(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 113255294 A
(43)申请公布日 2021.08.13

(21)申请号 202110797174.5

(22)申请日 2021.07.14

(71)申请人 北京邮电大学
地址 100876 北京市海淀区西土城路10号

(72)发明人 杜军平 于润羽 薛哲 徐欣

(74)专利代理机构 北京金咨知识产权代理有限公司 11612
代理人 薛海波

(51)Int.Cl.
G06F 40/126(2020.01)
G06F 40/295(2020.01)
G06F 40/30(2020.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

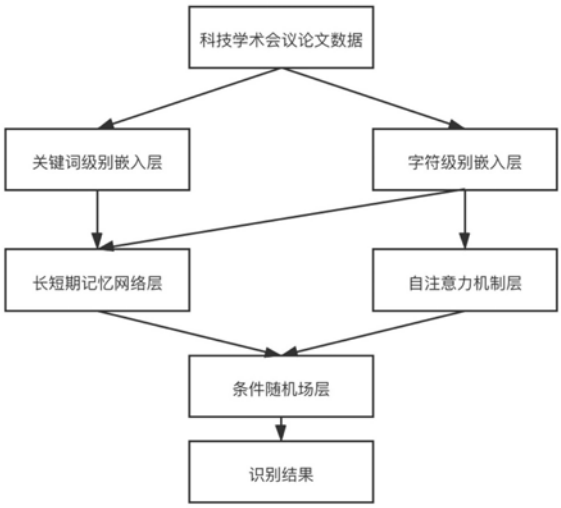
权利要求书2页 说明书11页 附图2页

(54)发明名称

命名实体识别模型训练方法、识别方法及装置

(57)摘要

本发明提供一种命名实体识别模型训练方法、识别方法及装置，所述模型训练方法的初始神经网络模型由通过结合关键字符级别编码和词级别编码对科技论文数据进行向量表示，将字符级别向量和词级别向量引入双向长短期记忆网络能够挖掘上下文关系，同时挖掘关键词的语义特征，提升了分词边界的准确性；通过将字符级别向量引入自注意力机制模型，能够更高效地捕捉数据内部相关性，提升命名实体识别的准确率。



CN 113255294 A

1.一种命名实体识别模型训练方法,其特征在于,包括:

获取多个科技论文数据,各科技论文数据包含一个或多个关键词,对各科技论文数据进行序列标注,以得到训练样本集;

获取初始神经网络模型,所述初始神经网络模型对所述科技论文数据的各单个中文字符进行字符级别编码得到相应的字符级别向量、对所述科技论文数据的各关键词进行词级别编码得到相应的词级别向量;将各字符级别向量和各词级别向量进行连接后输入至双向长短期记忆网络,由所述双向长短期记忆网络输出第一特征向量;将各字符级别向量输入自注意力机制模块,将所述自注意力机制模块输出与原始的各字符级别向量连接得到第二特征向量;将所述第一特征向量与所述第二特征向量进行融合,并输入条件随机场后输出命名实体识别结果;

采用所述训练样本集对所述初始神经网络模型进行训练,对所述双向长短期记忆网络、所述自注意力机制模块以及所述条件随机场的参数进行调整迭代,得到目标命名实体识别模型。

2.根据权利要求1所述的命名实体识别模型训练方法,其特征在于,所述初始神经网络模型采用word2vec模型获取各单个中文字符对应的字符级别向量以及各关键词对应的词级别向量。

3.根据权利要求2所述的命名实体识别模型训练方法,其特征在于,将各字符级别向量和各词级别向量进行连接后输入至双向长短期记忆网络,包括:

将单个字符对应的字符级别向量和词级别向量进行归一化求和得到该字符对应的第一输入序列,并输入至所述双向长短期记忆网络,计算式为:

$$c_j^c=\sum_b \alpha_j^c \otimes \tilde{c}_j^c+\alpha_{b,j}^w \otimes c_{b,j}^w ;$$

其中, c_j^c 为第j个字符对应的第一输入序列, $\tilde{c}_j^c$ 为第j个字符对应的字符级别向量, α_j^c 为 $\tilde{c}_j^c$ 对应的归一化系数, $c_{b,j}^w$ 为第j个字符所属关键词的词级别向量, $\alpha_{b,j}^w$ 为 $c_{b,j}^w$ 的归一化系数,b为第j个字符所属关键词的序数。

4.根据权利要求3所述的命名实体识别模型训练方法,其特征在于,将所述第一特征向量与所述第二特征向量进行融合,包括:

将所述第一特征向量与所述第二特征向量进行归一化求和,计算式为:

$$y_i=\alpha_i^h \otimes h_i+\alpha_i^a \otimes a_i ;$$

$$\alpha_i^h=\frac{e^{h_i}}{e^{h_i}+e^{a_i}} ;$$

$$\alpha_i^a = \frac{e^{a_i}}{e^{h_i} + e^{a_i}};$$

其中，y_i为所述科技论文数据第i个字符的特征值，h_i为所述科技论文数据第i个字符经所述双向长短期记忆网络输出的特征值，a_i为所述科技论文数据第i个字符经所述自注意力机制模块输出的特征值，α_i^h为h_i的归一化系数，α_i^a为a_i的归一化系数；e为自然底数。

5.根据权利要求4所述的命名实体识别模型训练方法，其特征在于，采用所述训练样本集对所述初始神经网络模型进行训练，包括：采用交叉熵函数作为损失函数，对所述双向长短期记忆网络、所述自注意力机制模块以及所述条件随机场的参数进行调整迭代。

6.根据权利要求1所述的命名实体识别模型训练方法，其特征在于，对各科技论文数据进行序列标注采用BIO标注。

7.根据权利要求2所述的命名实体识别模型训练方法，其特征在于，所述word2vec模型采用科技论文数据进行预训练。

8.一种命名实体识别方法，其特征在于，包括：

获取待处理的科技论文数据，将所述科技论文数据输入如权利要求1至7任意一项所述命名实体识别模型训练方法的目标命名实体识别模型中，输出命名实体识别结果。

9.一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，其特征在于，所述处理器执行所述程序时实现如权利要求1至8任一项所述方法的步骤。

10.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现权利要求1至8任一项所述方法的步骤。

命名实体识别模型训练方法、识别方法及装置

技术领域

[0001] 本发明涉及数据处理技术领域，尤其涉及一种命名实体识别模型训练方法、识别方法及装置。

背景技术

[0002] 科技大数据可以定义为与科研相关的活动产生的海量数据，其以论文数据为主体，具有数据规模大，内容专业化，特征属性繁多的特点。科技学术会议数据包含某个领域内的论文集合。以学术会议为单位进行画像的构建，可以帮助科研人员快速获得有价值的科研信息，而构建画像的核心工作即为命名实体识别。

[0003] 命名实体识别(Named Entity Recognition, NER)是自然语言处理领域中的一个重要研究方向，其目的是将给定文本中的实体按照预定义好的类别进行分类，是一种序列标注问题。学术会议论文数据的命名实体识别与通用领域的识别有一定区别，主要原因在于通用领域的数据集有较为严格的句子组成规范。但由于科研领域技术更新迭代快，导致论文数据集中有大量的专业术语。同时实体之间可能相互嵌套，增加了实体识别的难度。中文命名实体识别的效果和分词结果直接相关，如果在分词阶段发生错误，会严重影响识别效果。因此，亟需一种新的命名实体识别方法。

发明内容

[0004] 本发明实施例提供了一种命名实体识别模型训练方法、识别方法及装置，以消除或改善现有技术中存在的一个或更多个缺陷，解决中文科技论文分词效果较差，导致识别结果准确率低的问题。

[0005] 本发明的技术方案如下：

一方面，本发明提供一种命名实体识别模型训练方法，包括：

获取多个科技论文数据，各科技论文数据包含一个或多个关键词，对各科技论文数据进行序列标注，以得到训练样本集；

获取初始神经网络模型，所述初始神经网络模型对所述科技论文数据的各单个中文字符进行字符级别编码得到相应的字符级别向量、对所述科技论文数据的各关键词进行词级别编码得到相应的词级别向量；将各字符级别向量和各词级别向量进行连接后输入至双向长短期记忆网络，由所述双向长短期记忆网络输出第一特征向量；将各字符级别向量输入自注意力机制模块，将所述自注意力机制模块输出与原始的各字符级别向量连接得到第二特征向量；将所述第一特征向量与所述第二特征向量进行融合，并输入条件随机场后输出命名实体识别结果；

采用所述训练样本集对所述初始神经网络模型进行训练，对所述双向长短期记忆网络、所述自注意力机制模块以及所述条件随机场的参数进行调整迭代，得到目标命名实体识别模型。

[0006] 在一些实施例中，所述初始神经网络模型采用word2vec模型获取各单个中文字符

对应的字符级别向量以及各关键词对应的词级别向量。

[0007] 在一些实施例中，将各字符级别向量和各词级别向量进行连接后输入至双向长短期记忆网络，包括：

将单个字符对应的字符级别向量和词级别向量进行归一化求和得到该字符对应的第一输入序列，并输入至所述双向长短期记忆网络，计算式为：

$$c_j^c = \sum_b \alpha_j^c \otimes \tilde{c}_j^c + \alpha_{b,j}^w \otimes c_{b,j}^w \ ;$$

其中，c_j^c 为第j个字符对应的第一输入序列，$\tilde{c}_j^c$ 为第j个字符对应的字符级别向量，α_j^c 为 $\tilde{c}_j^c$ 对应的归一化系数，$c_{b,j}^w$ 为第j个字符所属关键词的词级别向量，$\alpha_{b,j}^w$ 为 $c_{b,j}^w$ 的归一化系数，b为第j个字符所属关键词的序数。

[0008] 在一些实施例中，将所述第一特征向量与所述第二特征向量进行融合，包括：

将所述第一特征向量与所述第二特征向量进行归一化求和，计算式为：

$$y_i = \alpha_i^h \otimes h_i + \alpha_i^a \otimes a_i \ ;$$

$$\alpha_i^h = \frac{e^{h_i}}{e^{h_i} + e^{a_i}} ;$$

$$\alpha_i^a = \frac{e^{a_i}}{e^{h_i} + e^{a_i}} ;$$

其中，y_i 为所述科技论文数据第i个字符的特征值，h_i 为所述科技论文数据第i个字符经所述双向长短期记忆网络输出的特征值，a_i 为所述科技论文数据第i个字符经所述自注意力机制模块输出的特征值，α_i^h 为 h_i 的归一化系数，α_i^a 为 a_i 的归一化系数；e为自然底数。

[0009] 在一些实施例中，采用所述训练样本集对所述初始神经网络模型进行训练，包括：采用交叉熵函数作为损失函数，对所述双向长短期记忆网络、所述自注意力机制模块以及所述条件随机场的参数进行调整迭代。

[0010] 在一些实施例中，对各科技论文数据进行序列标注采用BIO标注。

[0011] 在一些实施例中，所述word2vec模型采用科技论文数据进行预训练。

[0012] 另一方面，本发明提供一种命名实体识别方法，包括：

获取待处理的科技论文数据，将所述科技论文数据输入上述命名实体识别模型训练方法的目标命名实体识别模型中，输出命名实体识别结果。

[0013] 另一方面，本发明提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现上述方法的步骤。

[0014] 另一方面，本发明提供一种计算机可读存储介质，其上存储有计算机程序，其特征

在于，该程序被处理器执行时实现上述方法的步骤。

[0015] 本发明的有益效果至少是：

所述命名实体识别模型训练方法、识别方法及装置中，所述模型训练方法的初始神经网络模型由通过结合关键字符级别编码和词级别编码对科技论文数据进行向量表示，将字符级别向量和词级别向量引入双向长短期记忆网络能够挖掘上下文关系，同时挖掘关键词的语义特征，提升了分词边界的准确性；通过将字符级别向量引入自注意力机制模型，能够更高效地捕捉数据内部相关性，提升命名实体识别的准确率。

[0016] 本发明的附加优点、目的，以及特征将在下面的描述中将部分地加以阐述，且将对于本领域普通技术人员在研究下文后部分地变得明显，或者可以根据本发明的实践而获知。本发明的目的和其它优点可以通过在书面说明及其权利要求书以及附图中具体指出的结构实现到并获得。

[0017] 本领域技术人员将会理解的是，能够用本发明实现的目的和优点不限于以上具体所述，并且根据以下详细说明将更清楚地理解本发明能够实现的上述和其他目的。

附图说明

[0018] 此处所说明的附图用来提供对本发明的进一步理解，构成本申请的一部分，并不构成对本发明的限定。在附图中：

图1为本发明一实施例所述命名实体识别模型训练方法中初始神经网络模型工作逻辑示意图；

图2为本发明一实施例所述命名实体识别模型训练方法中字符级别向量和词级别向量连接结构结构示意图；

图3为本发明一实施例所述命名实体识别模型训练方法中初始神经网络模型结构示意图。

具体实施方式

[0019] 为使本发明的目的、技术方案和优点更加清楚明白，下面结合实施方式和附图，对本发明做进一步详细说明。在此，本发明的示意性实施方式及其说明用于解释本发明，但并不作为对本发明的限定。

[0020] 在此，还需要说明的是，为了避免因不必要的细节而模糊了本发明，在附图中仅仅示出了与根据本发明的方案密切相关的结构和/或处理步骤，而省略了与本发明关系不大的其他细节。

[0021] 应该强调，术语“包括/包含”在本文使用时指特征、要素、步骤或组件的存在，但并不排除一个或更多个其它特征、要素、步骤或组件的存在或附加。

[0022] 命名实体识别可以使用基于统计机器学习的方法，这首先需要用人工标注的语料进行有监督训练，然后利用训练好的机器学习模型实现预测。基于统计机器学习的模型有隐马尔可夫模型、最大熵模型、决策树、支持向量机等。但是，在基于统计机器学习的识别方法中，用于监督训练的数据需求巨大，需要人为的进行特征标注，代价太过昂贵，并且基于统计机器学习的方法对于不同形式或者领域的数据不能通用，泛化能力较差，相比于基于深度学习的识别方法，有较为明显的不足。

[0023] 基于深度学习的命名实体识别方式，主要包括卷积神经网络(CNN)和长短期记忆网络(LSTM)作为主体框架的识别方法。CNN主要用于处理英文，英文单词是由更细粒度的字母组成，这些字母潜藏着一些特征，但是在中文文本中，CNN的识别效果会受到一定的影响，同时对于序列标注任务来说，普通CNN卷积之后，为了覆盖上下文的序列信息，可能会导致卷积层数非常深，这样参数就会越来越多，模型庞大，难以训练。LSTM按照文本序列的输入处理上文的信息，而下文的信息对于科技学术会议论文数据的处理也有重要意义，也无法考虑到文本中的全局信息。与此同时，中文文本输入到长短期记忆网络中是按照字符为单位进行输入，但对于中文的命名实体识别来说，词语中同样蕴含着大量的语义信息，因此现有的技术并不能很好的充分挖掘语义中的文本信息，对于科技学术会议数据来说，分词阶段有可能产生错误，如将专有的技术词汇拆分成其他领域中的词汇，影响命名实体识别的准确率。

[0024] 所以，在中文命名实体识别过程中，大部分方法是基于字符级别编码，这种方式在通用领域的命名识别中取得了较好的效果，然而在学术论文数据中，由于专业词汇较多，采用这种方式很有可能产生错误的词语边界。仅采用字符级别编码无法挖掘到一串字符信息中的词级别的信息。

[0025] 为了解决这个问题，本发明引入论文关键词特征，提出关键词-字符编码方式，在编码阶段同时考虑到关键词级别和字符级别的语义信息，把字符级模型和词级别的模型相结合，降低歧义发生的概率。此外在LSTM+CRF(长短期记忆神经网络+条件随机场)为主体框架的基础上，在LSTM层引入自注意力机制，弥补长短期记忆网络无法考虑到全局信息的缺陷，最后将LSTM和注意力机制输出的结果进行融合再通过CRF进行标注，兼顾了字符之间的依赖关系，在论文数据集中取得了更好的识别效果。

[0026] 需要预先说明的是，字符级别编码是指将句子文本中的中文字符逐一单个进行向量化话，而词级别编码是将关键词进行整体的向量化。因此，对于一个句子中关键词内的字符就存在字符级别向量及其所属关键词的词级别向量。示例性的，对于句子“神经网络的文本分类”，按照字符级别可以拆分为“神、经、网、络、的、文、本、分、类”，同时也可以提取关键词“神经网络”和“文本分类”。

[0027] 一方面，本发明提供一种命名实体识别模型训练方法，包括步骤S101~S103：

步骤S101：获取多个科技论文数据，各科技论文数据包含一个或多个关键词，对各科技论文数据进行序列标注，以得到训练样本集。

[0028] 步骤S102：获取初始神经网络模型，初始神经网络模型对科技论文数据的各单个中文字符进行字符级别编码得到相应的字符级别向量、对科技论文数据的各关键词进行词级别编码得到相应的词级别向量；将各字符级别向量和各词级别向量进行连接后输入至双向长短期记忆网络，由双向长短期记忆网络输出第一特征向量；将各字符级别向量输入自注意力机制模块，将自注意力机制模块输出与原始的各字符级别向量连接得到第二特征向量；将第一特征向量与第二特征向量进行融合，并输入条件随机场后输出命名实体识别结果。

[0029] 步骤S103：采用训练样本集对初始神经网络模型进行训练，对双向长短期记忆网络、自注意力机制模块以及条件随机场的参数进行调整迭代，得到目标命名实体识别模型。

[0030] 在本实施例中，步骤S101首先配置训练样本集，以科技论文数据为样本主体，科技

论文数据可以是具有特定技术领域范围的科技学术会议数据，科技学术会议是一种以促进科学发展、学术交流、课题研究等学术性话题为主题的会议。学术会议一般都具有国际性、权威性、高知识性、高互动性等特点，学术会议会包含论文集，科技学术会议数据即指代其中的论文数据。每一个技学术会议的相关数据一般都是针对特定科技领域的，其以论文数据为主体，具有数据规模大，内容专业化，特征属性繁多的特点。在一些实施例中，构建训练样本可以以单个科技学术会议的数据为样本。为了提高通用性，也可以用多个不同科技学术会议的数据作为样本。具体的，样本可以为中文论文文本，其中，摘要部分记载有关键词，对中文论文文本进行序列标注，具体可以采用BIO标注法进行标注。

[0031] 在步骤S102中，构建了结合字符级别向量和词级别向量，并且联合应用双向长短期记忆神经网络和自注意力机制进行特征挖掘。具体的，参照图1和图3，将样本中的科技学术会议论文数据分别输入关键词级别嵌入层和字符级别嵌入层，分别进行词级别编码和字符级别编码。在一些实施例中，初始神经网络模型采用word2vec模型获取各单个中文字符对应的字符级别向量以及各关键词对应的词级别向量。在一些实施例中，word2vec模型可以采用科技论文数据进行预训练，以适应科技论文数据使用场景的需求。进一步地，对于一个句子文本，其中每个字符都经字符级别编码得到相应的字符级别向量，而论文摘要中的关键词经词级别编码得到词级别向量。

[0032] 具体的，对于科技论文数据，$s=[c_1,c_2,c_3,...,c_n]$ 可以表示为，其中ci表示句子中的第i个字符，每个字符经字符级别编码得到字符级别向量 $x^c=[x_1^c,x_2^c,x_3^c,...,x_n^c]$ ，表达式为：

$$x_i^c=e^c(c_i) \quad (1)$$

其中，e^c代表字符级别向量表示。

[0033] 对科技论文数据按照中文分词方式进行切分，得到n个词汇，表示为 $s=[w_1,w_2,w_3,...,w_n]$ ，每个关键词经词级别编码得到词级别向量 $x^w=[x_1^w,x_2^w,x_3^w,...,x_n^w]$ ，表达式为：

$$x_i^w=e^w(w_i) \quad (2)$$

其中，e^w代表词级别向量表示。

[0034] 如图2所示，对于文本“神经网络的文本分类”，可以按序标记为 x_1^c 至 x_9^c ，相应的，9个中文字符对应的字符级别向量分别为c_1~c_9，句中包括“神经网络”和“文本分类”两个关键词分别记录为 $\mathrm{W}_{1,4}^{\mathrm{k}}$ 和 $\mathrm{W}_{6,9}^{\mathrm{k}}$ ，两个关键词的词级别向量 $\mathrm{X}_{1,4}^{\mathrm{w}}$ 和 $\mathrm{X}_{6,9}^{\mathrm{w}}$ ，计算式可表示为：

$$x_{b,e}^{w}=e^{w}(w^{k}_{b,e}) \quad (3)$$

其中，e^{w}代表词级别向量表示，$w^{k}_{b,e}$为第b个字符至第e个字符构成的关键词，$x_{b,e}^{w}$表示第b个字符至第e个字符构成关键词的词级别向量。

[0035] 进一步地，将字符级别向量和词级别向量融合输入至双向长短期记忆神经网络，在考虑上下文信息的前提下，挖掘语义特征。示例性的，参照图2，字符级别向量c_1~c_9分别与相应的词向量进行连接融合得到得到特征序列h_1~h_9。

[0036] 具体的，步骤S102中，将各字符级别向量和各词级别向量进行连接后输入至双向长短期记忆网络，包括：将单个字符对应的字符级别向量和词级别向量进行归一化求和得到该字符对应的第一输入序列，并输入至双向长短期记忆网络，计算式为：

$$c_j^{c}=\sum_{b}\alpha_j^{c}\otimes\tilde{c}_j^{c}+\alpha_{b.j}^{w}\otimes c_{b,j}^{w} \quad (4)$$

其中，c_j^{c}为第j个字符对应的第一输入序列，$\tilde{c}_j^{c}$为第j个字符对应的字符级别向量，α_j^{c}为$\tilde{c}_j^{c}$对应的归一化系数，$c_{b,j}^{w}$为第j个字符所属关键词的词级别向量，$\alpha_{b.j}^{w}$为$c_{b,j}^{w}$的归一化系数，b为第j个字符所属关键词的序数。

[0037] 同时，将各字符级别向量输入至自注意力机制模块，以挖掘全局特征。

[0038] 由双向长短期记忆网络输出第一特征向量，将自注意力机制模块输出与原始的各字符级别向量连接得到第二特征向量，将第一特征向量与第二特征向量进行融合并输入条件随机场，以输出命名实体识别结果。

[0039] 在一些实施例的步骤102中，将第一特征向量与第二特征向量进行融合，包括：

将第一特征向量与第二特征向量进行归一化求和，计算式为：

$$y_i=\alpha_i^{h}\otimes h_i+\alpha_i^{a}\otimes a_i \quad (5)$$

$$\alpha_i^{h}=\frac{e^{h_i}}{e^{h_i}+e^{a_i}} \quad (6)$$

$$\alpha_i^{a}=\frac{e^{a_i}}{e^{h_i}+e^{a_i}} \quad (7)$$

其中，y_i为科技论文数据第i个字符的特征值，h_i为科技论文数据第i个字符经双向长短期记忆网络输出的特征值，a_i为科技论文数据第i个字符经自注意力机制模块

输出的特征值，α_i^h 为 h_i 的归一化系数，α_i^a 为 a_i 的归一化系数；e为自然底数。

[0040] 在步骤S103中，基于步骤S101中的训练样本集对步骤S102构件的初始神经网络模型进行训练和迭代。

[0041] 在一些实施例中，采用训练样本集对初始神经网络模型进行训练，包括：采用交叉熵函数作为损失函数，对双向长短期记忆网络、自注意力机制模块以及条件随机场的参数进行调整迭代。

[0042] 另一方面，本发明提供一种命名实体识别方法，包括步骤S201：

步骤S201：获取待处理的科技论文数据，将科技论文数据输入上述步骤S101~S103所述命名实体识别模型训练方法的目标命名实体识别模型中，输出命名实体识别结果。

[0043] 另一方面，本发明提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现上述方法的步骤。

[0044] 另一方面，本发明提供一种计算机可读存储介质，其上存储有计算机程序，其特征在于，该程序被处理器执行时实现上述方法的步骤。

[0045] 下面结合以具体实施例对本发明进行说明：

本实施例中提供一种初始神经网络模型，如图3所示，包括嵌入层、BiLSTM-SA网络层(双向长短期记忆网络和自注意力机制模块联合的网络层)以及CRF层(条件随机场)。该初始神经网路模型经训练样本集进行训练得到目标命名实体识别模型，训练样本集的每个样本包括一个科技论文，并以BIO标注作为标签。

[0046] 第一部分，在嵌入层冲，基于关键词-字符级别编码模型，对科技论文数据进行向量化：

基于字符级别编码模型对每一个中文字符逐一进行编码，可以采用word2vec模型，给定一个论文标题文本序列的示例为“基于神经网络的文本分类”，可以将其表示为 $s=[c_1,c_2,c_3,...,c_n]$，其中，c_i 表示句子中的第 i 个字符，每个字符经过公式(1)的变换，获得对应的输入字符级别向量 $x^c=[x_1^c,x_2^c,x_3^c,...,x_n^c]$ 。

[0047] 基于词级别编码模型对中文词汇的关键词进行编码，可以采用word2vec模型，同样给定文本序列的示例为“基于神经网络的文本分类”，按照常规的中文分词方式对其进行切分，然后按照词级别进行编码，句中包括“神经网络”和“文本分类”两个关键词分别记录为 $\mathrm{W}_{1,4}^{\mathrm{k}}$ 和 $\mathrm{W}_{6,9}^{\mathrm{k}}$ ，两个关键词的词级别向量 $\mathrm{X}_{1,4}^{\mathrm{w}}$ 和 $\mathrm{X}_{6,9}^{\mathrm{w}}$ ，计算式可表示为：

$$x_{b,e}^{w}=e^{w}(w^{k}{}_{b,e}) \quad (3)$$

其中，e^w 代表词级别向量表示，$w^k{}_{b,e}$ 为第b个字符至第e个字符构成的关键词，$x_{b,e}^w$ 表示第b个字符至第e个字符构成关键词的词级别向量。

[0048] 关键词-字符编码模型主要考虑到了科技学术会议中论文数据本身的特点。由于论文数据专业性强，因此常规的分词方式并不适用于论文数据集，如果采用基本的字词融合，可能会产生很多错误的边界，影响识别准确率。考虑到论文数据集中有关键词这一特

征，例如对于文本序列：基于神经网络的文本分类模型，在关键词字段中包含了：神经网络，文本分类等词汇，如果不考虑关键词信息，该句会被切分为：

$$s = 基于|神经|网络|的|文本|分类$$

对于本文想要识别的实体，显然产生了错误的词汇边界，因此要引入关键词特征，构建词典，对于例子中的文本序列，需要将其正确切分为：

$$s = 基于|神经网络|的|文本分类$$

第二部分，BiLSTM-SA网络层，融合双向长短期记忆网络和自注意力机制：

LSTM是一种特殊的RNN，与传统的RNN相比，LSTM同样是基于本层输入x_t和上一层输出h_{t-1}来计算本层输出h_t，但加入了输入门i_t、遗忘门f_t以及输出门o_t三个门和一个内部记忆单元c_t。

[0049] 第t层的更新计算公式为计算式8~13：

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (8)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (9)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (10)$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1}) \quad (11)$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t \quad (12)$$

$$h_t = o_t \otimes \tanh(c_t) \quad (13)$$

LSTM模型按照文本序列的输入处理上文的信息，而下文的信息对于科技学术会议论文数据的处理也有重要意义，因此，本实施例中采用BiLSTM，它由两层LSTM组成，向量表示层得到的向量按照正序作为正向LSTM的输入，即可以得到输出序列：

$$h_L = [h_{L1}, h_{L2}, h_{L3}, \ldots, h_{Ln}] \quad ;$$

再通过反向输入的方式，得到逆向LSTM输出序列：

$$h_R = [h_{R1}, h_{R2}, h_{R3}, \ldots, h_{Rn}] \quad ;$$

将两层的输出进行融合，得到包含上下文的特征$h_n = [h_L, h_R]$ 。

[0050] 具体的，将本实施例第一部分中的单个字符对应的字符级别向量和词级别向量进行归一化求和得到该字符对应的第一输入序列，并输入至双向长短期记忆网络，并最终得到序列h_1~h_n。

[0051] 第一输入序列的计算式为：

$$c_j^c = \sum_b \alpha_j^c \otimes \tilde{c}_j^c + \alpha_{b,j}^w \otimes c_{b,j}^w \quad (4)$$

其中，c_j^c 为第j个字符对应的第一输入序列，$\tilde{c}_j^c$为第j个字符对应的字符级别向量，α_j^c 为$\tilde{c}_j^c$对应的归一化系数，$c_{b,j}^w$ 为第j个字符所属关键词的词级别向量，$\alpha_{b,j}^w$ 为$c_{b,j}^w$的归一化系数，b为第j个字符所属关键词的序数。

[0052] BiLSTM在可以考虑到上下文的信息，但对于全局信息无法充分的表达，因此本模型将Self Attention机制作为BiLSTM模块的补充，提高命名实体识别的准确率。

[0053] Attention的计算如公式(14)所示。Q、K和V三个矩阵均来自同一输入，首先计算Q与K之间的点乘，然后除以一个尺度标度d_k，然后将其结果归一化，再乘以矩阵V就得到权重求和的表示。由于Attention本身就考虑到了全局的输入，因此直接利用字符级别编码进行输入。

[0054] $$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \quad (14)$$

Attention层的输入为字符编码的向量，输出通过式(14)的计算后，输出为$a=[a_1,a_2,a_3,...,a_n]$，然后对这两个输出进行融合操作，得到y_n后将其输入到CRF层中，获得命名实体识别最大概率的分类。

[0055] 具体的，假设BiLSTM-SA网络层的输出为y_n，在进行融合操作时采用归一化求和的形式，即：

$$y_i = \alpha_i^h \otimes h_i + \alpha_i^a \otimes a_i \quad (5)$$

$$\alpha_i^h = \frac{e^{h_i}}{e^{h_i} + e^{a_i}} \quad (6)$$

$$\alpha_i^a = \frac{e^{a_i}}{e^{h_i} + e^{a_i}} \quad (7)$$

其中，y_i为科技论文数据第i个字符的特征值，h_i为科技论文数据第i个字符经双向长短期记忆网络输出的特征值，a_i为科技论文数据第i个字符经自注意力机制模块输出的特征值，α_i^h为h_i的归一化系数，α_i^a为a_i的归一化系数；e为自然底数。

[0056] 第三部分，CRF层(条件随机场层)，进行序列标注，获得命名实体识别最大概率的分类。

[0057] 在预测当前标签时，CRF通常可以产生更高的标记精度。由于论文数据相邻字符之

间有较强的依赖关系，因此，在模型的最后一层，利用CRF来对前序层中得到的融合特征信息进行解码。

[0058] 获得LSTM+SA层的序列输出为 $y = [y_1, y_2, y_3, \cdots y_n]$ ，是输入文本的多个可能的标注序列，CRF的标记过程为：

$$S(x,y) = \sum_{i=1}^{N}\left(O_{i,y_i} + T_{y_{i-1},y_i}\right) \quad (15)$$

式中，$S(x,y)$ 为每个标注序列的评分，O_{i,y_i} 表示第 i 个单词标记为 y_i 个标签的概率，矩阵 T 是转移矩阵，$T_{i,j}$ 表示由标签 i 转移到标签 j 的概率，CRF在原语句为S的条件下，产生标记序列的概率为公式(15)。

[0059] 标记序列的似然公式为如下公式(16)：

$$p(y|S) = \frac{e^{S(X,y)}}{\sum_{\tilde{y}\in Y_x} s(X,\tilde{y})} \quad (16)$$

$$\log\left(p(y|S)\right) = s(X,y) - \log\left(\sum_{\tilde{y}\in Y_x} e^{s(X,\tilde{y})}\right) \quad (17)$$

$$y^* = argmax\,[s(X,\tilde{y})] \quad (18)$$

式中，Y_x 表示所有可能的标记集合，最终的解码阶段通过标准 Viterbi 算法，求解最大的概率，最后通过公式(18)预测出最优的命名实体识别序列 y^* 。

[0060] 本实施例针对科技学术会议数据，提出了结合关键词-字符、BiLSTM和自注意力机制的命名实体识别算法，整体模型由向量表示层、双向长短期记忆网络-自注意力层和条件随机场层构成。该算法可以挖掘到文本中潜在的语义信息，减少了中文分词的边界问题带来的识别错误，同时考虑到了全局的文本信息，可以对科技学术会议中论文数据的命名实体进行有效的识别，提高命名实体识别的准确率和召回率。基于识别出的命名实体，结合论文数据中结构化的数据获取到的关联关系可以对学术会议数据构建精准画像，为科研人员进行科研信息的获取以及进行科研决策提供良好的数据支撑。

[0061] 综上所述，所述命名实体识别模型训练方法、识别方法及装置中，所述模型训练方法的初始神经网络模型由通过结合关键字符级别编码和词级别编码对科技论文数据进行向量表示，将字符级别向量和词级别向量引入双向长短期记忆网络能够挖掘上下文关系，同时挖掘关键词的语义特征，提升了分词边界的准确性；通过将字符级别向量引入自注意力机制模型，能够更高效地捕捉数据内部相关性，提升命名实体识别的准确率。

[0062] 本领域普通技术人员应该可以明白，结合本文中所公开的实施方式描述的各示例性的组成部分、系统和方法，能够以硬件、软件或者二者的结合来实现。具体究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本发明的

范围。当以硬件方式实现时，其可以例如是电子电路、专用集成电路(ASIC)、适当的固件、插件、功能卡等等。当以软件方式实现时，本发明的元素是被用于执行所需任务的程序或者代码段。程序或者代码段可以存储在机器可读介质中，或者通过载波中携带的数据信号在传输介质或者通信链路上传送。“机器可读介质”可以包括能够存储或传输信息的任何介质。机器可读介质的例子包括电子电路、半导体存储器设备、ROM、闪存、可擦除ROM(EROM)、软盘、CD-ROM、光盘、硬盘、光纤介质、射频(RF)链路，等等。代码段可以经由诸如因特网、内联网等的计算机网络被下载。

[0063] 还需要说明的是，本发明中提及的示例性实施例，基于一系列的步骤或者装置描述一些方法或系统。但是，本发明不局限于上述步骤的顺序，也就是说，可以按照实施例中提及的顺序执行步骤，也可以不同于实施例中的顺序，或者若干步骤同时执行。

[0064] 本发明中，针对一个实施方式描述和/或例示的特征，可以在一个或更多个其它实施方式中以相同方式或以类似方式使用，和/或与其他实施方式的特征相结合或代替其他实施方式的特征。

[0065] 以上所述仅为本发明的优选实施例而已，并不用于限制本发明，对于本领域的技术人员来说，本发明实施例可以有各种更改和变化。凡在本发明的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

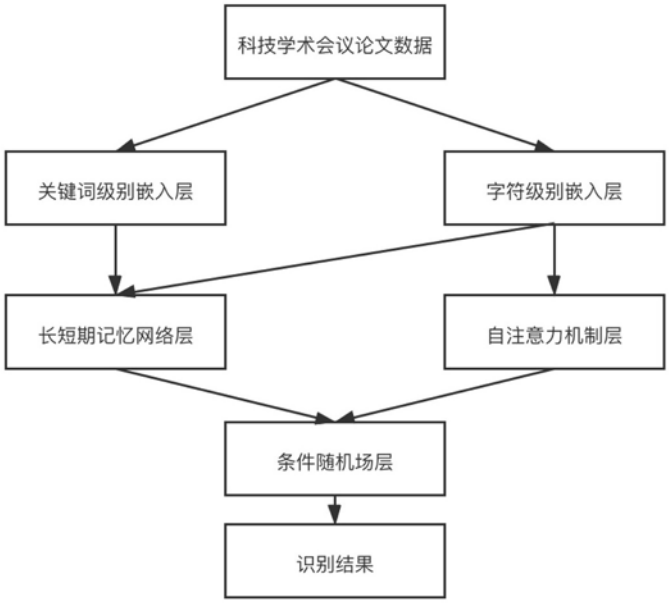

科技学术会议论文数据
关键词级别嵌入层
字符级别嵌入层
长短期记忆网络层
自注意力机制层
条件随机场层
识别结果

图1

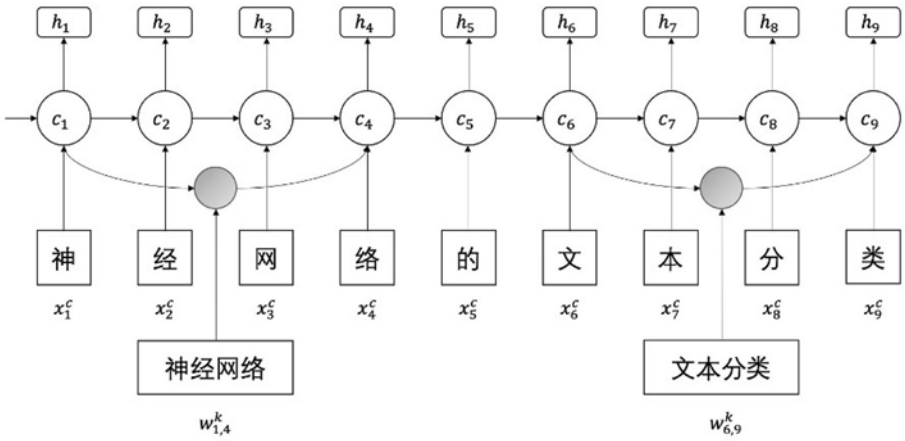

h_1 h_2 h_3 h_4 h_5 h_6 h_7 h_8 h_9
c_1 c_2 c_3 c_4 c_5 c_6 c_7 c_8 c_9
神 经 网 络 的 文 本 分 类
x_1^c x_2^c x_3^c x_4^c x_5^c x_6^c x_7^c x_8^c x_9^c
神经网络
文本分类
$w_{1,4}^k$
$w_{6,9}^k$

图2

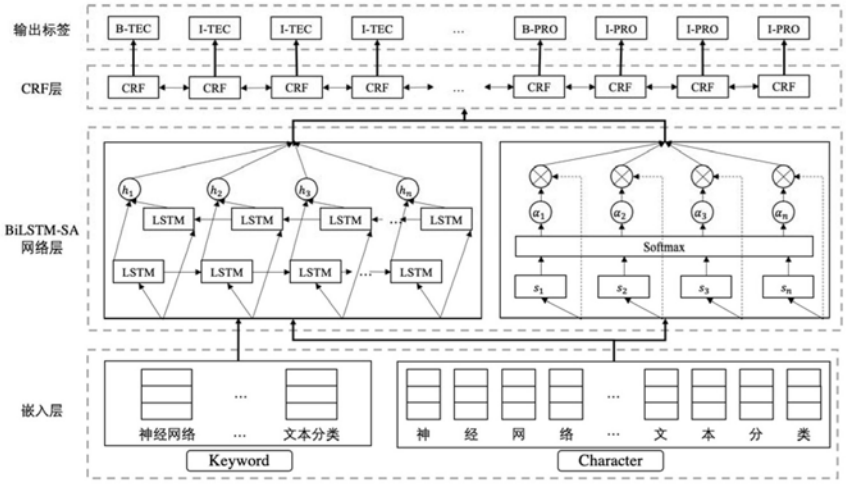

输出标签
B-TEC
I-TEC
I-TEC
I-TEC
...
B-PRO
I-PRO
I-PRO
I-PRO
CRF层
CRF
CRF
CRF
CRF
...
CRF
CRF
CRF
CRF
BiLSTM-SA
网络层
LSTM
Softmax
嵌入层
神经网络
...
文本分类
Keyword
神
经
网
络
...
文
本
分
类
Character

图3