(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115687607 A
(43)申请公布日 2023.02.03

(21)申请号 202110850152.0

(22)申请日 2021.07.27

(71)申请人 中移系统集成有限公司
地址 050000 河北省石家庄市青园街220号
申请人 中移雄安信息通信科技有限公司
中国移动通信集团有限公司

(72)发明人 雷泽华

(74)专利代理机构 北京路浩知识产权代理有限公司 11002
专利代理师 李文清

(51)Int.Cl.
G06F 16/35(2019.01)
G06V 30/14(2022.01)
G06V 30/19(2022.01)
G06V 10/80(2022.01)

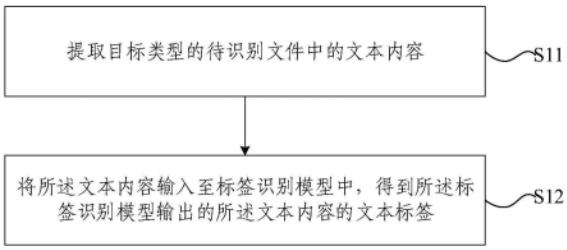
权利要求书2页 说明书11页 附图4页

(54)发明名称

文本标签识别方法及系统

(57)摘要

本发明提供一种文本标签识别方法及系统，首先提取目标类型的待识别文件中的文本内容；然后将所述文本内容输入至标签识别模型中，得到所述标签识别模型输出的所述文本内容的文本标签。采用的标签识别模型中，特征提取层可以对文本内容进行多粒度特征提取，可以从不同层次对文本内容进行描述，进而使得通过不同粒度的特征向量融合得到的文本特征向量更准确的描述文本内容，使得最终通过文本特征向量得到的分类结果更加准确，提高了分类的准确率。



CN 115687607 A

1.一种文本标签识别方法,其特征在于,包括:

提取目标类型的待识别文件中的文本内容;

将所述文本内容输入至标签识别模型中,得到所述标签识别模型输出的所述文本内容的文本标签;

其中,所述标签识别模型包括特征提取层和分类层,所述特征提取层用于对所述文本内容进行多粒度特征提取,得到不同粒度的特征向量,并将不同粒度的特征向量进行融合,得到文本特征向量;所述分类层用于对所述文本特征向量进行分类;所述标签识别模型基于所述目标类型的多个文本样本训练得到。

2.根据权利要求1所述的文本标签识别方法,其特征在于,所述特征提取层具体用于:

基于BRET模型,确定所述文本内容的字向量;

基于所述字向量,确定文本内容的字粒度特征向量、词粒度特征向量以及词条粒度特征向量;

基于命名实体识别模型,确定所述文本内容的命名实体粒度特征向量;

将所述字粒度特征向量、所述词粒度特征向量、所述词条粒度特征向量以及所述命名实体粒度特征向量进行融合,得到所述文本特征向量。

3.根据权利要求2所述的文本标签识别方法,其特征在于,所述命名实体识别模型为基于BiLSTM和CRF的命名实体识别模型。

4.根据权利要求1所述的文本标签识别方法,其特征在于,所述方法还包括:

若判断获知多个文本样本中存在目标文本样本,所述目标文本样本未携带有文本样本标签,则基于预先确定的所述目标类型的标签集合,确定所述目标文本样本的文本样本标签。

5.根据权利要求4所述的文本标签识别方法,其特征在于,所述标签集合基于如下方法确定:

基于聚类算法,对多个文本样本对应的文本特征向量进行聚类,生成多个类簇;

基于各簇类中包含的文本特征向量,确定备选标签集合;

基于所述备选标签集合,确定所述标签集合。

6.根据权利要求1-5中任一项所述的文本标签识别方法,其特征在于,所述待识别文件的文件格式包括excel文件、csv文件、word文件及pdf文件。

7.根据权利要求6所述的文本标签识别方法,其特征在于,所述提取目标类型的待识别文件中的文本内容,之后还包括:

对所述文本内容进行数据清洗。

8.一种文本标签识别装置,其特征在于,包括:

提取模块,用于提取目标类型的待识别文件中的文本内容;

识别模块,用于将所述文本内容输入至标签识别模型中,得到所述标签识别模型输出的所述文本内容的文本标签;

其中,所述标签识别模型包括特征提取层和分类层,所述特征提取层用于对所述文本内容进行多粒度特征提取,得到不同粒度的特征向量,并将不同粒度的特征向量进行融合,得到文本特征向量;所述分类层用于对所述文本特征向量进行分类;所述标签识别模型基于所述目标类型的多个文本样本训练得到。

9.一种电子设备,包括存储器、处理器及存储在所述存储器上并可在所述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7任一项所述文本标签识别方法的步骤。

10.一种非暂态计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现如权利要求1至7任一项所述文本标签识别方法的步骤。

文本标签识别方法及系统

技术领域

[0001] 本发明涉及文本分类技术领域,尤其涉及一种文本标签识别方法及系统。

背景技术

[0002] 随着机器的智能化发展,通过机器进行文本分类对于文本的应用至关重要。在很多场景下,都涉及到文本分类。比如说,在网络论坛中,需要对用户发表的帖子进行分类,以在对应标签(如家庭情感)的论坛版块中,对该帖子进行展示。

[0003] 文本分类,即文本标签识别。现有技术中,在对文本标签进行识别时,通常先对文本进行分词,去停用词等操作后,进行特征提取,随后使用随机森林构建分类器,进行标签识别。由于现有技术中在进行特征提取时,通常无法提取出多粒度特征,这将导致提取出的特征用于标签识别时的识别准确率不高。

[0004] 为此,现急需提供一种文本标签识别方法。

发明内容

[0005] 本发明提供一种文本标签识别方法及系统,用以解决现有技术中存在的缺陷。

[0006] 本发明提供一种文本标签识别方法,包括:

[0007] 提取目标类型的待识别文件中的文本内容;

[0008] 将所述文本内容输入至标签识别模型中,得到所述标签识别模型输出的所述文本内容的文本标签;

[0009] 其中,所述标签识别模型包括特征提取层和分类层,所述特征提取层用于对所述文本内容进行多粒度特征提取,得到不同粒度的特征向量,并将不同粒度的特征向量进行融合,得到文本特征向量;所述分类层用于对所述文本特征向量进行分类;所述标签识别模型基于所述目标类型的多个文本样本训练得到。

[0010] 根据本发明提供的一种文本标签识别方法,所述特征提取层具体用于:

[0011] 基于BRET模型,确定所述文本内容的字向量;

[0012] 基于所述字向量,确定文本内容的字粒度特征向量、词粒度特征向量以及词条粒度特征向量;

[0013] 基于命名实体识别模型,确定所述文本内容的命名实体粒度特征向量;

[0014] 将所述字粒度特征向量、所述词粒度特征向量、所述词条粒度特征向量以及所述命名实体粒度特征向量进行融合,得到所述文本特征向量。

[0015] 根据本发明提供的一种文本标签识别方法,所述命名实体识别模型为基于BiLSTM和CRF的命名实体识别模型。

[0016] 根据本发明提供的一种文本标签识别方法,所述方法还包括:

[0017] 若判断获知多个文本样本中存在目标文本样本,所述目标文本样本未携带有文本样本标签,则基于预先确定的所述目标类型的标签集合,确定所述目标文本样本的文本样本标签。

[0018] 根据本发明提供的一种文本标签识别方法，所述标签集合基于如下方法确定：

[0019] 基于聚类算法，对多个文本样本对应的文本特征向量进行聚类，生成多个类簇；

[0020] 基于各簇类中包含的文本特征向量，确定备选标签集合；

[0021] 基于所述备选标签集合，确定所述标签集合。

[0022] 根据本发明提供的一种文本标签识别方法，所述待识别文件的文件格式包括excel文件、csv文件、word文件及pdf文件。

[0023] 根据本发明提供的一种文本标签识别方法，所述提取目标类型的待识别文件中的文本内容，之后还包括：

[0024] 对所述文本内容进行数据清洗。

[0025] 本发明还提供一种文本标签识别装置，包括：

[0026] 提取模块，用于提取目标类型的待识别文件中的文本内容；

[0027] 识别模块，用于将所述文本内容输入至标签识别模型中，得到所述标签识别模型输出的所述文本内容的文本标签；

[0028] 其中，所述标签识别模型包括特征提取层和分类层，所述特征提取层用于对所述文本内容进行多粒度特征提取，得到不同粒度的特征向量，并将不同粒度的特征向量进行融合，得到文本特征向量；所述分类层用于对所述文本特征向量进行分类；所述标签识别模型基于所述目标类型的多个文本样本训练得到。

[0029] 本发明还提供一种电子设备，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如上述任一种所述文本标签识别方法的步骤。

[0030] 本发明还提供一种非暂态计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现如上述任一种所述文本标签识别方法的步骤。

[0031] 本发明提供的文本标签识别方法及系统，首先提取目标类型的待识别文件中的文本内容；然后将所述文本内容输入至标签识别模型中，得到所述标签识别模型输出的所述文本内容的文本标签。采用的标签识别模型中，特征提取层可以对文本内容进行多粒度特征提取，可以从不同层次对文本内容进行描述，进而使得通过不同粒度的特征向量融合得到的文本特征向量更准确的描述文本内容，使得最终通过文本特征向量得到的分类结果更加准确，提高了分类的准确率。

附图说明

[0032] 为了更清楚地说明本发明或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0033] 图1是本发明提供的文本标签识别方法的流程示意图之一；

[0034] 图2是本发明提供的不同文件格式的待识别文件的文本内容提取方法的流程示意图；

[0035] 图3是本发明提供的标签识别模型中特征提取层的工作流程示意图；

[0036] 图4是本发明提供的命名实体识别模型的结构示意图；

[0037] 图5是本发明提供的文本标签识别方法的流程示意图之二；

[0038] 图6是本发明提供的文本标签识别装置的结构示意图；

[0039] 图7是本发明提供的电子设备的结构示意图。

具体实施方式

[0040] 为使本发明的目的、技术方案和优点更加清楚，下面将结合本发明中的附图，对本发明中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。

[0041] 图1为本发明实施例中提供的一种文本标签识别方法的流程示意图，如图1所示，该方法包括：

[0042] S1，提取目标类型的待识别文件中的文本内容；

[0043] S2，将所述文本内容输入至标签识别模型中，得到所述标签识别模型输出的所述文本内容的文本标签；

[0044] 其中，所述标签识别模型包括特征提取层和分类层，所述特征提取层用于对所述文本内容进行多粒度特征提取，得到不同粒度的特征向量，并将不同粒度的特征向量进行融合，得到文本特征向量；所述分类层用于对所述文本特征向量进行分类；所述标签识别模型基于所述目标类型的多个文本样本训练得到。

[0045] 具体地，本发明实施例中提供的文本标签识别方法，其执行主体为服务器，该服务器可以是本地服务器，也可以是云端服务器，本地服务器具体可以是计算机等，本发明实施例中对此不作具体限定。

[0046] 首先执行步骤S1，提取目标类型的待识别文件中的文本内容。目标类型是指待识别文件的类型，例如可以是政务文件，也可以是舆情文件、社交网络文件等，本发明实施例中对此不作具体限定。待识别文件中的文本内容是指待识别文件中以文本形式表示的内容。

[0047] 如图2所示，待识别文件的文件格式可以包括excel文件、csv文件、word文件及pdf文件等。对于不同的文件格式，提取文本内容的方式也有所不同。excel文件和csv文件均具备较高的规整性，即文件信息包含着csv文件名、表格属性信息以及内部标题和文本，因此可以直接提取到文本内容。word格式是常见的文本格式，利用python-docx模块可以将word文件中的段落、文本、字体等都看作对象进行提取。对于word文件中的表格内容，python-docx同样可以实现提取，利用.tables函数可以获取word文件中所有表格对象的列表，随后进行提取以获取表格样式、单元格对象及文字信息。对于word文件中的图片，可以利用光学字符识别(Optical Character Recognition，OCR)技术提取文本内容。对于pdf文件，通常需要分为两种类型，第一种为文本转化而成的pdf文件，即文本pdf；另一种为通过扫描得到的图片插入制成的pdf，即图片pdf。针对文本pdf，可以使用PyPDF2、pdfminer、textract、slate等库提取文本；使用pdfplumber、camelot等库提取表格；针对图片pdf，可以先将图片pdf转化为图片，之后再利用OCR技术提取内容，如pytesseract库。

[0048] 然后执行步骤S2，将文本内容输入至标签识别模型中，通过标签识别模型对文本内容进行识别，得到文本内容的文本标签。文本标签可以是文本内容中的关键信息，也可以

是关键信息的概括总结性信息。文本标签可以为一个，也可以为多个，本发明实施例中对此不作具体限定。

[0049] 本发明实施例中，采用的标签识别模型可以为神经网络模型，其结构上可以包括特征提取层和分类层，通过特征提取层可以对文本内容进行多粒度特征提取，得到不同粒度的特征向量，并将不同粒度的特征向量进行融合，得到文本特征向量。其中，不同粒度可以包括字粒度、词粒度、词条粒度以及命名实体粒度等。相应地，不同粒度的特征向量可以分别为字粒度特征向量、词粒度特征向量、词条粒度特征向量以及命名实体特征向量等。将不同粒度的特征向量进行融合，可以得到文本特征向量。文本特征向量可以是用于描述文本内容的特征向量，通过不同粒度的特征向量融合得到的文本特征向量描述文本内容，可以有效减少文本信息的丢失。通过分类层可以实现对文本特征向量进行分类，分类层可以通过分类器实现文本标签的输出。

[0050] 本发明实施例中，标签识别模型可以通过目标类型的多个文本样本训练得到，文本样本中可以包含有多个。文本样本可以具有多个文本样本标签，文本样本标签可以通过文本样本携带，也可以通过标注得到。

[0051] 若用(x,y)表示任意一个具有文本样本标签的文本样本，其中$y=(y1,y2,...,yq)\in\{-1,+1\}^q$，q为文本样本标签的数量。训练标签识别模型的过程，可以等价于学习一个联合概率分布p(x,y)，用$H_p(x,y)$表示给定联合概率分布p(x,y)时(x,y)的信息熵，信息熵最大时对应的标签识别模型即是可以应用的模型。即有：

[0052] $\max_p H_p(x,y)$

[0053] $E_p[f_k(x,y)]=F_k(k\in K)$

[0054] 其中，$f_k(x,y)$为特征函数，描述x和y之间的一个事实k，满足这个事实k时返回1，否则返回0。约束的目的是希望联合概率分布p(x,y)上，特征函数的期望能够等于一个希望值F_k，这个值通常通过训练集来估计，随后对优化问题进行求解。

[0055] 本发明实施例中提供的文本标签识别方法，首先提取目标类型的待识别文件中的文本内容；然后将所述文本内容输入至标签识别模型中，得到所述标签识别模型输出的所述文本内容的文本标签。采用的标签识别模型中，特征提取层可以对文本内容进行多粒度特征提取，可以从不同层次对文本内容进行描述，进而使得通过不同粒度的特征向量融合得到的文本特征向量更准确的描述文本内容，使得最终通过文本特征向量得到的分类结果更加准确，提高了分类的准确率。

[0056] 在上述实施例的基础上，本发明实施例中提供的文本标签识别方法，所述特征提取层具体用于：

[0057] 基于BRET模型，确定所述文本内容的字向量；

[0058] 基于所述字向量，确定文本内容的字粒度特征向量、词粒度特征向量以及词条粒度特征向量；

[0059] 基于命名实体识别模型，确定所述文本内容的命名实体粒度特征向量；

[0060] 将所述字粒度特征向量、所述词粒度特征向量、所述词条粒度特征向量以及所述命名实体粒度特征向量进行融合，得到所述文本特征向量。

[0061] 具体地，本发明实施例中，特征提取层可以包括特征提取模块、命名实体识别模块和特征融合模块。在得到文本特征向量的过程中，特征提取模块可以实现从文本到向量的

映射,命名实体识别模块可以实现对于文本内容中命名实体的识别和提取,特征融合模块可以将特征提取模块以及命名实体识别模块得到的特征向量进行融合,进而得到文本特征向量。

[0062] 特征提取模块可以根据BRET(Bidirectional Encoder Representation from Transformers)模型,确定文本内容的字向量。BERT起源于预训练的上下文表示学习,包括半监督序列学习(Semi-supervised Sequence Learning),生成预训练(Generative Pre-Training)。与之前的模型不同,BERT是一种深度双向的、无监督的语言表示,且仅使用纯文本语料库进行预训练的模型。

[0063] 特征提取过程中需要考虑其本身特性。首先选择的短词条文本数据为数据的文件名,其所具备信息量较大,绝大多数数据的标签划分也和文件名息息相关;其次文本内容中,可以考虑谈及其内部特征作为短词条数据,例如:标题、表格属性信息、文件最后署名信息等;最后将其余的文本内容数据作为研究对象进行特征提取。

[0064] 特征提取考虑选择预训练模型实现对文本内容的表示。BERT是一种预训练语言表示的方法,其在大量文本语料(维基百科)上训练了一个通用的“语言理解”模型,然后该模型可以去执行其他下游NLP任务。BERT比之前的预训练方法表现更出色,因为它是第一个用在预训练NLP上的无监督的、深度双向系统。无监督意味着BERT只需要用纯文本语料来训练,这点非常重要,因为海量的文本语料可以在各种语言的网络公开得到。预训练表示可以是上下文无关的,也可以是上下文相关的,而且,上下文相关的表示可以是单向的或双向的。上下文相关的模型会基于句子中的其他词生成每一个词的表示。BERT是建立在最近的预训练相关表示工作例如ELMO和GPT之上,但是这些模型都是单向的或浅双向的,这意味着每个词只会和它左边或右边的词相关。而BERT的这种自编码器的形式可以有效解决这个问题。本发明实施例中,使用BERT模型进行文本内容的特征提取,首先在海量数据中抽取短词条文本数据和内容数据,在其中各随机抽取10%的数据进行BERT模型的微调(fune-tuning),然后将依据fune-tuning后的权重将文本内容的特征全部提取到某一中间文件中。

[0065] 文本内容将在分词后生成词向量,随后这些词向量将会送入12层transformer编码层进行编码,BERT模型将会对15%的词进行随机掩码并基于周围词来预测掩码词,从而完成fune-tuning。

[0066] 然后,根据字向量,可以确定文本内容的字粒度特征向量、词粒度特征向量以及词条粒度特征向量。

[0067] 文本内容中的字粒度特征向量的提取可以通过利用BERT模型生成中文的字向量之后,将文本内容中每个字所对应的字向量按照顺序拼接起来得到文本内容的字向量矩阵,即字粒度特征向量:

[0068] $z_{1:m}=[z_1,z_2,z_3,\ldots,z_p,\ldots z_m]^T\in R^{m\times r}$

[0069] 其中,m是文本内容中字的数量,r表示向量的维度,z_p表示第p个字的字向量。

[0070] 文本内容中的词粒度特征向量的提取可以通过利用BERT模型生成中文的字向量之后,随后把词语中的每个字向量进行累计求平均实现,随后通过构建当前词的上下文语境,将词表示为具有相同维度的低维稠密词向量。将每个词对应的词向量按照文本内容中的顺序拼接起来得到文本内容的词向量矩阵,即词粒度特征向量。

[0071] $x_{1:n}=[x_1,x_2,x_3,\ldots,x_i,\ldots x_n]^T\in R^{n\times t}$

[0072] 其中,n表示文本内容中词的数量,t表示词向量的维度,x_i表示第i个词的词向量。

[0073] 文本内容中的词条粒度特征向量的提取可以将词条中每个字所对应的字向量按照顺序拼接起来得到词条的字向量矩阵,即词条粒度特征向量。

[0074] $w_{1:m}=[w_1,w_2,w_3,\ldots,w_p,\ldots w_m]^T\in R^{m\times r}$

[0075] 其中,m是文本内容的词条之中字的数量,r表示向量的维度,w_p表示词条之中第p个字的字向量。

[0076] 然后,命名实体识别模块可以根据命名实体识别模型,确定文本内容的命名实体粒度特征向量。

[0077] 命名实体粒度特征向量在利用命名实体识别模型实现提取之后,形成了“机构名”、“地名”、“城市名”等信息,其本身与标签之间存在直接对应关系,所以可以考虑将命名实体与标签之间建立直接匹配模型,即若命名实体和标签比对成功,即赋予标签。同时,命名实体也利用字向量进行提取。

[0078] $y_{1:m}=[y_1,y_2,y_3,\ldots,y_p,\ldots y_m]^T\in R^{m\times r}$

[0079] 其中,m是命名实体中字的数量,r表示字向量的维度,y_p表示命名实体之中第p个字的字向量。

[0080] 最后,可以通过特征融合模块将字粒度特征向量、词粒度特征向量、词条粒度特征向量以及命名实体粒度特征向量进行融合,得到文本特征向量。通过特征融合模块的融合,可以降低文本特征向量的维度。在特征融合模块中,首先通过连接前向隐藏状态和后向隐藏状态,以获得给定单词的向量信息,该向量以单词为中心总结了整个句子的信息;其次将单词通过一个单隐含层MLP(Multi-Layer Perceptron)进行馈送,以获得作为单词的隐藏表示;然后将单词的重要性度量与单词上下文向量的相似性进行对比,并通过softmax函数获得归一化的重要性权重;最后,计算单词的加权和,得到句子权重下的整体向量表示,即文本特征向量。

[0081] 本发明实施例中,特征提取层在得到文本内容之后,利用卷积神经网络学习文本中各个粒度微妙的特征。卷积神经网络接受的输入是矩阵形式,故需要将非结构化的文本表示为向量矩阵形式。为充分提取文本中的语义信息,考虑分别从“词”、“命名实体”、“字”和“词条”粒度入手。如图3所示,可以将文件内容分为文件名数据、属性数据以及除文件名数据、属性数据之外的剩余数据。针对文件名这一短词条数据进行特征提取,即对其分词数据、命名实体、全部词条构造特征向量进行表示。文件名中无实际含义的情况下,针对属性数据这若干分词数据进行特征提取,即对其分词数据和命名实体构造特征向量进行表示。除去文件名数据和属性数据的剩余数据,对剩余文本数据的分词数据、命名实体以及文本构造特征向量进行表示。

[0082] 本发明实施例中,通过特征提取层,可以实现多层次特征提取以及对全局特征信息和局部特征信息的融合,可以有效减少文本信息丢失。

[0083] 在上述实施例的基础上,本发明实施例中提供的文本标签识别方法,所述命名实体识别模型为基于BiLSTM和CRF的命名实体识别模型。

[0084] 具体地,本发明实施例中,命名实体识别模型可以通过基于BiLSTM和CRF构建,即经由BERT模型得到的字向量通过BiLSTM编码层生成对应高维度隐向量,之后经过softmax

层后便可生成实体标签的概率分布，并由CRF层来控制标签序列整体的出现概率从而得到最合理的标签序列，提取出主题、地点等实体。

[0085] 如图4所示，本发明实施例中采用的命名实体识别模型的结构中包括BiLSTM编码层(Bi-LSTM encoder)以及CRF层(CRF Layer)。BiLSTM编码层包括第一LSTM单元以及第二LSTM单元，第一LSTM单元包括I_1、I_2、I_3、I_4，第二LSTM单元包括r_1、r_2、r_3、r_4，下标1、2、3、4分别对应于字向量(Word embeddings)Mark、Watney、visited、Mars。最后各字向量对应的BiLSTM编码层的输出分别为c_1、c_2、c_3、c_4。BiLSTM编码层的输出经CRF层可以得到B-PER、E-PER、O、S-LOC。

[0086] 在上述实施例的基础上，本发明实施例中提供的文本标签识别方法，所述方法还包括：

[0087] 若判断获知多个文本样本中存在目标文本样本，所述目标文本样本未携带有文本样本标签，则基于预先确定的所述目标类型的标签集合，确定所述目标文本样本的文本样本标签。

[0088] 具体地，本发明实施例中，如果文本样本存在未携带有文本样本标签的目标文本样本时，则可以通过文本聚类和标签发现模块确定目标文本样本的文本样本标签。即对于目标文本样本的文本样本标签，可以根据预先确定的目标类型的标签集合进行确定。在目标类型的标签集合中可以存储有目标类型的大量文本样本中的文本样本标签。通过将目标文本样本与标签集合中存储的文本样本标签对应的各文本样本进行匹配，并将与目标文本样本匹配成功的文本样本的文本样本标签作为目标文本样本的文本样本标签。

[0089] 本发明实施例中，给出了未携带有文本样本标签的目标文本样本的文本样本标签确定方法，可以保证标签识别模型的训练过程顺利进行。

[0090] 在上述实施例的基础上，本发明实施例中提供的文本标签识别方法，所述标签集合基于如下方法确定：

[0091] 基于聚类算法，对多个文本样本对应的文本特征向量进行聚类，生成多个类簇；

[0092] 基于各簇类中包含的文本特征向量，确定备选标签集合；

[0093] 基于所述备选标签集合，确定所述标签集合。

[0094] 具体地，本发明实施例中，在确定标签集合时，可以使用聚类算法，对多个文本样本对应的文本特征向量进行聚类，生成多个类簇。聚类算法可以采用K-means聚类算法实现。然后可以从各簇类中包含的文本特征向量中，提取出标签，并将提取到的标签作为备选标签构建备选标签集合。即有：

[0095] $z_{1:m}=[z_1,z_2,z_3,\ldots z_m]^T\in R^{m\times r}$

[0096] $$\arg\min_{s}\sum_{i=1}^{k}\sum_{z\epsilon s_i}||z-\mu_i||^2=\arg\min_{s}\sum_{i=1}^{k}|s_i|Var\,(s_i)$$

[0097] 其中，z为特征融合模块输出的多个文本样本对应的文本特征向量，μ_i为s_i中的平均值点，s为簇类内所有文本特征向量平方和最小的k个簇类的集合，Var(s_i)为s_i的方差。

[0098] 由于通过聚类算法得到的备选标签集合中的备选标签有些缺乏现实意义或合理性解释，所以需要对备选标签集合中各备选标签进行审核，以确定最终的标签集合。审核的方式可以是人工审核，也可以是设定规则并通过审核模块进行审核，本发明实施例中对此

不作具体限定。对于审核通过的簇类即为最终确定的标签，对于审核未通过的簇类则舍弃。

[0099] 本发明实施例中，采用聚类算法，对多个文本样本对应的文本特征向量进行聚类，确定备选标签集合，并通过对备选标签集合中的备选标签进行审核确定最终的标签集合。不仅可以提高标签集合中标签的准确性，还可以使标签集合中标签具有现实意义以及合理性解释。

[0100] 在上述实施例的基础上，本发明实施例中提供的文本标签识别方法，所述提取目标类型的待识别文件中的文本内容，之后还包括：

[0101] 对所述文本内容进行数据清洗。

[0102] 具体地，本发明实施例中，在提取出待识别文件中的文本内容之后，由于文本内容中难免存在较多噪声，可利用的文本内容绝大多数都是高度非结构化，因此为了获得更好的文本标签提取效果，使用干净的文本内容至关重要。其中，数据清洗包括以下技术：

[0103] HTML字符转换：文本内容通常包含了大量html实体比如<&gt:&，嵌入在文本内容中。必须去掉这些实体，考虑用正则表达式直接删除；

[0104] 编码数据统一：编码是一个信息转换的过程，将复杂符号转换成简单易于理解的字符。文本内容可能受到不同形式的编码，例如“GBK”，“UTF-16”，“UTF-8”等。因此，为了更好的分析，必须让所有的文本内容保持标准的编码格式。统一的编码格式可以是“UTF-8”。

[0105] 移除标点符号：当后续数据分析需要在单词水平上被数据驱动时，标点符号并不含有实际意义，应该被移除。

[0106] 如图5所示，在上述实施例的基础上，本发明实施例中提供的文本标签识别方法的流程示意图，该方法包括：

[0107] 获取文本样本；

[0108] 提取文本样本中的文本内容；

[0109] 将文本内容输入至特征提取模块进行特征提取，得到词向量以及字粒度特征向量、词粒度特征向量、词条粒度特征向量；

[0110] 将词向量输入至命名实体识别模块，得到文本内容的命名实体粒度特征向量；

[0111] 将字粒度特征向量、词粒度特征向量、词条粒度特征向量以及命名实体粒度特征向量输入至特征融合模块进行融合，得到文本特征向量；

[0112] 判断文本样本是否携带有文本样本标签，如果没有，将其作为目标文本样本，将目标文本样本的文本特征向量输入至文本聚类和标签发现模块，确定目标文本样本的文本样本标签，并得到标注后的目标文本样本。如果有则将文本特征向量输入至分类器，通过模型训练得到标签识别模型。

[0113] 对于待识别文件，提取待识别文本中的文本内容，并将提取到的文本内容输入至标签识别模型，得到待识别文本中的文本标签。

[0114] 综上所述，本发明实施例中提供的文本标签识别方法，引入了标签识别模型，标签识别模型中包含有文本聚类与发现模块，从而可以针对不同的应用场景和数据，通过深度学习建立不同的标签集合。分类器可以针对特征融合模块得到的文本特征向量进行分类，达到了更高的准确率；基于BiLSTM和CRF的命名实体识别模块，可以针对各种数据集提取命名实体，达到更好的模型训练效果；文本聚类与发现模块弥补了目前技术方案中的空白，实现了基于深度学习对数据集的标签确定；多种数据格式文本内容提取的方法，优化了现有

技术只针对指定的文本格式输入的不足。

[0115] 如图6所示,在上述实施例的基础上,本发明实施例中提供了一种文本标签识别装置,包括:提取模块61和识别模块62。

[0116] 提取模块61,用于提取目标类型的待识别文件中的文本内容;

[0117] 识别模块62,用于将所述文本内容输入至标签识别模型中,得到所述标签识别模型输出的所述文本内容的文本标签;

[0118] 其中,所述标签识别模型包括特征提取层和分类层,所述特征提取层用于对所述文本内容进行多粒度特征提取,得到不同粒度的特征向量,并将不同粒度的特征向量进行融合,得到文本特征向量;所述分类层用于对所述文本特征向量进行分类;所述标签识别模型基于所述目标类型的多个文本样本训练得到。

[0119] 具体地,本发明实施例中提供的文本标签识别装置中各模块的作用与上述方法类实施例中各步骤的操作流程是一一对应的,实现的效果也是一致的,具体参见上述实施例,本发明实施例中对此不再赘述。

[0120] 在上述实施例的基础上,本发明实施例中提供的文本标签识别装置,所述识别模块具体用于:

[0121] 基于BRET模型,确定所述文本内容的字向量;

[0122] 基于所述字向量,确定文本内容的字粒度特征向量、词粒度特征向量以及词条粒度特征向量;

[0123] 基于命名实体识别模型,确定所述文本内容的命名实体粒度特征向量;

[0124] 将所述字粒度特征向量、所述词粒度特征向量、所述词条粒度特征向量以及所述命名实体粒度特征向量进行融合,得到所述文本特征向量。

[0125] 在上述实施例的基础上,本发明实施例中提供的文本标签识别装置,所述命名实体识别模型为基于BiLSTM和CRF的命名实体识别模型。

[0126] 在上述实施例的基础上,本发明实施例中提供的文本标签识别装置,所述装置还包括文本聚类和标签发现模块,用于:

[0127] 若判断获知多个文本样本中存在目标文本样本,所述目标文本样本未携带有文本样本标签,则基于预先确定的所述目标类型的标签集合,确定所述目标文本样本的文本样本标签。

[0128] 在上述实施例的基础上,本发明实施例中提供的文本标签识别装置,所述文本聚类和标签发现模块还用于:

[0129] 基于聚类算法,对多个文本样本对应的文本特征向量进行聚类,生成多个类簇;

[0130] 基于各簇类中包含的文本特征向量,确定备选标签集合;

[0131] 基于所述备选标签集合,确定所述标签集合。

[0132] 在上述实施例的基础上,本发明实施例中提供的文本标签识别装置,所述待识别文件的文件格式包括excel文件、csv文件、word文件及pdf文件。

[0133] 在上述实施例的基础上,本发明实施例中提供的文本标签识别装置,所述装置还包括数据预处理模块,用于:

[0134] 对所述文本内容进行数据清洗。

[0135] 图7示例了一种电子设备的实体结构示意图,如图7所示,该电子设备可以包括:处

理器(processor)710、通信接口(Communications Interface)720、存储器(memory)730和通信总线740,其中,处理器710,通信接口720,存储器730通过通信总线740完成相互间的通信。处理器710可以调用存储器730中的逻辑指令,以执行上述各实施例提供的文本标签识别方法,该方法包括:提取目标类型的待识别文件中的文本内容;将所述文本内容输入至标签识别模型中,得到所述标签识别模型输出的所述文本内容的文本标签;其中,所述标签识别模型包括特征提取层和分类层,所述特征提取层用于对所述文本内容进行多粒度特征提取,得到不同粒度的特征向量,并将不同粒度的特征向量进行融合,得到文本特征向量;所述分类层用于对所述文本特征向量进行分类;所述标签识别模型基于所述目标类型的多个文本样本训练得到。

[0136] 此外,上述的存储器730中的逻辑指令可以通过软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的部分可以以软件产品的形式体现出来,该计算机软件产品存储在一个存储介质中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本发明各个实施例所述方法的全部或部分步骤。而前述的存储介质包括:U盘、移动硬盘、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、磁碟或者光盘等各种可以存储程序代码的介质。

[0137] 另一方面,本发明还提供一种计算机程序产品,所述计算机程序产品包括存储在非暂态计算机可读存储介质上的计算机程序,所述计算机程序包括程序指令,当所述程序指令被计算机执行时,计算机能够执行上述各实施例提供的文本标签识别方法,该方法包括:提取目标类型的待识别文件中的文本内容;将所述文本内容输入至标签识别模型中,得到所述标签识别模型输出的所述文本内容的文本标签;其中,所述标签识别模型包括特征提取层和分类层,所述特征提取层用于对所述文本内容进行多粒度特征提取,得到不同粒度的特征向量,并将不同粒度的特征向量进行融合,得到文本特征向量;所述分类层用于对所述文本特征向量进行分类;所述标签识别模型基于所述目标类型的多个文本样本训练得到。

[0138] 又一方面,本发明还提供一种非暂态计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现以执行上述各实施例提供的文本标签识别方法,该方法包括:提取目标类型的待识别文件中的文本内容;将所述文本内容输入至标签识别模型中,得到所述标签识别模型输出的所述文本内容的文本标签;其中,所述标签识别模型包括特征提取层和分类层,所述特征提取层用于对所述文本内容进行多粒度特征提取,得到不同粒度的特征向量,并将不同粒度的特征向量进行融合,得到文本特征向量;所述分类层用于对所述文本特征向量进行分类;所述标签识别模型基于所述目标类型的多个文本样本训练得到。

[0139] 以上所描述的装置实施例仅仅是示意性的,其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的,作为单元显示的部件可以是或者也可以不是物理单元,即可以位于一个地方,或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下,即可以理解并实施。

[0140] 通过以上的实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，如ROM/RAM、磁碟、光盘等，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行各个实施例或者实施例的某些部分所述的方法。

[0141] 最后应说明的是：以上实施例仅用以说明本发明的技术方案，而非对其限制；尽管参照前述实施例对本发明进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本发明各实施例技术方案的精神和范围。

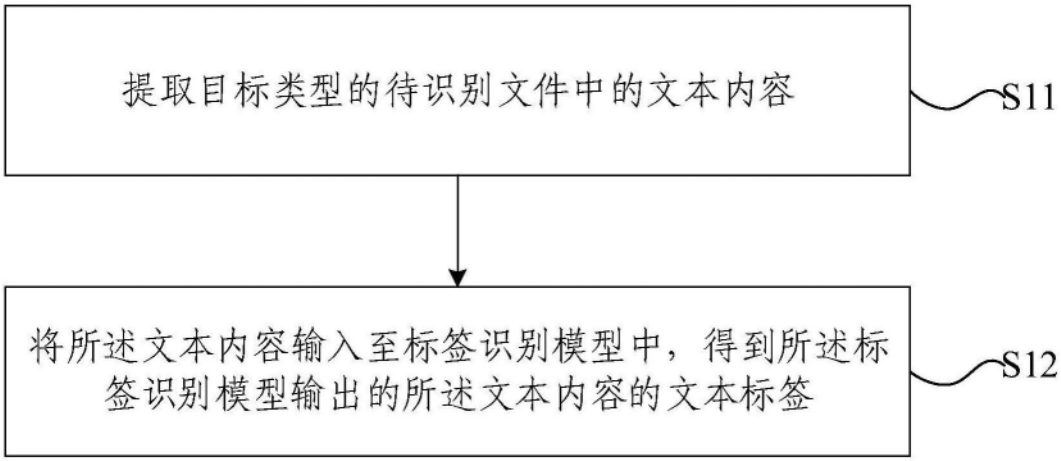

提取目标类型的待识别文件中的文本内容
S11
将所述文本内容输入至标签识别模型中，得到所述标签识别模型输出的所述文本内容的文本标签
S12

图1

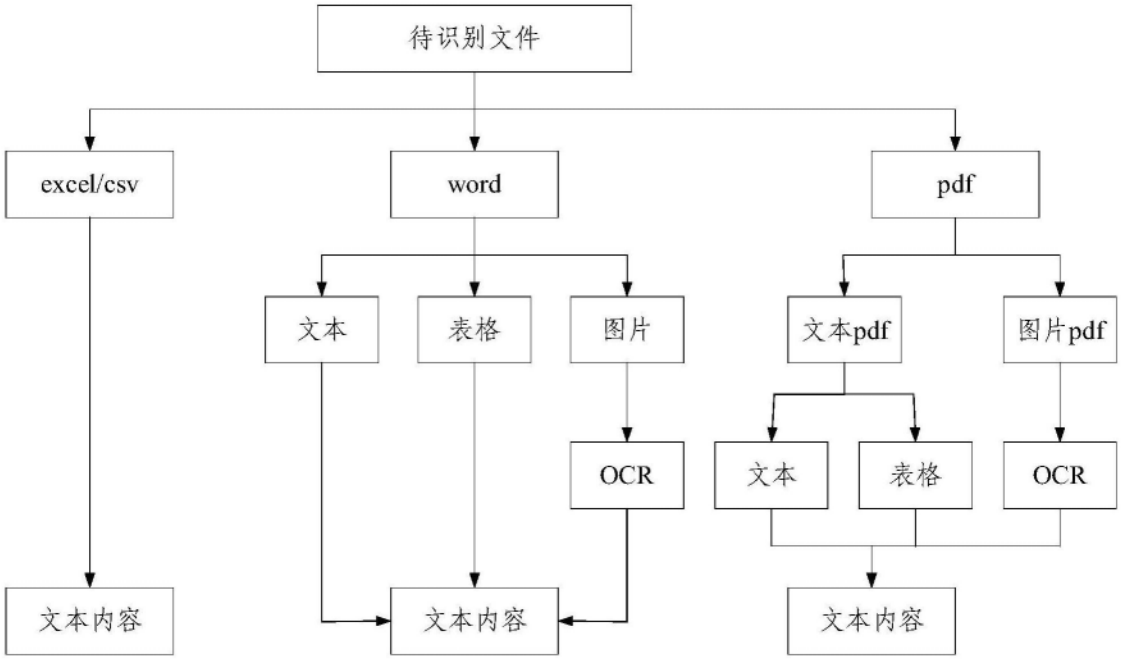

待识别文件
excel/csv
word
pdf
文本
表格
图片
文本pdf
图片pdf
OCR
文本
表格
OCR
文本内容
文本内容
文本内容

图2

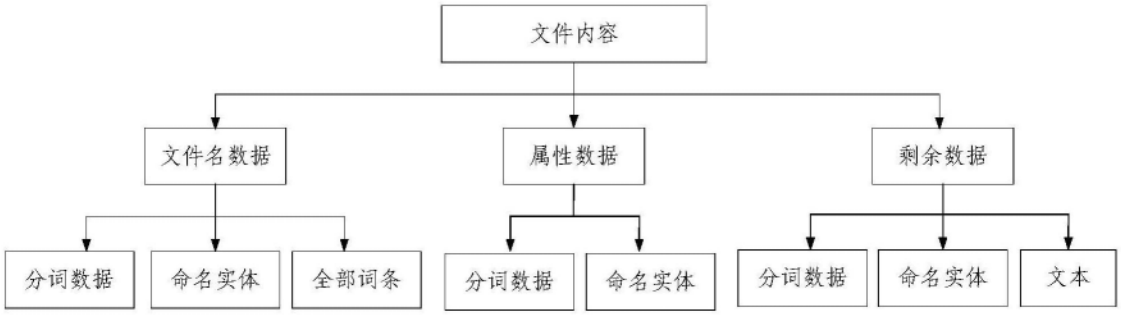

文件内容
文件名数据
属性数据
剩余数据
分词数据
命名实体
全部词条
分词数据
命名实体
分词数据
命名实体
文本

图3

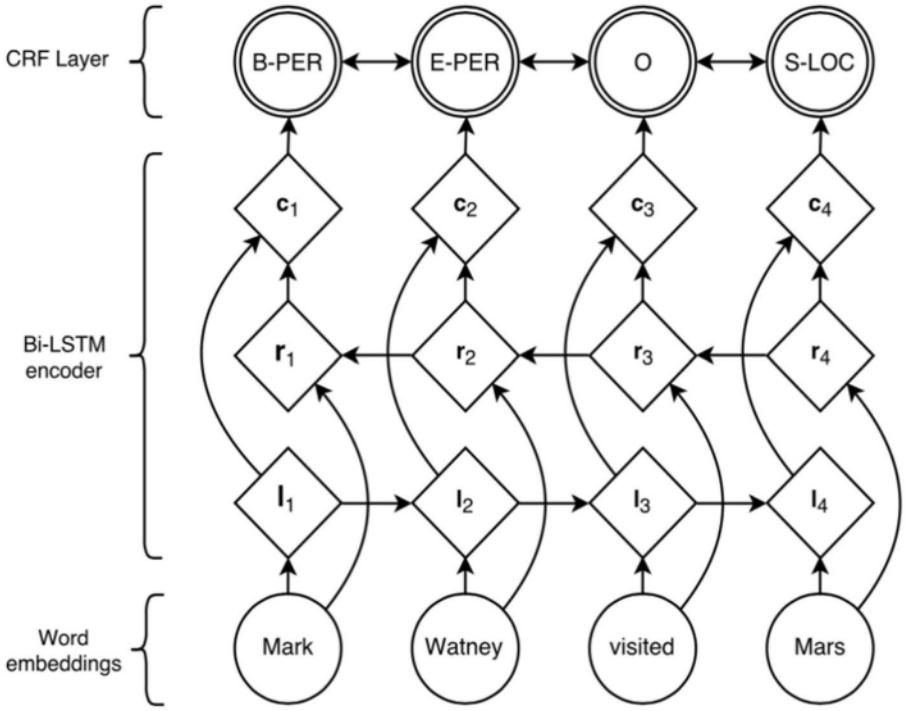

CRF Layer
Bi-LSTM encoder
Word embeddings
B-PER
E-PER
O
S-LOC
c1
c2
c3
c4
r1
r2
r3
r4
l1
l2
l3
l4
Mark
Watney
visited
Mars

图4

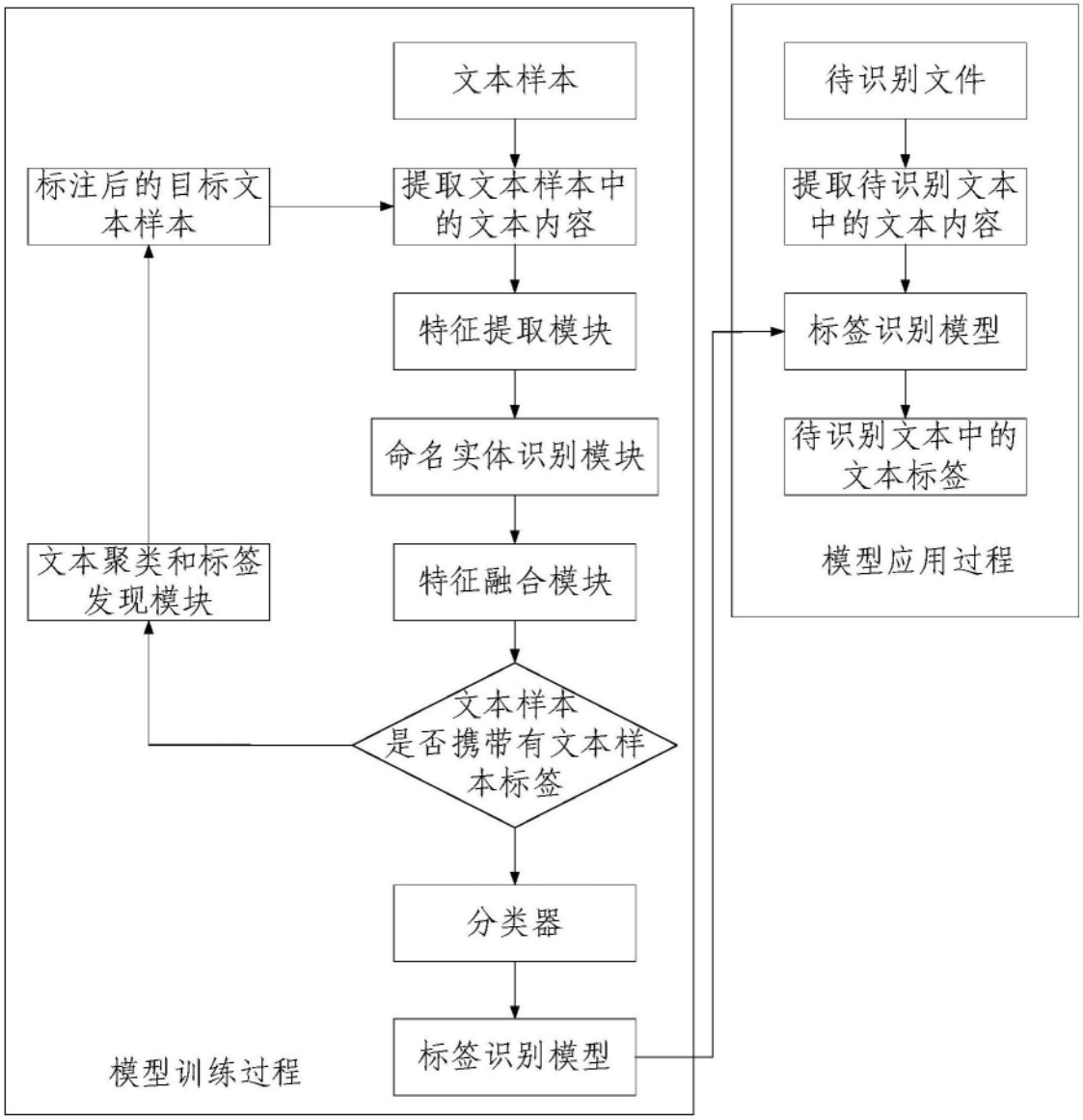

文本样本
标注后的目标文本样本
提取文本样本中的文本内容
特征提取模块
命名实体识别模块
文本聚类和标签发现模块
特征融合模块
文本样本是否携带有文本样本标签
分类器
标签识别模型
模型训练过程
待识别文件
提取待识别文本中的文本内容
标签识别模型
待识别文本中的文本标签
模型应用过程

图5


提取模块
61
识别模块
62

图6

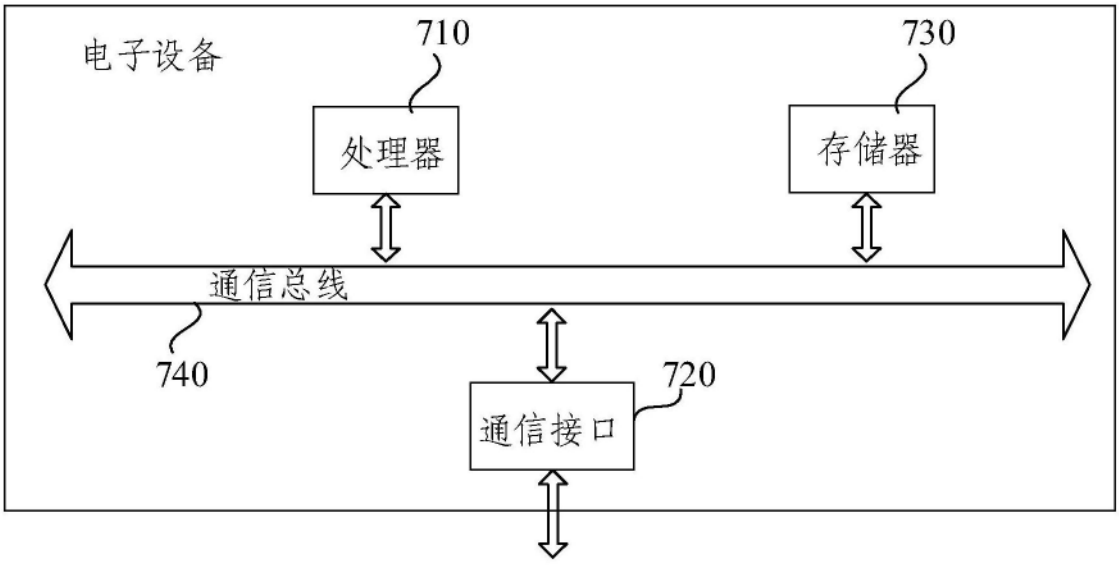


图7