output layer
b). Fine-tuning
Parameters transferring
Convolution
Softmax
Near
<e1>
Palmyra
</e1>
,
one
strike
destroye
d an
ISIL
-
held
<e2>
tank
</e2>
.
Word
Position
Input layer
Embedding layer
BiLSTM layer (Fine-tuning)
CNN layer
Output layer

图3

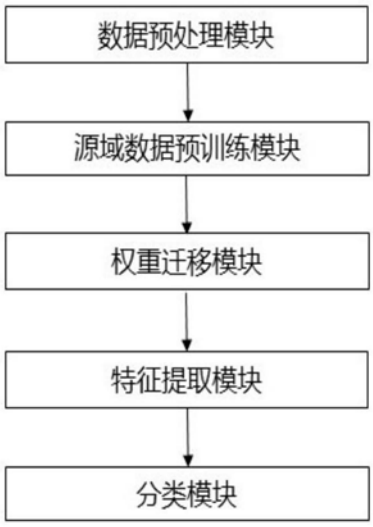

图4