(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 110287482 A
(43)申请公布日 2019.09.27

(21)申请号 201910455093.X

(22)申请日 2019.05.29

(71)申请人 西南电子技术研究所(中国电子科技集团公司第十研究所)
地址 610036 四川省成都市金牛区茶店子东街48号

(72)发明人 代翔 崔莹 黄细凤 孙涛 李强

(74)专利代理机构 成飞(集团)公司专利中心 51121
代理人 郭纯武

(51)Int.Cl.
G06F 17/27(2006.01)
G06K 9/62(2006.01)

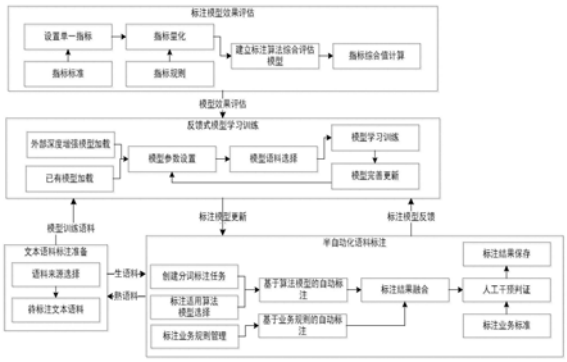
权利要求书2页 说明书6页 附图2页

(54)发明名称

半自动化分词语料标注训练装置

(57)摘要

本发明一种半自动化分词语料标注训练装置,旨在解决分词语料标注及训练过程中使用语料存在的弊端。本发明通过下述技术方案予以实现:文本语料标注准备模块对待标注语料、分词语料的管理,通过基于集成词典的双向最大匹配分词、CRF、JIEBA、等多种分词算法,将生语料分词标注工作提交给半自动化语料分词标注模块,创建分词标注任务,选择标注适用算法模型,开展自动标注,在自动标注结果融合的基础上,将文本语料标注准备模块产生的训练模型语料和标注模型反馈至反馈式模型学习训练模块,选择和模型学习训练,调用统一训练模型接口生成核心词典,更新分词训练模型表,建立标注算法综合评估模型对模型标注效果进行评估,完成新的分词标注任务。



CN 110287482 A

1.一种半自动化分词语料标注训练装置,包括:文本语料标注准备模块、半自动化语料分词标注模块、反馈式模型学习训练模块和分词标注模型效果评估模块,其特征在于:文本语料标注准备模块为标注任务提供准备,通过对不同来源的数据进行区分和语料来源选择,按来源或主题对待标注语料数据进行单一分词的预标注处理,实现对待标注语料、分词语料的管理,然后通过基于集成词典的双向最大匹配分词、条件随机场CRF、结巴JIEBA中文分词、双向LSTM网络、BI-LSTM多种分词算法,将生语料分词标注工作提交给半自动化语料分词标注模块;半自动化语料分词标注模块针对不同标注使用需求及语料特点,创建分词标注任务,选择标注适用算法模型,按标注业务规则管理开展自动标注,从集成词典的双向最大匹配分词、CRF、JIEBA、BI-LSTM多种分词算法中所选择的一种分词算法模型和业务规则,完成每一类标注任务的自动标注,基于算法模型的自动标注结果和业务规则的自动标注结果进行标注结果融合;在自动标注结果融合的基础上,依据标注业务标准进行人工干预判证,保存标注结果,将文本语料标注准备模块产生的训练模型语料和标注模型反馈至反馈式模型学习训练模块,根据已有模型和外部深度增强模型进行加载模型参数设置、模型语料选择和模型学习训练,将模型完善更新后再返回模型参数设置;调用统一训练模型接口Train生成核心词典和N-gram核心词典后,按统一模型接入接口导入外部算法模型,对模型进行更新或导出,保存包含核心词典和N-gram词典文件的分词模型文件,并更新分词训练模型表,建立标注算法综合评估模型,对模型标注效果进行评估,通过模型更新与语料标注之间的不断迭代,使用训练好的模型对平台中用于分词标注的模型进行更新,完成新的分词标注任务。

2.如权利要求1所述的半自动化分词语料标注训练装置,其特征在于:分词标注模型效果评估模块标注根据指标标准构建设置单一指标算法,按照指标计算规则对指标进行量化,根据不同标注任务采用组织相应指标构建标注算法综合评估模型,完成指标综合值计算,对标注模型效果进行反馈。

3.如权利要求1所述的半自动化分词语料标注训练装置,其特征在于:半自动化语料分词标注模块针对不同标注使用需求及语料特点,自主选择适配算法并开展自动标注,通过人工判证环节实现标注结果的干预判证。

4.如权利要求1所述的半自动化分词语料标注训练装置,其特征在于:文本语料标注准备模块根据不同来源语料创建分词标注任务;半自动化语料分词标注模块针对每一类标注任务选择效果适配的算法模型,在分词标注任务中,根据语料自动标注效果配置CRF、JIEBA、BI-LSTM算法选择CRF、JIEBA、BI-LSTM分词算法完成自动标注。

5.如权利要求1所述的半自动化分词语料标注训练装置,其特征在于:模型学习训练模块针对特殊标注任务创建业务标注规则,并对标注业务规则进行管理,并且标注业务规则包括业务字典和正则表达式;反馈式模型学习训练模块针对内外部标注模型算法,提供模型学习训练、反馈更新能力,采用标注业务规则对语料进行自动标注。

6.如权利要求1所述的半自动化分词语料标注训练装置,其特征在于:分词标注模型效果评估模块对基于算法模型的自动标注结果和基于业务规则的自动标注结果进行融合处理;在自动标注融合处理结果基础上,依据标注业务标准,人工对标注结果进行修改、确认和保存。

7.如权利要求1所述的半自动化分词语料标注训练装置,其特征在于:文本语料标注准

备模块对不同来源语料选择和管理，按不同标注任务保存为待标注的文本语料，即生语料；在半自动化语料分词标注模块中，创建相应的分词标注任务，并选择适用的标注算法模型，基于所选的算法模型对分词任务语料进行自动预标注，同时，针对数据所处领域的特殊性，结业相关业务规则进行基于业务规则的自动预标注，采用投票法对两类标注结果进行融合。

8.如权利要求1所述的半自动化分词语料标注训练装置，其特征在于：半自动化分词语料标注模块通过反馈式模型学习训练模块进行模型训练和更新，对标注所使用的已有模型进行反馈式模型学习训练，或采用外部深度加强模型进行反馈式模型学习训练；设置分词标注模型参数；选择分词模型训练所需熟语料进行模型学习训练。

9.如权利要求1所述的半自动化分词语料标注训练装置，其特征在于：在分词模型训练处理流中：模型语料选择模块选取用于做分词模型训练的语料，选择CRF、JIEBA、BI-LSTM分词算法进行训练，调用分词训练模型接口Train，生成核心词典和N-gram核心词典，使模型准确度达到最佳；判断是否保存分词模型，使用已标注语料数据对CRF、BI-LSTM可训练算法进行离线训练，按统一分词训练模型接入接口导入外部算法模型，对模型进行更新或导出，保存包含核心词典和N-gram词典文件的分词模型文件，并更新分词训练模型表。

10.如权利要求9所述的半自动化分词语料标注训练装置，其特征在于：分词模型更新后，启动分词服务，选择CRF、JIEBA、BI-LSTM分词算法训练，配置文件增加新分词开关，判断是否更新分词模型，是则读取指定分词模型，获取分词模型名称，否则读取分词训练模型表和读取算法自带核心词典，合并词典，更新分词训练模型表和完善模型，并将更新后的标注模型反馈给半自动化分词语料标注模块的算法自带核心词典，使用训练好的模型对平台中用于分词标注的模型进行更新，完成新的分词标注任务。

半自动化分词语料标注训练装置

技术领域

[0001] 本发明涉及文本挖掘技术领域,尤其涉及分词语料半自动化标注训练装置。

背景技术

[0002] 词是最小的、能够独立活动的、有意义的语言成分,但汉语中词语之间没有明显的区分标记,因此,中文词语分析是中文信息处理的基础与关键。分词的准确度和词性标注的准确度密切相关,有机地将分词过程和词性标注过程融合在一起,有利于消除歧义和提高整体效率。中文句子是由连续的字组成,字与字之间没有空格分离。词性标注是指为句子中的每个词确定一个合适的词性的过程。中文分词又是中文信息处理的第一道“工序”,在许多应用领域(文本分词、事件抽取、文本摘要、信息检索等)中扮演着极其重要的角色。分词和词性标注都是对语料进行基本的处理,统称为语料分词标注。然而有标注的分词语料很少,对于分词所在的大任务效果的提高是间接的,在一个实际系统中,不同分词错误的影响是非常不一样的,另外分词语料获得的成本非常昂贵,人工很难熟练地按照某一个标准前后一致地去标注生语料,使得在大数据量大计算能力的今天,分词语料的规模相当有限。词性标注在信息处理流程上是紧接着分词之后的步骤,而且所采用的算法原理与分词类似,所以在很多系统的实现中,常常对分词和词性标注进行一体化的处理。然而,目前领域内分词语料相对匮乏,且分词语料标注工作目前主要通过人工标注来完成,全人工对语料做词性标注就像蚂蚁一样忙忙碌碌,是非常耗费时的,并且存在语料标注质量差、标注过程繁琐、标注效率低、人力资源成本高等问题。同时,已有分词语料标注工具存在标注方法单一、无法对标注方法模型进行自动更新等弊端,因此,迫切需要一套能够辅助人工标注语料的半自动分词标注和训练平台来解决以上问题。如果有一个半自动化的分词标注方法和基于该方法设计的半自动化标注装置,能够对待处理的分词语料完全自动化地,迅速给出一篇预标注结果,这样才甚好。

[0003] 近年来,伴随大数据采集获取手段的高速发展,从数据中挖掘最大化价值变得尤为急迫,这对大数据的智能化分析提出了全新需求。在此背景下,机器学习、深度学习等技术在大数据应用上迅猛发展并获得了巨大成功,其技术底层使用的模型算法更多需要依赖于大量的数据标注语料作为基础训练支撑。海量数据语料标注工作对算法模型的训练有着重要影响,同时作为大数据分析过程中的基础性工作,主要支撑了大数据日常研发、算法调优、演示验证等环节,是大数据挖掘分析的核心基础。分词的关键取决于词典,目前结巴JIEBA提供的词典虽然并不是非常全,但是对于一般的应用已经足够了。结巴(jieba)插件,可以对一段中文进行分词,有三种分词模式,可以适应不同需求。中文分词(Chinese Word Segmentation)指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。

现有的分词算法可分为三大类:基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。基于字符串匹配的分词方法:这种方法又叫做机械分词方法,它是按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行配,若在词典中

找到某个字符串，则匹配成功(识别出一个词)。

1)正向最大匹配法(由左到右的方向)

2)逆向最大匹配法(由右到左的方向)：

3)最少切分(使每一句中切出的词数最小)

4)双向最大匹配法(进行由左到右、由右到左两次扫描)

基于理解的分词方法：这种分词方法是通过让计算机模拟人对句子的理解，达到识别词的效果。其基本思想就是在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。它通常包括三个部分：分词子系统、句法语义子系统、总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断，即它模拟了人对句子的理解过程。这种分词方法需要使用大量的语言知识和信息。由于汉语语言知识的笼统、复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于理解的分词系统还处在试验阶段。

基于统计的分词方法：给出大量已经分词的文本，利用统计机器学习模型学习词语切分的规律(称为训练)，从而实现对未知文本的切分。例如最大概率分词方法和最大熵分词方法等。随着大规模语料库的建立，统计机器学习方法的研究和发展，基于统计的中文分词方法渐渐成为了主流方法。

主要统计模型：N元文法模型(N-gram)，隐马尔可夫模型(Hidden Markov Model，HMM)，最大熵模型(ME)，条件随机场模型(Conditional Random Fields，CRF)等。词法分析是NLP的一项重要的基础技术，包括分词、词性标注、实体识别等，其主要算法结构为基于Bi-LSTM-CRF算法体系。用CRF是为获取全局最优的输出序列，相当于对lstm信息的再利用。从网络结构上来讲，Bi-LSTM-CRF套用的还是CRF这个大框架，只不过把LSTM在每个t时刻在第i个tag上的输出，看作是CRF特征函数里的“点函数”(只与当前位置有关的特征函数)，然后“边函数”(与前后位置有关的特征函数)还是用CRF自带的。这样就将线性链CRF里原始形式的特征函数(线性)变成LSTM的输出f_l(非线性)，这就在原始CRF中引入了非线性，可以更好的拟合数据。Bi-LSTM即双向LSTM，较单向的LSTM，Bi-LSTM能更好地捕获句子中上下文的信息。Bi-LSTM其实就是两个基于长短期记忆LSTM，只不过反向的LSTM是把输入的数据先reverse首尾转置一下，然后跑一个正常的LSTM，然后再把输出结果reverse一次使得与正向的LSTM的输入对应起来。

发明内容

[0004] 本发明的发明目的针对现有技术存在的不足之处，着眼于解决上述分词语料标注及训练过程中使用语料存在的弊端，提出一种半自动化分词语料标注训练装置。

[0005] 本发明的上述目的可以通过以下措施能达到，一种半自动化分词语料标注训练装置，包括：文本语料标注准备模块、半自动化语料分词标注模块、反馈式模型学习训练模块和分词标注模型效果评估模块，其特征在于：文本语料标注准备模块为标注任务提供准备，通过对不同来源的数据进行区分和语料来源选择，按来源或主题对待标注语料数据进行单一分词的预标注处理，实现对待标注语料、分词语料的管理，然后通过基于集成词典的双向最大匹配分词、CRF、JIEBA、BI-LSTM等多种分词算法，将生语料分词标注工作提交给半自动化语料分词标注模块；半自动化语料分词标注模块针对不同标注使用需求及语料特点，创

建分词标注任务，选择标注适用算法模型，按标注业务规则管理开展自动标注，基于集成词典的双向最大匹配分词、CRF、JIEBA、BI-LSTM等多种分词算法中所选择的一种分词算法模型和业务规则完成每一类标注任务的自动标注，基于算法模型的自动标注结果和基于业务规则的自动标注结果进行标注结果融合；在自动标注结果融合的基础上，依据标注业务标准进行人工干预判证，保存标注结果，将文本语料标注准备模块产生的训练模型语料和标注模型反馈至反馈式模型学习训练模块，根据已有模型和外部深度增强模型加载进行模型参数设置、模型语料选择和模型学习训练，将模型完善更新后再返回模型参数设置；调用统一训练模型接口Train生成核心词典和N-gram核心词典后，按统一模型接入接口导入外部算法模型，对模型进行更新或导出，保存包含核心词典和N-gram词典文件的分词模型文件，并更新分词训练模型表，建立标注算法综合评估模型对模型标注效果进行评估，通过模型更新与语料标注之间的不断迭代，使用训练好的模型对平台中用于分词标注的模型进行更新，完成新的分词标注任务；分词标注模型效果评估模块根据指标标准构建设置单一指标算法，按照指标计算规则对指标进行量化，根据不同标注任务采用组织相应指标构建标注算法综合评估模型，完成指标综合值计算，对标注模型效果进行反馈。

[0006] 本发明相比于现有技术具有如下的有益效果：

本发明通过建立标注算法综合评估模型对模型标注效果进行评估，反馈分词模型学习训练使模型达到最好效果，用于后续新增标注任务，通过模型更新与语料标注之间的不断迭代提高语料分词标注质量和算法模型效果。系统可针对不同标注使用需求及语料特点，提供基于自主选择适配算法和多算法融合的自动标注方式，多算法融合自动标注采用投票方法对多算法结果进行融合处理，在忽略相关性的条件下，集成方法的性能优于单一方法，通过该方法进行的预标注工作可降低人工标注过程繁复度，减轻人力工作成本；

本发明通过对不同来源的数据进行区分，实现对分词语料的管理；引入人工判证环节，通过集成基于词典的双向最大匹配分词、基于CRF分词、基于CRF+Bi-LSTM分词、JIEBA分词等算法，针对不同分词语料，在标注过程中提供适用的标注算法可选择，对待标注语料数据进行单一分词方法的预标注处理或多种分词方法融合的预标注处理，其中多种分词方法结果融合采用投票方法；支持实时后台分词算法模型的自动反馈调整，大幅提高语料标注效率和准确率；

本发明针对不同分词语料，通过集成基于词典的双向最大匹配分词、基于CRF分词、基于CRF+Bi-LSTM分词等多种分词算法，在标注过程中提供适用的标注算法可选择，对待标注语料数据进行单一分词方法的预标注处理或多分词方法融合的预标注处理，多分词方法结果融合采用投票方法；当标注任务完成后，使用标注语料对分词模型进行重新训练。最后通过人工确认环节对分词标注语料进行确认提交，完成语料分词标注工作。通过建立标注算法综合评估模型对模型标注效果进行评估，反馈分词模型学习训练使模型达到最好效果，用于后续新增标注任务，通过模型更新与语料标注之间的不断迭代提高语料分词标注质量和算法模型效果。

[0007] 本发明可以通过搭建Bi-LSTM网络实现序列标注，并可实现分词，准确率可达到95％左右，通过人工确认环节对分词标注语料进行修改、确认、提交，完成语料分词标注工作；当标注任务完成后，使用标注语料对分词模型进行重新训练；系统支持通过友好的人机交互式标注界面，简化用户标注操作流程；

本发明提供统一的分词模型接入标准，支持外部模型的导入、训练和使用。可在各种电子设备中应用。

附图说明

[0008] 图1是本发明半自动化分词语料标注、训练装置的工作原理示意图。

[0009] 图2是图1的分词模型训练处理流程图。

[0010] 为使本发明的目的、技术方案和优点更加清楚，下面结合实施方式和附图，对本发明作进一步地详细描述。

具体实施方式

[0011] 参阅图。在以下描述的优选实施例中，一种半自动化分词语料标注训练装置，包括：文本语料标注准备模块、半自动化语料分词标注模块、反馈式模型学习训练模块和分词标注模型效果评估模块，其特征在于：文本语料标注准备模块为标注任务提供准备，通过对不同来源的数据进行区分和语料来源选择，按来源或主题对待标注语料数据进行单一分词的预标注处理，实现对待标注语料、分词语料的管理，然后通过基于集成词典的双向最大匹配分词、条件随机场CRF、JIEBA、双向LSTM网络、BI-LSTM等多种分词算法、将生语料分词标注工作提交给半自动化语料分词标注模块；半自动化语料分词标注模块针对不同标注使用需求及语料特点，创建分词标注任务，选择标注适用算法模型，按标注业务规则管理开展自动标注，基于集成词典的双向最大匹配分词、CRF、JIEBA、BI-LSTM等多种分词算法中所选择的一种分词算法模型和业务规则完成每一类标注任务的自动标注，基于算法模型的自动标注结果和基于业务规则的自动标注结果进行标注结果融合；在自动标注结果融合的基础上，依据标注业务标准进行人工干预判证，保存标注结果，将文本语料标注准备模块产生的训练模型语料和标注模型反馈至反馈式模型学习训练模块，根据已有模型和外部深度增强模型加载进行模型参数设置、模型语料选择和模型学习训练，将模型完善更新后再返回模型参数设置；调用统一训练模型接口Train生成核心词典和N-gram核心词典后，按统一模型接入接口导入外部算法模型，对模型进行更新或导出，保存包含核心词典和N-gram词典文件的分词模型文件，并更新分词训练模型表，建立标注算法综合评估模型对模型标注效果进行评估，通过模型更新与语料标注之间的不断迭代，使用训练好的模型对平台中用于分词标注的模型进行更新，完成新的分词标注任务。分词标注模型效果评估模块根据指标标准构建设置单一指标算法，按照指标计算规则对指标进行量化，根据不同标注任务采用组织相应指标构建标注算法综合评估模型，完成指标综合值计算，对标注模型效果进行反馈。

[0012] 文本语料标注准备模块：完成对待标注语料按来源或主题进行管理，为标注任务提供准备；半自动化语料分词标注模块针对不同标注使用需求及语料特点，自主选择适配算法并开展自动标注，通过人工判证环节实现标注结果的干预判证，具体步骤如下：

文本语料标注准备模块根据不同来源语料创建分词标注任务；文本语料标注准备模块根据不同来源语料创建分词标注任务；半自动化语料分词标注模块针对每一类标注任务选择效果适配的算法模型，在分词标注任务中，根据语料自动标注效果配置条件随机场CRF、JIEBA、双向LSTM网络BI-LSTM算法选择CRF、JIEBA、BI-LSTM其中一种算法完成自动标注。为了建一个条件随机场CRF，首先要定义一个特征函数集，每个特征函数都以整个句子s，当前

位置i,位置i和i-1的标签为输入,然后为每一个特征函数赋予一个权重,然后针对每一个标注序列l,对所有的特征函数加权求和,必要的话,可以把求和的值转化为一个概率值。CRF的转移矩阵A由神经网络的CRF层近似得到,而P矩阵也就是发射矩阵由Bi-LSTM近似得到。

[0013] 模型学习训练模块针对特殊标注任务创建业务标注规则,并对标注业务规则进行管理,这里标注业务规则主要包括业务字典和正则表达式。

[0014] 反馈式模型学习训练模块针对内外部标注模型算法,提供模型学习训练、反馈更新能力,采用标注业务规则对语料进行自动标注。

[0015] 分词标注模型效果评估模块对基于算法模型的自动标注结果和基于业务规则的自动标注结果进行融合处理;在自动标注融合处理结果基础上,依据标注业务标准,人工对标注结果进行修改、确认和保存。标注人员通过文本语料标注准备模块对不同来源语料选择和管理,按不同标注任务保存为待标注的文本语料,即生语料;在半自动化语料分词标注模块中,创建相应的分词标注任务,并选择适用的标注算法模型,基于所选的算法模型对分词任务语料进行自动预标注,同时,针对数据所处领域的特殊性,结业相关业务规则进行基于业务规则的自动预标注,采用投票法对两类标注结果进行融合。基于标注的业务标准,通过人工干预判证环节对标注融合结果进行修改、调整,最终保存成为分词熟语料,为反馈式模型学习训练提供模型训练所需语料。

[0016] 半自动化分词语料标注模块中所使用的模型通过反馈式模型学习训练模块进行模型训练和更新,具体的:可对标注所使用的已有模型进行反馈式模型学习训练,也可采用外部深度加强模型进行反馈式模型学习训练;设置分词标注模型参数;选择分词模型训练所需熟语料进行模型学习训练。

[0017] 参阅图2。为半自动化分词语料标注训练装置详细工作流程。在分词模型训练处理流中:模型训练使用者通过模型语料选择模块选取用于做分词模型训练的语料,选择CRF、JIEBA、BI-LSTM分词算法训练,调用分词训练模型接口Train,生成核心词典和N-gram核心词典,使模型准确度达到最佳。判断是否保存分词模型,使用已标注语料数据对CRF、BI-LSTM等可训练算法进行离线训练,按统一分词训练模型接入接口导入外部算法模型,对模型进行更新或导出,保存包含核心词典和N-gram词典文件的分词模型文件,并更新分词训练模型表。分词模型更新后,启动分词服务,选择CRF、JIEBA、BI-LSTM分词算法训练,配置文件增加新分词开关,判断是否更新分词模型,是则读取指定分词模型,获取分词模型名称,否则读取分词训练模型表和读取算法自带核心词典,合并词典,更新分词训练模型表和完善模型,并将更新后的标注模型反馈给半自动化分词语料标注模块的算法自带核心词典,使用训练好的模型对平台中用于分词标注的模型进行更新,完成新的分词标注任务。

[0018] 分词标注模型效果评估模块提供模型评估指标构建标注、构建规则、指标量化等方法,支持通过自动构建标注算法综合评估模型对模型标注效果进行评估,具体步骤如下:分词标注模型效果评估模块根据指标标准构建设置单一指标算法;按照指标计算规则对指标进行量化,根据不同标注任务采用组织相应指标构建标注算法综合评估模型;完成指标综合值计算,对标注模型效果进行反馈。

[0019] 本实施例对通过本装置进行分词标注分词语料库标注的基本评价指标包括切分准确率precision,切分召回率Recall,F测度、交集型歧义准确率、组合型歧义准确率、兼类

标注准确率等。具体定义如下：$切分准确率Precision = \frac{正确切分（标注）的词语总数}{切分（标注）得到的词语总数} \times 100\%$

$切分召回率Recall = \frac{正确切分（标注）的词语总数}{标准语料库的词语总数} \times 100\%$，$F值 = \frac{2 \cdot P \cdot R}{P+R}$

其中，F表示F值，即为正确率和召回率的调和平均值，P表示准确率，R表示召回率。

[0020] 准确率和召回率一般称反比的关系。通过某些方法提高准确率，会导致召回率下降，反之亦然。为了定义应用系统对于准确率和召回率的不同需求，可以给出一个权重值对其进行加权的考量，从而得到E值：$E值 = 1 - \frac{1+b^2}{\frac{b^2}{P}+\frac{1}{R}}$

其中，b为加入的权重，b越大，则表示E值的考量中准确率的权重越大，反之则召回率的权重越大。

[0021] 切分歧义也是汉语自动分词算法的一个难点，为了考察算法对歧义消解的能力，特别对存在歧义的部分做单独的考察指标。具体的，针对交集型和组合型两种不同的歧义类型：交集型歧义和组合型歧义，分别定义精确率如下：

$$交集型歧义准确率Precision = \frac{正确切分的交集型歧义词语总数}{标准语料库中的的交集型歧义词语总数} \times 100\%$$

$$组合型歧义准确率Precision = \frac{正确切分的组合型歧义词语总数}{标准语料库中的的组合型歧义词语总数} \times 100\%$$

与切分歧义类似，词性标注也存在着自己的“歧义词”，称为兼类词。如果一个词拥有两个或者更多的不同词性，就称为兼类词。显然，兼类词的标注是词性标注的重点和难点，对此，定义一个专门的指标兼类词标注精确率来对其进行考察，具体定义如下

$$兼类词标注准确率Precision = \frac{正确标注的兼类词总数}{语料库中的兼类词总数} \times 100\%$$

通过对待标注语料按来源或主题进行管理，为标注任务提供准备；通过集成CRF、JIEBA、BI-LSTM等多种分词处理算法，完成分词语料的半自动化标注，在标注过程中提供适用的标注算法可选择，对待标注语料数据进行分词预标注处理；最后通过人工确认环节对标注语料进行修改、确认和提交，完成语料标注工作。当标注任务完成后，使用标注语料对模型进行重新训练。通过建立标注算法综合评估模型对模型标注效果进行评估，反馈模型学习训练使模型达到最好效果，用于后续新增标注任务，通过模型更新与语料标注之间的不断迭代提高语料标注质量和算法模型效果。

[0022] 以上所述为本发明较佳实施例，应该注意的是上述实施例对本发明进行说明，然而本发明并不局限于此，并且本领域技术人员在脱离所附权利要求的范围情况下可设计出替换实施例。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。

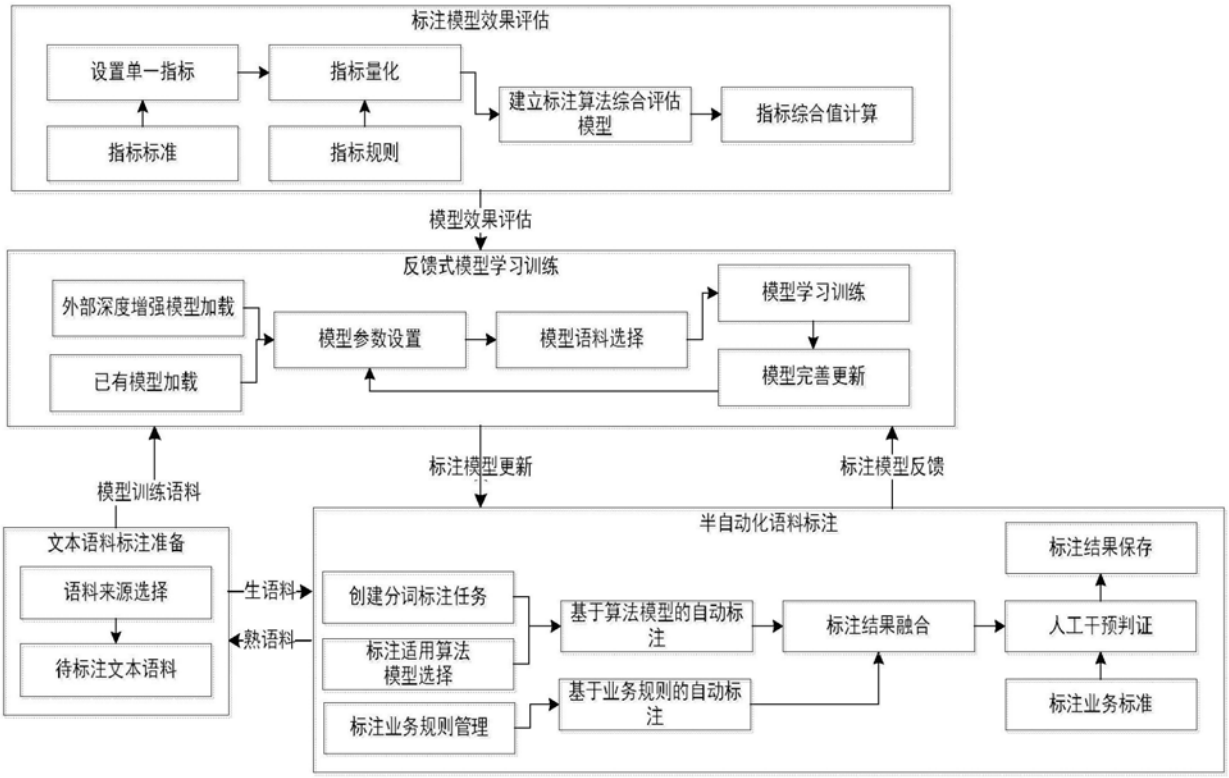

标注模型效果评估
设置单一指标
指标量化
建立标注算法综合评估模型
指标综合值计算
指标标准
指标规则
模型效果评估
反馈式模型学习训练
外部深度增强模型加载
已有模型加载
模型参数设置
模型语料选择
模型学习训练
模型完善更新
模型训练语料
标注模型更新
标注模型反馈
文本语料标注准备
语料来源选择
待标注文本语料
生语料
熟语料
半自动化语料标注
创建分词标注任务
标注适用算法模型选择
标注业务规则管理
基于算法模型的自动标注
基于业务规则的自动标注
标注结果融合
人工干预判证
标注结果保存
标注业务标准

图1

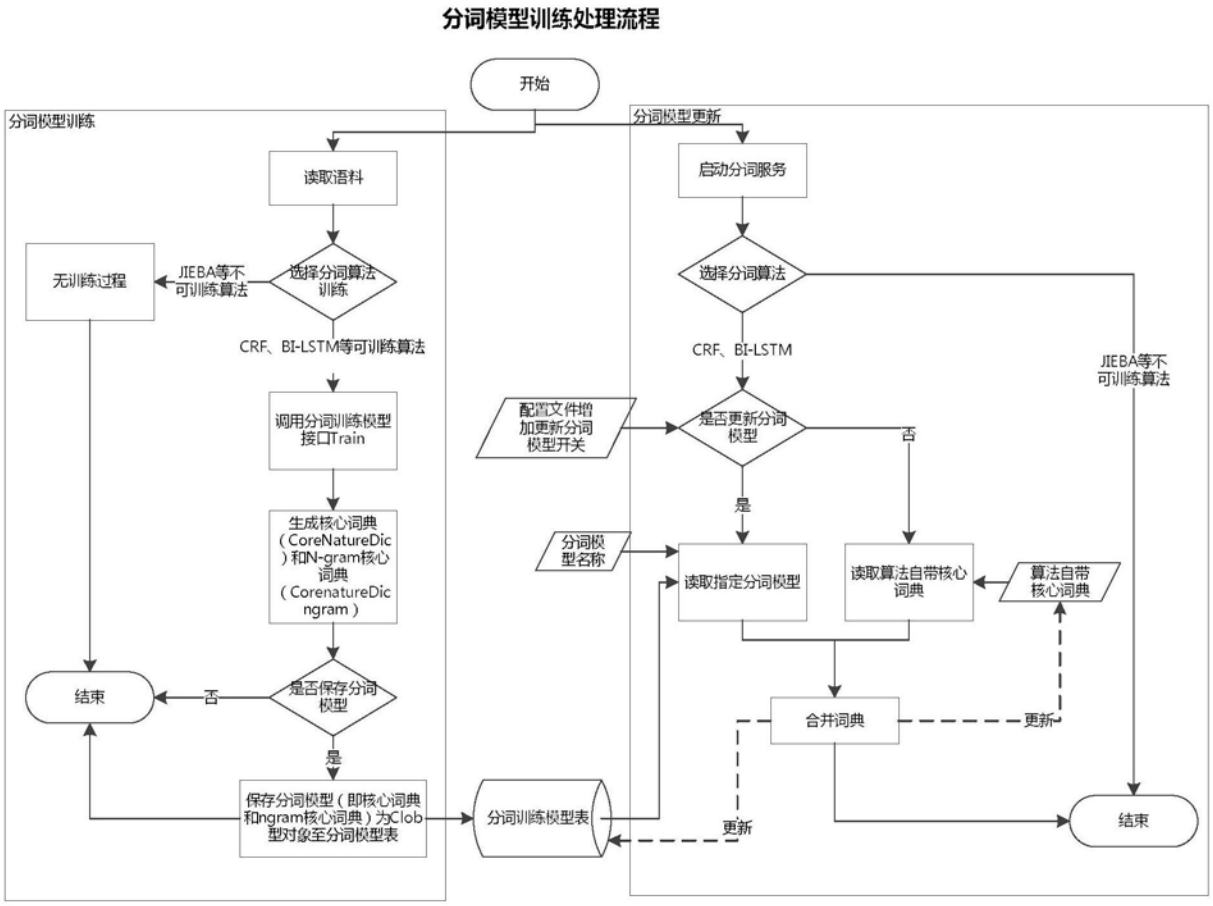

分词模型训练处理流程
开始
分词模型训练
读取语料
选择分词算法训练
JIEBA等不可训练算法
无训练过程
CRF、BI-LSTM等可训练算法
调用分词训练模型接口Train
生成核心词典（CoreNatureDic）和N-gram核心词典（CorenatureDicngram）
是否保存分词模型
否
结束
是
保存分词模型（即核心词典和ngram核心词典）为Clob型对象至分词模型表
分词训练模型表
分词模型更新
启动分词服务
选择分词算法
CRF、BI-LSTM
JIEBA等不可训练算法
配置文件增加更新分词模型开关
是否更新分词模型
否
是
分词模型名称
读取指定分词模型
读取算法自带核心词典
算法自带核心词典
合并词典
更新
更新
结束

图2