(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 110807324 A
(43)申请公布日 2020.02.18

(21)申请号 201910955093.6

(22)申请日 2019.10.09

(71)申请人 四川长虹电器股份有限公司
地址 621000 四川省绵阳市高新区绵兴东路35号

(72)发明人 孙云云 刘楚雄 唐军

(74)专利代理机构 四川省成都市天策商标专利事务所 51213
代理人 郭会

(51)Int.Cl.
G06F 40/289(2020.01)
G06F 40/295(2020.01)
G06F 40/30(2020.01)
G06F 16/35(2019.01)
G06F 16/36(2019.01)
G06K 9/62(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

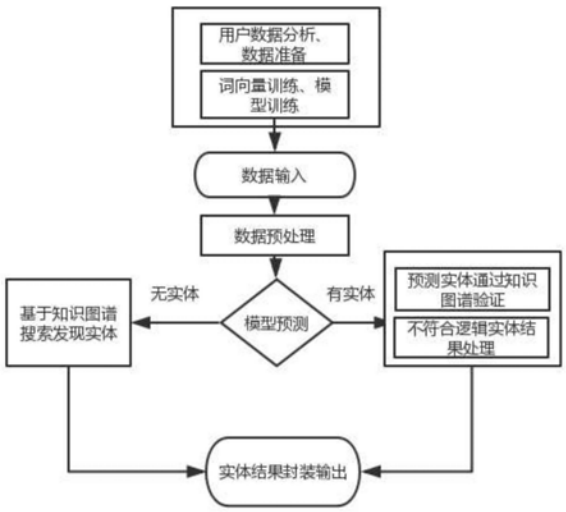
权利要求书2页 说明书8页 附图1页

(54)发明名称

一种基于IDCNN-crf与知识图谱的影视实体识别方法

(57)摘要

本发明公开了一种基于IDCNN-crf与知识图谱的影视实体识别方法,包括以下步骤:A.收集影视数据信息;B.收集大量通过语音转换为文本的用户搜索影视的数据并进行数据分析得到用于模型训练的训练数据;C.对实体识别模型进行训练;D.采集需要进行预测的预测数据,并在进行数据预处理后输入实体识别模型进行预测;E.对模型预测结果进行验证处理并输出。本发明的方法能够解决标注数据少及文本简短、口语化的影视文本数据的实体识别问题。



CN 110807324 A

1.一种基于IDCNN-crf与知识图谱的影视实体识别方法,其特征在于,包括以下步骤:

A.收集影视数据信息;

B.收集大量通过语音转换为文本的用户搜索影视的数据并进行数据分析得到用于模型训练的训练数据;

C.对实体识别模型进行训练;

D.采集需要进行预测的预测数据,并在进行数据预处理后输入实体识别模型进行预测;

E.对模型预测结果进行验证处理并输出。

2.根据权利要求1所述的一种基于IDCNN-crf与知识图谱的影视实体识别方法,其特征在于,所述步骤A具体为:从各大影视数据源实时收集影视数据信息,爬取各影视数据的包括影视名、演员、角色、人物关系的实体信息,建立影视专用的知识图谱。

3.根据权利要求1所述的一种基于IDCNN-crf与知识图谱的影视实体识别方法,其特征在于,所述步骤B包括:将采集的用户数据进行频次统计、k-Means聚类分析,并将用户说法意图相似的语句聚类到一起结合频次分析及聚类分析的结果,预知用户常用的影视搜查语句并打上标签作为训练数据。

4.根据权利要求3所述的一种基于IDCNN-crf与知识图谱的影视实体识别方法,其特征在于,所述实体识别模型由特征表示层、dropout层、IDCNN层和CRF层构成,其中,特征表示层是由词向量和字符向量组成,所述字符向量是通过LM模型训练得到的,词向量是结巴分词后按‘0/1/2/3’编码后的向量,其长度为输入文本的长度值,模型初始的参数是通过word2vec训练得的100维预训练字符向量,通过将词向量和字符级向量进行拼接以表示单词在特定语义空间下的特征;

所述dropout层用于对输入的特征进行dropout处理以预防过拟合,所述IDCNN层具体是对输入的特征分别编码当前时刻的上文和下文信息;再将两者的编码信息合并构成待解码的得分信息;所述CRF层用于将IDCNN层的输出得分作为输入,同时引入转移得分矩阵,根据序列得分选择最优的标签序列。

5.根据权利要求4所述的一种基于IDCNN-crf与知识图谱的影视实体识别方法,其特征在于,所述步骤B中还包括进行词向量训练,具体包括:

对训练数据进行包括去除特殊标点符号、英文大小写转换的预处理,然后将处理后的数据使用gensim工具包的word2vec训练,训练为维度100维的字符向量,对训练数据的句子进行结巴分词并编码得到词向量,将词向量与word2vec训练出的字符向量按一定权重相加得到最终的词向量,并将最终得到的词向量作为双向IDCNN网络的初始参数。

6.根据权利要求4所述的一种基于IDCNN-crf与知识图谱的影视实体识别方法,其特征在于,在所述步骤C中进行实体识别模型训练前还包括从训练数据中筛选出包含各个标签的常用数据,并由人工按BIO标准进行训练数据标注。

7.根据权利要求6所述的一种基于IDCNN-crf与知识图谱的影视实体识别方法,其特征在于,进行实体识别模型训练时具体包括以下步骤:

C1.将所有标注的训练数据按a、b、c的比例划分为训练数据集、测试数据集和验证数据集,其中,a+b+c=1;

C2.在训练数据集中,以句子为单位,将一个含有n个字的句子记作:$x=(x_1,x_2,\ldots,$

x_n)，其中，x_i表示句子的第i个字在字典中的id，根据x_i得到每个字的word2Id向量，其中，word2Id是通过统计训练数据集的字符个数，并按此方法得到的字符数据集；在字符数据集中按字符出现频率降序进行编码，得到字符对应的唯一的id编号数据集Word2ID，其中，未在Word2ID中出现的字符ID置0，用‘<UNK>’标记；

C3.在实体识别模型的特征表示层利用预训练或随机初始化的向量矩阵将句子中的每个字x_i由wordid向量映射为低维稠密的字向量，其中，$x_i \in R^2$；

C4.在实体识别模型的dropout层设置dropout以缓解过拟合，且dropout设置为0.5；

C5.将一个句子的各个字的字符向量序列(x_1,x_2,...,x_n)作为IDCNN层的输入，建立基于IDCNN的深度学习模型，随机抽取batch_size进行参数训练，并将膨胀算子计算的卷积矩阵组合，采用dropout正则化模型参数，在每个批次的训练中将隐层神经元随机保留一半，得到每个字符对应的非归一化的对数概率logits值，其中，$\mathrm{logits}=\ln(\frac{p}{1-p})$，p是每个字符属于某一标签的概率，由logits值将概率p由[0,1]映射到[-∞,+∞]；

C6.在实体识别模型的CRF层中进行句子级的序列标注，其中，CRF层的参数是一个(k+2)×(k+2)的矩阵A，k是不同标签的数目，A_{ij}表示从第i个标签到第j个标签的转移得分，对每个句子的各标签进行打分，记一个长度等于句子长度的标签序列y即y＝(y_1,y_2,...,y_n)，则对于句子x的标签y的打分为$score(x,y)=\sum_{i=1}^{n}P_{i,y_i}+\sum_{i=1}^{n+1}A_{y_{i-1}y_i}$，其中，$P_{i,yi}$是双向LSTM网络输出的分数矩阵，$P_{i,yi}$的大小为n×k，k是不同标签的数目，$P_{i,yi}$对应于句子中第i个单词的第$y_i$个标签的分数；最后，挑选分值最高的一个标签作为该单元的标签。

8.根据权利要求7所述的一种基于IDCNN-crf与知识图谱的影视实体识别方法，其特征在于，所述步骤C1中，a＝0.7，b＝0.2，c＝0.1。

9.根据权利要求7所述的一种基于IDCNN-crf与知识图谱的影视实体识别方法，其特征在于，所述步骤E中，对预测结果的验证处理具体是将预测结果分为有实体及无实体两类，并分别进行预测结果无实体验证处理及预测结果有实体验证处理；

其中，所述预测结果无实体验证处理包括去掉前后冗余部分和特定实体后再去搜索知识图谱得到对应的实体结果，若有对应实体结果则将其作为最终输出的预测结果，若无则进行模糊搜索；

所述预测结果有实体验证处理包括将对应实体在搜索知识图谱中验证是否有真实存在对应的实体，若有，则将其作为最终的，若无，则转为预测结果无实体验证处理。

一种基于IDCNN-crf与知识图谱的影视实体识别方法

技术领域

[0001] 本发明涉及深度学习自然语言处理技术领域,特别涉及一种基于IDCNN-crf 与知识图谱的影视实体识别方法。

背景技术

[0002] 智能电视已进入了快速发展,视频领域也积累了大量的影片、演员等非结 构化的用户数据。原语义识别系统是语音识别后的文本做简单的数据处理后, 去媒资数据库模糊搜索,由于媒资库数据量大,搜索比较耗时,且准确率不高, 一些噪音数据也有可能被识别成影片名输出,而且不能满足用户多轮请求的需 求,用户体验很差。对语音识别后文本的语义的识别即命名识别是智能电视的 关键技术之一。如何用一种有效地方式准确的提取影视实体,以帮助用户快速 找到其心仪的影视剧,成为一个重要的需求。

[0003] 目前命名实体识别在自然语言处理中较多采用深度神经网络方法,将语料 标注与词向量特征结合,通过减少人工特征在模型中的比重,减少命名实体识 别系统对于大型语料库的依赖;并通过概率统计降低规则方法的复杂度,有效 提高了模型的性能。在实际工程中主要采用长短期记忆(Long Short Term Memory,LSTM)神经网络及卷积神经网络(Convolutional Neural Networks,CNN) 等深度学习算法。目前对于命名实体识别任务,表现效果最好的算法基本上采 用双向LSTM(Bidirectional LSTM),避免了模型庞大的参数优化问题。采用 BLSTM通过词向量、字符向量等特征,建立Embedding层,再经过双向的LSTM 层,最后为CRF层。该模型集成了深度学习方法的优势,无需特征工程,仅使 用字符向量就可以达到很好的效果。对于序列标注,CNN有一个不足,就是卷 积之后,末层神经元可能只是得到了原始输入数据中一小块的信息。而对NER 来讲,整个输入句子中每个字都有可能对当前位置的标注产生影响,即所谓的 长距离依赖问题。为了覆盖到全部的输入信息就需要加入更多的卷积层,导致 层数越来越深,参数越来越多。而为了防止过拟合又要加入更多的Dropout之类 的正则化,带来更多的超参数,整个模型变得庞大且难以训练。但biLSTM又 有个问题,在对GPU并行计算的利用上不如CNN那么强大。

[0004] 因此,提出了dilated CNN模型,意思是“膨胀的”CNN。其想法并不复杂: 正常CNN的filter,都是作用在输入矩阵一片连续的区域上,不断sliding做卷 积。dilated CNN为这个filter增加了一个dilation width,作用在输入矩阵的时候, 会skip所有dilation width中间的输入数据;而filter本身的大小保持不变,这样 filter获取到了更广阔的输入矩阵上的数据,看上去就像是“膨胀”了一般。而 且与其他领域相比,影视领域涉及实体类别复杂,包含的实体种类也千差万别,” 扫毒”和电影‘扫毒’,看似同一实体却属于不同实体类型,而且实体的命名方 式无法统一,用户普通话不标准,平翘舌不分,同一实体不同表达方式等,都 对语音识别后的命名实体识别产生很大影响。

发明内容

[0005] 本发明的目的是克服上述背景技术中不足,提供一种基于IDCNN-crf与知 识图谱

的影视实体识别方法,能够解决标注数据少及文本简短、口语化的影视 文本数据的实体识别问题。

[0006] 为了达到上述的技术效果,本发明采取以下技术方案:

[0007] 一种基于IDCNN-crf与知识图谱的影视实体识别方法,包括以下步骤:

[0008] A.收集影视数据信息;

[0009] B.收集大量通过语音转换为文本的用户搜索影视的数据并进行数据分 析得到用于模型训练的训练数据;

[0010] C.对实体识别模型进行训练;

[0011] D.采集需要进行预测的预测数据,并在进行数据预处理后输入实体识 别模型进行预测;

[0012] E.对模型预测结果进行验证处理并输出。

[0013] 进一步地,所述步骤A具体为:从各大影视数据源实时收集影视数据信息, 爬取各影视数据的包括影视名、演员、角色、人物关系的实体信息,建立影视 专用的知识图谱;如可从豆瓣、百度百科等,爬取各影视名、演员、角色、人 物关系等各实体信息,建立影视专用的知识图谱,其中,知识图谱的建立和维 护不是本方案的重点,本方案只借助知识图谱做进一步验证,因此对应的具体 步骤在此不再赘述。

[0014] 进一步地,所述步骤B包括:将采集的用户数据进行频次统计、k-Means 聚类分析,并将用户说法意图相似的语句聚类到一起结合频次分析及聚类分析 的结果,预知用户常用的影视搜查语句并打上标签作为训练数据;具体的,对 从电视端采集的大量用户数据做频次统计、k-Means聚类分析,经过测试调参这 里选择15个聚类点,会将用户说法意图相似的语句聚类到一起结合频次分析及 聚类分析的结果,大概预知用户常用的影视搜查语句,确定要识别的实体类型。

[0015] 进一步地,所述实体识别模型由特征表示层、dropout层、IDCNN层和CRF 层构成,其中,特征表示层即嵌入层是由词向量和字符向量组成,所述字符向 量是通过LM模型训练得到的,词向量是结巴分词后按'0/1/2/3'编码后的向量, 其长度为输入文本的长度值,模型初始的参数是通过word2vec训练得的100维 预训练字符向量,通过将词向量和字符级向量进行拼接以表示单词在特定语义 空间下的特征;

[0016] 所述dropout层用于对输入的特征进行dropout(随机失活)处理以预防过 拟合,所述IDCNN层具体是对输入的特征分别编码当前时刻的上文和下文信息; 再将两者的编码信息合并构成待解码的得分信息;具体的,实体识别模型是4 个大的相同结构的Dilated CNN block拼在一起,每个block里面是dilation width 为1,1,2的三层Dilated卷积层,所以叫做Iterated Dilated CNN,IDCNN层可 对输入句子的每一个字生成一个logits;所述CRF层用于将IDCNN层的输出得 分作为输入,同时引入转移得分矩阵,根据序列得分选择最优的标签序列。

[0017] 进一步地,所述步骤B中还包括进行词向量训练,具体包括:对训练数据 进行包括去除特殊标点符号、英文大小写转换的预处理,然后将处理后的数据 使用gensim工具包的word2vec训练,训练为维度100维的字符向量,对训练数 据的句子进行结巴分词并编码得到词向量,将词向量与word2vec训练出的字符 向量按一定权重相加得到最终的词向量,并将最终得到的词向量作为双向 IDCNN网络的初始参数;

[0018] 本方案中，用大量真实数据训练的词向量一定程度上解决了标注数据少的 情况下使用深度神经网络做实体识别的问题，IDCNN神经网络的初始参数不再 是没有意义的随机参数，大量数据训练的词向量会得到中文字偏旁等初始信息 作为神经网络的底层输入，本文还加入了优化后的词向量，一句话中一个单独 的字符可能没有实际意思，而正确的分词向量却对整句话起着重要作用，将字 符向量和词向量结合更能体现文本的整体特征。

[0019] 进一步地，在所述步骤C中进行实体识别模型训练前还包括从训练数据中 筛选出包含各个标签的常用数据，并由人工按BIO标准进行训练数据标注。

[0020] 进一步地，进行实体识别模型训练时具体包括以下步骤：

[0021] C1.将所有标注的训练数据按a、b、c的比例划分为训练数据集、测试数据 集和验证数据集，其中，a+b+c＝1；

[0022] C2.在训练数据集中，以句子为单位，将一个含有n个字的句子记作：$x=(x_1, x_2,...,x_n)$，其中，x_i表示句子的第i个字在字典中的id，根据x_i得到每个字 的word2Id向量，其中，word2Id是通过统计训练数据集的字符个数，并按此方 法得到的字符数据集；在字符数据集中按字符出现频率降序进行编码，得到字 符对应的唯一的id编号数据集Word2ID，其中，未在Word2ID中出现的字符ID 置0，用‘<UNK>’标记；

[0023] C3.在实体识别模型的特征表示层利用预训练或随机初始化的向量矩阵将句 子中的每个字x_i由wordid向量映射为低维稠密的字向量，其中，$x_i \in R^2$；

[0024] C4.在实体识别模型的dropout层设置dropout以缓解过拟合，且dropout设 置为0.5；

[0025] C5.将一个句子的各个字的字符向量序列$(x_1,x_2,...,x_n)$作为IDCNN层的输入，建立基于IDCNN的深度学习模型，随机抽取batch_size进行参数训练，并将膨 胀算子计算的卷积矩阵组合，采用dropout正则化模型参数，在每个批次的训练 中将隐层神经元随机保留一半，得到每个字符对应的非归一化的对数概率logits 值，其中，$\text{logits}=\ln(\frac{p}{1-p})$，p是每个字符属于某一标签的概率，由logits值将概 率p由[0,1]映射到[-∞,+∞]；

[0026] 在本实施例的IDCNN层中，dilated width会随着层数的增加而指数增加，这样随着层数的增加，参数数量是线性增加的，而receptive field却是指数增加 的，可以很快覆盖到全部的输入数据；

[0027] C6.在实体识别模型的CRF层中进行句子级的序列标注，其中，CRF层的 参数是一个(k+2)×(k+2)的矩阵A，k是不同标签的数目，A_{ij}表示从第i个标签 到第j个标签的转移得分，对每个句子的各标签进行打分，记一个长度等于句子 长度的标签序列y即$y=(y_1, y_2,...,y_n)$，则对于句子x的标签y的打分为 $score(x,y)=\sum_{i=1}^{n}P_{i,y_i}+\sum_{i=1}^{n+1}A_{y_{i-1}y_i}$，其中，P是双向LSTM网络输出的分数矩阵，P 的大小为n×k，k是不同标签的数目，$P_{i,yi}$对应于句子中第i个单词的第y_i个标 签的分数；最后，挑选分值最高的一个标签作为该单元的标签；由此可看出整 个序列的打分等于各个位置的打分之和。

[0028] 作为优选，CRF层可以为最后预测的标签添加一些约束来保证预测的标签 是合法的，且在训练数据训练过程中，这些约束可以通过CRF层自动学习到，有了这些约束，标签序列预测中非法序列出现的概率将会大大降低。

[0029] 进一步地,所述步骤C1中,a＝0.7,b＝0.2,c＝0.1。

[0030] 进一步地,所述步骤E中,对预测结果的验证处理具体是将预测结果分为 有实体及无实体两类,并分别进行预测结果无实体验证处理及预测结果有实体 验证处理;其中,所述预测结果无实体验证处理包括去掉前后冗余部分和特定 实体后再去搜索知识图谱得到对应的实体结果,若有对应实体结果则将其作为 最终输出的预测结果,若无则进行模糊搜索;所述预测结果有实体验证处理包 括将对应实体在搜索知识图谱中验证是否有真实存在对应的实体,若有,则将 其作为最终的,若无,则转为预测结果无实体验证处理。

[0031] 本发明与现有技术相比,具有以下的有益效果:

[0032] 本发明的基于IDCNN-crf与知识图谱的影视实体识别方法,通过对大量用 户数据的分析、人工数据标注、模型训练;获取待识别文本的字符向量和词向 量,对字符向量和词向量进行加权求和,得到加权求和结果;将加权求和结果 输入至IDCNN模型中进行处理,得到文本特征序列;将文本特征序列输入至目 标CRF模型中进行处理,得到待识别文本的命名实体识别结果;将命名实体识 别结果去影视知识图谱中查询进一步验证结果,避免不合逻辑的实体结果,其 中,获取待识别文本的字符向量和词向量之后,通过对字符向量和词向量进行 加权求和,更好的利用了字符特征和词特征信息,通过采用膨胀的卷积神经网络模型尽可能记住全句的信息来对当前字做标注,再结合CRF模型进行处理及 知识图谱的实体验证,从而提高了命名实体识别的准确率,可以实现从大规模、表达口语化的语音文本中准确高效地提取实体,有助于用户在电视端进行影视 实体的搜索,提高用户体验。

附图说明

[0033] 图1是本发明的基于IDCNN-crf与知识图谱的影视实体识别方法流程示意 图。

[0034] 图2是本发明的一个实施例的实体识别模型的结构示意图。

具体实施方式

[0035] 下面结合本发明的实施例对本发明作进一步的阐述和说明。

[0036] 实施例:

[0037] 实施例一:

[0038] 现用比较多的实体识别方法都是基于biLSTM,biLSTM本质是一个序列模 型,但在对GPU并行计算的利用上不如CNN那么强大,且在应用于线上系统 时,随着用户增多,对模型训练及预测时间要求也比较高,在高并发下模型的 性能、处理时间尤为重要,本实施例即提供了一种基于IDCNN-crf与知识图谱 的影视实体识别方法,其IDCNN-crf模型在训练及预测时间均优于bilstm+crf, 能够解决标注数据少及文本简短、口语化的影视文本数据的实体识别问题。

[0039] 具体的,如图1所示,本实施例的基于IDCNN-crf与知识图谱的影视实体 识别方法包括以下步骤:

[0040] 步骤1.从各大影视数据源实时收集影视数据信息,如,豆瓣、百度百科等, 爬取各影视名、演员、角色、人物关系等各实体信息,建立影视专用的知识图 谱。

[0041] 步骤2.从电视端收集通过语音转换为文本的用户搜索影视的数据;分析收 集到的数据,对有一定规律的(指用户常用的搜索影视的句式)、用户常用的搜 索语句打标签,

用于训练模型并进行词向量训练。

[0042] 具体包括从大量采集到的用户数据中通过K-means聚类、频次、用户行为 数据等分析用户影视搜索的基本需求，如常用搜索句式、按什么条件搜索视频 等，结合业务需求，确定实体类别及命名；然后人工按BIO标准标注训练数据，由于没现成可用的标注数据，利用大量用户真实数据及word2vec语言模型训练 100维度的字符向量及分词向量，并按一定权重拼接文本的字符向量和词向量，作为双向IDCNN的底层输入。

[0043] 具体的，本实施例中，对从电视端采集的大量用户数据做频次统计、k-Means 聚类分析，经过测试调参本实施例中选择15个聚类点，会将用户说法意图相似 的语句聚类到一起结合频次分析及聚类分析的结果，大概预知用户常用的影视 搜查语句，确定要识别的实体类型、标签统一命名，本实施例中，目前有27个 标签。

[0044] 其中，词向量训练前需要对数据进行预处理，包括去除特殊标点符号、英 文大小写转换等，再将处理后的数据大量用户规范数据使用gensim工具包的 word2vec训练，训练为维度100维的字符向量；词向量对句子进行结巴分词并 编码，例如“我/想/看/刘德华/的/天下无贼，编码为“0/0/0/123/0/1223”，最后将 词向量与word2vec训练出的字符向量按一定权重相加，最终得到词向量作为双 向IDCNN网络的初始参数。

[0045] 作为优选，在词向量训练中还涉及到对于分词的优化，比如‘流淌的美好 生活’结巴分词不会记录这是影片名为一组分词结果，本方案将热门影片作为 结巴的自定义字典，这样提高分词和实体识别的效果。最后再将词向量和字符 级向量进行拼接作为词向量层。

[0046] 本方案中，用大量真实数据训练的词向量一定程度上解决了标注数据少的 情况下使用深度神经网络做实体识别的问题，IDCNN神经网络的初始参数不再 是没有意义的随机参数，大量数据训练的词向量会得到中文字偏旁等初始信息 作为神经网络的底层输入，本方案还加入了优化后的词向量，一句话中一个单 独的字符可能没有实际意思，而正确的分词向量却对整句话起着重要作用，将 字符向量和词向量结合更能体现文本的整体特征。

[0047] 步骤3.进行实体识别模型的训练。

[0048] 具体的，如图2所示，本实施例中，实体识别模型主要由嵌入层、dropout 层、IDCNN层和CRF层4部分构成。

[0049] 其中，嵌入层(即特征表示层)：主要由词向量和字符向量组成；字符向量 是通过LM模型训练得到的100维向量，词向量对句子进行结巴分词并编码， 例如“我/想/看/刘德华/的/天下无贼，编码为“0/0/0/123/0/1223”。

[0050] dropout层主要用于在特征输入IDCNN网络层之前做随机失活dropout，从 而一定程度缓解过拟合。

[0051] IDCNN层具体是对输入的特征分别编码当前时刻的上文和下文信息；再将 两者的编码信息合并构成待解码的得分信息。本模型是4个大的相同结构的Dilated CNN block拼在一起，每个block里面是dilation width为1,1,2的三层 Dilated卷积层，所以叫做Iterated Dilated CNN，IDCNN对输入句子的每一个字 生成一个logits值作为待解码的得分信息。

[0052] CRF层则是用于接受IDCNN的输出得分作为输入，同时引入转移得分矩 阵，根据序列得分选择全最优的标签序列。

[0053] 具体的，在进行模型训练时，具体包括以下步骤：

[0054] 先将所有标注的训练数据按0.7、0.2、0.1的比例划分为训练数据集、测试 数据集和验证数据集。其中，测试数据集和验证数据集分别是用于对模型的测 试及对于测试结果的验证，具体的测试及验证过程类似于模型的训练过程，且 本方案的重点在于对模型的训练过程，因此对于模型的测试及验证此处不再赘 述。然后以句子为单位，将一个含有n个字的句子(字的序列)记作：$x=(x_1,x_2,...,x_n)$；其中，x_i表示句子的第i个字在字典中的id，根据x_i得到每个字 的word2Id向量，其中，word2Id是通过统计训练数据集的字符个数，并按此方 法得到的字符数据集；在字符数据集中按字符出现频率降序进行编码，得到字 符对应的唯一的id编号数据集Word2ID，其中，未在Word2ID中出现的字符ID 置0，用‘<UNK>’标记。

[0055] 模型的第一层是嵌入层(特征表示层)，利用预训练或随机初始化的 embedding矩阵将句子中的每个字x_i由wordid向量映射为低维稠密的字向量 (character embedding)，$x_i \in R^2$是embedding的维度。

[0056] 模型的第二层是dropout层，在输入下一层之前，设置dropout以缓解过拟 合，本实施例中dropout设置为0.5。

[0057] 模型的第三层是IDCNN层，正常CNN的filter，都是作用在输入矩阵一片 连续的区域上，不断sliding做卷积。而dilated CNN为这个filter增加了一个 dilation width(膨胀宽度)，作用在输入矩阵的时候，会skip所有dilation width 中间的输入数据；而filter本身的大小保持不变，这样filter获取到了更广阔的输 入矩阵上的数据，看上去就像是“膨胀”了一般。则具体使用时，dilated width 会随着层数的增加而指数增加。这样随着层数的增加，参数数量是线性增加的， 而receptive field却是指数增加的，可以很快覆盖到全部的输入数据。

[0058] 最终，将一个句子的各个字的字符向量序列$(x_1,x_2,...,x_n)$作为IDCNN层的输 入，建立基于IDCNN的深度学习模型，随机抽取batch_size进行参数训练，并 将膨胀算子计算的卷积矩阵组合，采用dropout正则化模型参数，在每个批次的 训练中将隐层神经元随机保留一半，得到每个字符对应的非归一化的对数概率 logits值，其中，$\text{logits}=\ln(\frac{p}{1-p})$，p是每个字符属于某一标签的概率，由logits 值将概率p由[0,1]映射到[-∞,+∞]。

[0059] 模型的第四层是CRF层，进行句子级的序列标注。CRF层的参数是一个 (k+2)×(k+2)的矩阵A，k为句子中一个字的向量的长度，如本实施例中，k＝100； A_{ij}表示从第i个标签到第j个标签的转移得分，进而在一个位置进行标注的时候 可以利用此前已经标注过的标签，之所以要加2是应为要为句子首部添加一个 起始状态以及为句子尾部添加一个终止状态，对每个句子的各标签进行打分， 记一个长度等于句子长度的标签序列$y=(y_1,y_2,...,y_n)$，则对于句子x的标签y的 打分为$score(x,y)=\sum_{i=1}^{n}P_{i,y_i}+\sum_{i=1}^{n+1}A_{y_{i-1}y_i}$，其中，P是双向LSTM网络输出的分数矩 阵，P的大小为n×k，k是不同标签的数目，$P_{i,yi}$对应于句子中第i个单词的第 y_i个标签的分数；最后，挑选分值最高的一个标签作为该单元的标签。则可以看 出整个序列的打分等于各个位置的打分之和，而每个位置的打分由两部分得到， 一部分是由双向LSTM网络输出的分数矩阵P决定，另一部分则由CRF的转移 矩阵A决定。

[0060] CRF层可以为最后预测的标签添加一些约束来保证预测的标签是合法的。作为优

选,在训练数据训练过程中,这些约束可以通过CRF层自动学习到,有 了这些约束,标签序列预测中非法序列出现的概率将会大大降低,由于IDCNN 的输出为单元的每一个标签分值,我们可以挑选分值最高的一个作为该单元的 标签。

[0061] 具体的,在实施例中,模型训练主要分为以下3部分:

[0062] 第一,输入字/词向量表示。

[0063] 使用wordId向量表示每个字符,分词向量按‘0/1/2/3’方式编码,例如“我 /想/看/刘德华/的/天下无贼,编码为“0/0/0/123/0/1223”。将从单个字(单个字母) 中提取一些含义,从词向量中获取句子及部分上下文的含义。对每一个字,需 要构建一个向量来获取这个字的意思以及对实体识别有用的一些特征,本方案 中这个向量由Word2Vec训练的字向量和从词向量中提取出特征的向量按权重 堆叠而成的。

[0064] 第二,上下文字的语义表示。

[0065] 对上下文中的每一个字,需要有一个有意义的向量表示。使用IDCNN来获 取上下文中字的向量表示。利用IDCNN的膨胀机制,可以迅速的扫描获取输入 文本每个字的上下文信息。

[0066] 第三,进行实体标签的预测。

[0067] 这一阶段计算标签得分,使用每个字对应的logits来做最后预测,可以使用 一个全连接神经网络来获取每个实体标签的得分。记一个长度等于句子长度的 标签序列y=($y_1,y_2,...,y_n$),则对于句子x的标签y的打分为 $score(x,y)=\sum_{i=1}^{n}P_{i,y_i}+\sum_{i=1}^{n+1}A_{y_{i-1}y_i}$,其中,P是双向LSTM网络输出的分数矩阵,P的大 小为n×k,k是不同标签的数目,P_{i,y_i}对应于句子中第i个单词的第y_i个标签的 分数;最后,挑选分值最高的一个标签作为该单元的标签。

[0068] 本实施例中,具体使用了线性crf给实体标签的得分:softmax方法是做局部 选择,没有利用周围的标签来帮助决策。例如:“张三”,当我们给了三“I-actor” 这个标签后,这应该帮助我们决定“张”对应I-actor的起始位置。线性CRF定 义了全局得分。最后,将训练的模型及相应参数保存。

[0069] 步骤4.采集需要进行预测的预测数据,并在进行数据预处理。

[0070] 数据预处理时主要是去特殊符号等;将文本数据处理为模型预测要求的格 式,即将文本转换为wordId词向量,维度是训练数据词库字典的长度。

[0071] 步骤5.进行模型预测。

[0072] 将处理后的数据输入经训练的模型进行预测,预测结果可能的情况如下:

[0073] (1)、看张二谋导演的红X梁

[0074] O B-director I-director I-director O O B-movie_name I-movie_name I-movie_name

[0075] (2)、刘小四的大话X游

[0076] B-actor I-actor I-actor O B-movie_name I-movie_name I-movie_name

[0077] (3)、张三那期的快乐X本营

[0078] B-actor I-actor O O O B-movie_name I-movie_name I-movie_name I-movie_name I-movie_name

[0079] (4)、王二可不可以不悲伤

[0080] O O O O O O B-movie_name I-movie_name I-movie_name

[0081] (5)、钱某参加的综艺

[0082] O O O O O O O

[0083] 步骤6.预测结果的验证与处理。

[0084] 具体包括预测结果无实体处理及预测结果有实体验证处理。其中,如对于 上述第(5)种预测结果即没有预测出实体情况处理方法如下:

[0085] 首先,进行数据处理,去除前后冗余部门‘我想看’,‘我要看’、‘播放’、‘有吗’等,然后再对film集/季/部、版本、语言等实体规则提取,由于事先维 护了语言、版本、国家等不长变动这些数据同时在存于知识图谱,类似{‘英语’: ‘英语’,‘英文’:‘英语’,‘外语’:‘英语’}形式,会将其所有同义词考虑在 内。将对应的实体用正则匹配后并将实体替换为空,如‘我想看速度与XX英 文版’,如果模型没有预测实体结果,去掉前后冗余部分和特定实体后‘速度与 激情英’再去搜索知识图谱得到对应的实体结果。如果还没找到实体会分词后 再模糊搜索。

[0086] 如对于上述第(1)、(2)、(3)、(4)种预测结果即有实体结果标签处理方法如下:

[0087] 将对应实体搜索知识图谱验证是否有真实存在这样的实体,如(2)中刘小四 实际没有演过大话X游,将向用户推荐刘小四的其他电影,而不是返回用户没 找到该影片。预测结果(3)实际用户张三最新一期参加的快乐X本营节目,此是 实体抽象关系的挖掘,能够更好的满足用户需求。知识图谱验证进一步提高了 实体的效果。对(4)这种虽然有实体结果,但在知识图谱中没找到对应的影视名 称实体视为预测失败,再执行预测结果没实体处理。

[0088] 步骤7.实体结果封装输出。

[0089] 本步骤中还包含对不符合逻辑的实体预测结果处理,如对于‘刘某华第三 集’的识别结果为actor:刘某华,season:;则在结果输出时会将删除season 实体,将“刘某华”作为识别结果封装。

[0090] 可以理解的是,以上实施方式仅仅是为了说明本发明的原理而采用的示例 性实施方式,然而本发明并不局限于此。对于本领域内的普通技术人员而言, 在不脱离本发明的精神和实质的情况下,可以做出各种变型和改进,这些变型 和改进也视为本发明的保护范围。

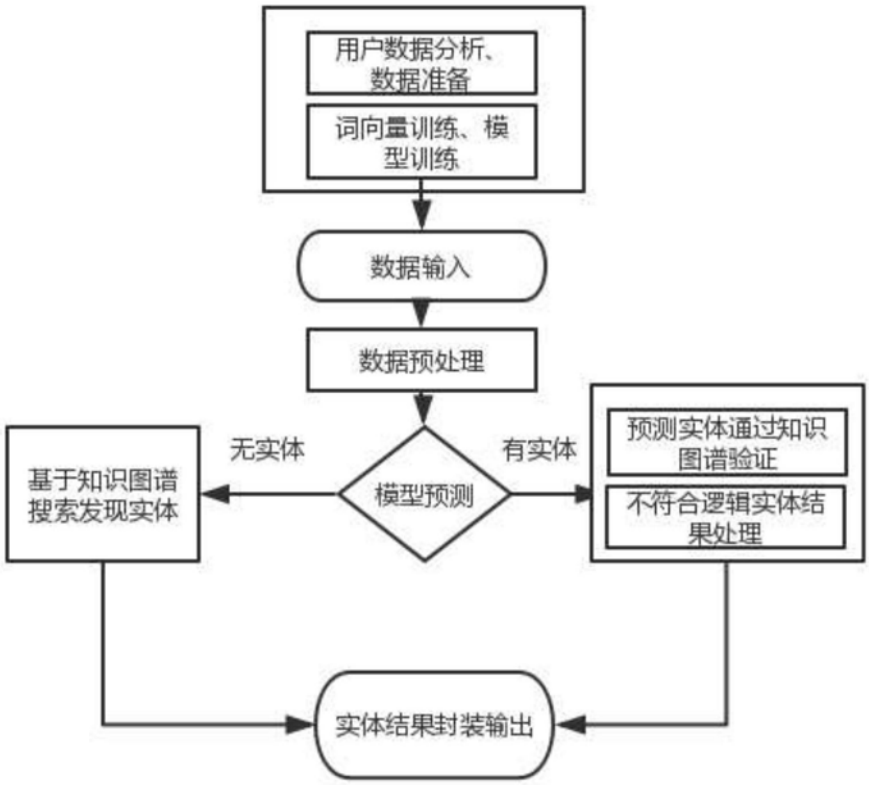

用户数据分析、数据准备
词向量训练、模型训练
数据输入
数据预处理
无实体
有实体
基于知识图谱搜索发现实体
模型预测
预测实体通过知识图谱验证
不符合逻辑实体结果处理
实体结果封装输出

图1

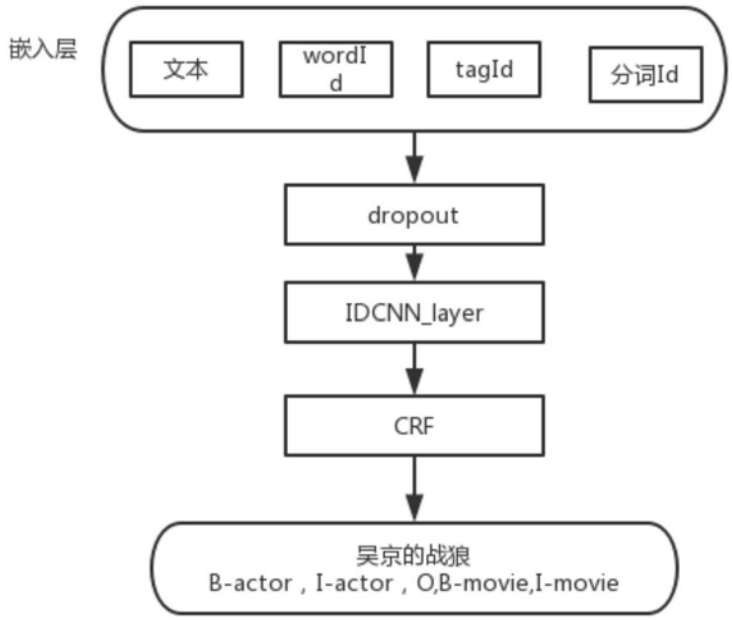

嵌入层
文本
wordId
tagId
分词Id
dropout
IDCNN_layer
CRF
吴京的战狼
B-actor , I-actor , O,B-movie,I-movie

图2