(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 117610552 B
(45)授权公告日 2024.11.19

(21)申请号 202311564031.5

(22)申请日 2023.11.22

(65)同一申请的已公布的文献号
申请公布号 CN 117610552 A

(43)申请公布日 2024.02.27

(73)专利权人 北京麦麦趣耕科技有限公司
地址 100020 北京市朝阳区光华路22号6层2单元708

(72)发明人 姚明磊 李楠 孙奥 翟斗号 刘家林

(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 张然

(51)Int.Cl.
G06F 40/279(2020.01)
G06F 40/211(2020.01)
G06N 3/0455(2023.01)
G06N 3/08(2023.01)

(56)对比文件
CN 103544246 A,2014.01.29
CN 105138510 A,2015.12.09

审查员 张镭

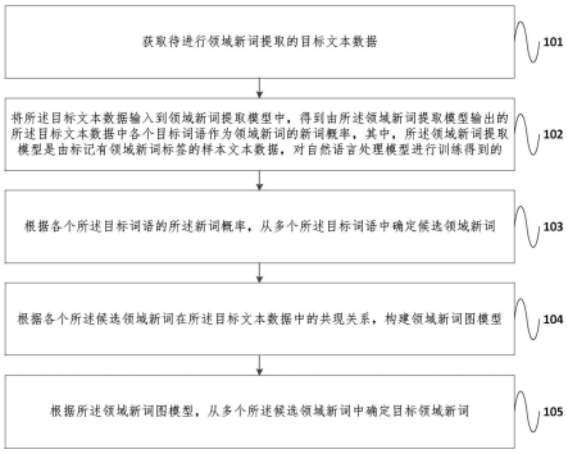
权利要求书3页 说明书14页 附图2页

(54)发明名称

领域新词提取方法、装置、电子设备及存储介质

(57)摘要

本发明提供一种领域新词提取方法、装置、电子设备及存储介质,涉及人工智能技术领域,该方法包括:获取待进行领域新词提取的目标文本数据;将目标文本数据输入到领域新词提取模型中,得到由领域新词提取模型输出的目标文本数据中各个目标词语作为领域新词的新词概率,其中,领域新词提取模型是由标记有领域新词标签的样本文本数据,对自然语言处理模型进行训练得到的;根据各个目标词语的新词概率,从多个目标词语中确定候选领域新词;根据各个候选领域新词在目标文本数据中的共现关系,构建领域新词图模型;根据领域新词图模型,从多个候选领域新词中确定目标领域新词。本发明更为准确地提取文本中的领域新词。



CN 117610552 B

1.一种领域新词提取方法,其特征在于,包括:

获取待进行领域新词提取的目标文本数据;

将所述目标文本数据输入到领域新词提取模型中,得到由所述领域新词提取模型输出的所述目标文本数据中各个目标词语作为领域新词的新词概率,其中,所述领域新词提取模型是由标记有领域新词标签的样本文本数据,对自然语言处理模型进行训练得到的;

根据各个所述目标词语的所述新词概率,从多个所述目标词语中确定候选领域新词;

根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型;

根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词。

2.根据权利要求1所述的领域新词提取方法,其特征在于,所述领域新词提取模型通过以下步骤训练得到:

获取多种行业领域对应的样本文本数据,并对所述样本文本数据进行文本预处理,得到预处理后的样本文本数据,其中,文本预处理至少包括分句处理、分词处理、去除停用词处理以及去除标点符号处理;

通过领域新词标签,标记出各个所述文本预处理后的样本文本数据中的样本领域新词,构建得到各种所述行业领域的样本文本数据集;

通过所述样本文本数据集,对预训练的BERT模型进行训练,若训练结果满足预设条件,得到所述领域新词提取模型。

3.根据权利要求1所述的领域新词提取方法,其特征在于,所述根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型,包括:

将各个所述候选领域新词和所述目标文本数据中其它目标词语作为节点,并根据所述目标文本数据的上下文信息,判断所述目标文本数据中的各个所述节点之间在是否存在共现关系;

若判断获知存在所述共现关系,将所述共现关系作为所述节点之间的边,并根据各个所述边两端的所述节点在所述目标文本数据中的共现频率,确定各个所述边的权重,构建得到所述领域新词图模型。

4.根据权利要求3所述的领域新词提取方法,其特征在于,所述根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词,包括:

基于PageRank算法,根据各个所述节点在所述领域新词图模型中的位置,以及各个所述节点之间的连接关系,对所述领域新词图模型中的各个所述节点进行重要性评估,得到所述领域新词图模型中各个所述节点的重要性得分;

将所述领域新词图模型中所述重要性得分满足预设新词重要性阈值的所述候选领域新词确定为所述目标领域新词。

5.根据权利要求4所述的领域新词提取方法,其特征在于,所述获取待进行领域新词提取的目标文本数据,包括:

在预设文本数据采集时段内,获取多个所述目标文本数据,且各个所述目标文本数据之间的文本内容是不同的,以通过各个所述目标文本数据,构建得到对应的所述领域新词图模型;

在所述基于PageRank算法,根据各个所述节点在所述领域新词图模型中的位置,以及

各个所述节点之间的连接关系,对所述领域新词图模型中的各个所述节点进行重要性评估,得到所述领域新词图模型中各个所述节点的重要性得分之后,所述方法还包括:

获取各个所述目标文本数据对应的所述领域新词图模型;

通过图对齐技术和图嵌入技术,将所有所述领域新词图模型进行融合处理,得到领域新词融合图;

获取所述领域新词融合图中各个所述候选领域新词对应的词关联关系对,其中,所述词关联关系对是由所述领域新词融合图中存在所述边的所述候选领域新词形成的;

根据所述领域新词融合图中各个所述节点的所述重要性得分,获取各个所述词关联关系对的词关联得分;

将所述词关联得分满足预设词关联阈值的所述词关联关系对确定为目标词关联关系对,并根据所述目标词关联关系对,生成对应的词云;

在所述根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词之后,所述方法还包括:

将所述目标领域新词和所述词云存储至领域新词库。

6.根据权利要求5所述的领域新词提取方法,其特征在于,所述根据各个所述目标词语的所述新词概率,从多个所述目标词语中确定候选领域新词,包括:

基于所述新词概率从大到小的顺序,将所有所述目标词语进行排序,并根据排序结果,选取前N个所述目标词语作为待定词语;

获取各个所述待定词语在所述目标文本数据中的出现次数,并判断所述出现次数是否大于预设频次阈值,若大于,则将所述待定词语确定为所述候选领域新词,其中,所述预设频次阈值是根据所述领域新词库中所有领域新词在所有历史文本数据中的出现次数生成的;所述历史文本数据为所述领域新词库中在历史时刻新增领域新词时,获取到的新增的领域新词对应的文本数据;

在所述将所述目标领域新词和所述词云存储至领域新词库之后,所述方法还包括:

将所述目标领域新词对应的目标文本数据存储至所述领域新词库,得到更新后的领域新词库;

基于所述更新后的领域新词库,对所述预设频次阈值进行更新,得到更新后的预设频次阈值。

7.根据权利要求6所述的领域新词提取方法,其特征在于,在所述获取各个所述待定词语在所述目标文本数据中的出现次数,并判断所述出现次数是否大于预设频次阈值,若大于,则将所述待定词语确定为所述候选领域新词之后,所述方法还包括:

将获取到的多个所述候选领域新词构建为候选领域新词集合,并对所述候选领域新词集合进行筛选处理,所述筛选处理包括:

判断所述候选领域新词集合中任意一个候选领域新词是否属于其它候选领域新词的子串,若属于,则将属于所述子串的候选领域新词从所述候选领域新词集合中删除,得到筛选处理后的候选领域新词集合;

所述根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型,包括:

根据所述筛选处理后的候选领域新词集合中的各个所述候选领域新词在所述目标文

本数据中的共现关系,构建领域新词图模型。

8.一种领域新词提取装置,其特征在于,包括:

文本数据输入模块,用于获取待进行领域新词提取的目标文本数据;

预测模块,用于将所述目标文本数据输入到领域新词提取模型中,得到由所述领域新词提取模型输出的所述目标文本数据中各个目标词语作为领域新词的新词概率,其中,所述领域新词提取模型是由标记有领域新词标签的样本文本数据,对自然语言处理模型进行训练得到的;

候选新词筛选模块,用于根据各个所述目标词语的所述新词概率,从多个所述目标词语中确定候选领域新词;

图模型构建模块,用于根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型;

图模型处理模块,用于根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词。

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计算机程序时实现如权利要求1至7任一项所述领域新词提取方法。

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任一项所述领域新词提取方法。

领域新词提取方法、装置、电子设备及存储介质

技术领域

[0001] 本发明涉及人工智能技术领域,尤其涉及一种领域新词提取方法、装置、电子设备及存储介质。

背景技术

[0002] 领域新词是指在特定领域或专业中出现的新词汇或术语,随着科技发展,不同领域和行业都在不断涌现新的概念、技术和现象,领域新词在专业交流、学术研究、行业报道和领域内部沟通中起到重要的角色,有助于精确传达特定领域的信息,提高沟通效率,并促进该领域的进一步发展和创新。

[0003] 在学术研究和文献分析中,领域新词的提取可以帮助研究人员追踪和理解最新的研究成果和话题;而在信息检索和文本挖掘中,也可以利用领域新词来提高搜索的效果和准确性。然而,由于领域新词的特殊性和专业性,现有针对领域新词的提取方式,如通过人工的方式来标注和识别领域中的新词,依赖于人工经验和专业知识,容易受到主观因素的影响,并且无法处理大规模的文本数据;或,通过频率统计,分析文本中词语的出现频率和特定领域的背景知识,确定频率较高但在通用语料库中较少出现的词语作为领域新词,该方法无法准确判断一个词是否是新词,而且对于多义词的判别也存在困难。

[0004] 因此,现在亟需一种领域新词提取方法、装置、电子设备及存储介质来解决上述问题。

发明内容

[0005] 针对现有技术存在的问题,本发明提供一种领域新词提取方法、装置、电子设备及存储介质。

[0006] 本发明提供一种领域新词提取方法,包括:

[0007] 获取待进行领域新词提取的目标文本数据;

[0008] 将所述目标文本数据输入到领域新词提取模型中,得到由所述领域新词提取模型输出的所述目标文本数据中各个目标词语作为领域新词的新词概率,其中,所述领域新词提取模型是由标记有领域新词标签的样本文本数据,对自然语言处理模型进行训练得到的;

[0009] 根据各个所述目标词语的所述新词概率,从多个所述目标词语中确定候选领域新词;

[0010] 根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型;

[0011] 根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词。

[0012] 根据本发明提供的一种领域新词提取方法,所述领域新词提取模型通过以下步骤训练得到:

[0013] 获取多种行业领域对应的样本文本数据,并对所述样本文本数据进行文本预处

理,得到预处理后的样本文本数据,其中,文本预处理至少包括分句处理、分词处理、去除停用词处理以及去除标点符号处理;

[0014] 通过领域新词标签,标记出各个所述文本预处理后的样本文本数据中的样本领域新词,构建得到各种所述行业领域的样本文本数据集;

[0015] 通过所述样本文本数据集,对预训练的BERT模型进行训练,若训练结果满足预设条件,得到所述领域新词提取模型。

[0016] 根据本发明提供的一种领域新词提取方法,所述根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型,包括:

[0017] 将各个所述候选领域新词和所述目标文本数据中其它目标词语作为节点,并根据所述目标文本数据的上下文信息,判断所述目标文本数据中的各个所述节点之间在是否存在共现关系;

[0018] 若判断获知存在所述共现关系,将所述共现关系作为所述节点之间的边,并根据各个所述边两端的所述节点在所述目标文本数据中的共现频率,确定各个所述边的权重,构建得到所述领域新词图模型。

[0019] 根据本发明提供的一种领域新词提取方法,所述根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词,包括:

[0020] 基于PageRank算法,根据各个所述节点在所述领域新词图模型中的位置,以及各个所述节点之间的连接关系,对所述领域新词图模型中的各个所述节点进行重要性评估,得到所述领域新词图模型中各个所述节点的重要性得分;

[0021] 将所述领域新词图模型中所述重要性得分满足预设新词重要性阈值的所述候选领域新词确定为所述目标领域新词。

[0022] 根据本发明提供的一种领域新词提取方法,所述获取待进行领域新词提取的目标文本数据,包括:

[0023] 在预设文本数据采集时段内,获取多个所述目标文本数据,且各个所述目标文本数据之间的文本内容是不同的,以通过各个所述目标文本数据,构建得到对应的所述领域新词图模型;

[0024] 在所述基于PageRank算法,根据各个所述节点在所述领域新词图模型中的位置,以及各个所述节点之间的连接关系,对所述领域新词图模型中的各个所述节点进行重要性评估,得到所述领域新词图模型中各个所述节点的重要性得分之后,所述方法还包括:

[0025] 获取各个所述目标文本数据对应的所述领域新词图模型;

[0026] 通过图对齐技术和图嵌入技术,将所有所述领域新词图模型进行融合处理,得到领域新词融合图;

[0027] 获取所述领域新词融合图中各个所述候选领域新词对应的词关联关系对,其中,所述词关联关系对是由所述领域新词融合图中存在所述边的所述候选领域新词形成的;

[0028] 根据所述领域新词融合图中各个所述节点的所述重要性得分,获取各个所述词关联关系对的词关联得分;

[0029] 将所述词关联得分满足预设词关联阈值的所述词关联关系对确定为目标词关联关系对,并根据所述目标词关联关系对,生成对应的词云;

[0030] 在所述根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词

之后,所述方法还包括:

[0031] 将所述目标领域新词和所述词云存储至领域新词库。

[0032] 根据本发明提供的一种领域新词提取方法,所述根据各个所述目标词语的所述新词概率,从多个所述目标词语中确定候选领域新词,包括:

[0033] 基于所述新词概率从大到小的顺序,将所有所述目标词语进行排序,并根据排序结果,选取前N个所述目标词语作为待定词语;

[0034] 获取各个所述待定词语在所述目标文本数据中的出现次数,并判断所述出现次数是否大于预设频次阈值,若大于,则将所述待定词语确定为所述候选领域新词,其中,所述预设频次阈值是根据所述领域新词库中所有领域新词在所有历史文本数据中的出现次数生成的;所述历史文本数据为所述领域新词库中在历史时刻新增领域新词时,获取到的新增的领域新词对应的文本数据;

[0035] 在所述将所述目标领域新词和所述词云存储至领域新词库之后,所述方法还包括:

[0036] 将所述目标领域新词对应的目标文本数据存储至所述领域新词库,得到更新后的领域新词库;

[0037] 基于所述更新后的领域新词库,对所述预设频次阈值进行更新,得到更新后的预设频次阈值。

[0038] 根据本发明提供的一种领域新词提取方法,在所述获取各个所述待定词语在所述目标文本数据中的出现次数,并判断所述出现次数是否大于预设频次阈值,若大于,则将所述待定词语确定为所述候选领域新词之后,所述方法还包括:

[0039] 将获取到的多个所述候选领域新词构建为候选领域新词集合,并对所述候选领域新词集合进行筛选处理,所述筛选处理包括:

[0040] 判断所述候选领域新词集合中任意一个候选领域新词是否属于其它候选领域新词的子串,若属于,则将属于所述子串的候选领域新词从所述候选领域新词集合中删除,得到筛选处理后的候选领域新词集合;

[0041] 所述根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型,包括:

[0042] 根据所述筛选处理后的候选领域新词集合中的各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型。

[0043] 本发明还提供一种领域新词提取装置,包括:

[0044] 文本数据输入模块,用于获取待进行领域新词提取的目标文本数据;

[0045] 预测模块,用于将所述目标文本数据输入到领域新词提取模型中,得到由所述领域新词提取模型输出的所述目标文本数据中各个目标词语作为领域新词的新词概率,其中,所述领域新词提取模型是由标记有领域新词标签的样本文本数据,对自然语言处理模型进行训练得到的;

[0046] 候选新词筛选模块,用于根据各个所述目标词语的所述新词概率,从多个所述目标词语中确定候选领域新词;

[0047] 图模型构建模块,用于根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型;

[0048] 图模型处理模块,用于根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词。

[0049] 本发明还提供一种电子设备,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如上述任一种所述领域新词提取方法。

[0050] 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如上述任一种所述领域新词提取方法。

[0051] 本发明提供的领域新词提取方法、装置、电子设备及存储介质,通过由自然语言处理模型训练得到的领域新词提取模型,对文本数据中的词语进行新词概率预测,再根据预测得到的新词概率,选取候选领域新词进行领域新词图模型构建,最终通过领域新词图模型获取文本数据中的领域新词,从而更为准确地提取文本中的领域新词。

附图说明

[0052] 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进行简单地介绍,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。

[0053] 图1为本发明提供的领域新词提取方法的流程示意图;

[0054] 图2为本发明提供的基于图算法的领域新词的关系构建过程示意图;

[0055] 图3为本发明提供的基于深度学习模型的候选领域新词的提取过程示意图;

[0056] 图4为本发明提供的领域新词提取装置的结构示意图;

[0057] 图5为本发明提供的电子设备的结构示意图。

具体实施方式

[0058] 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。

[0059] 在特定行业或领域内,需要精确地识别并提取领域新词(如专业术语、行业特定语汇、新兴词汇以及非标准化的口语表达)。这些领域新词能够帮助专业人员及时获取领域最新信息,深入了解行业发展趋势、事件动态以及社区舆论态势,从而为相关研究提供精准的素材和数据支持。

[0060] 在目前的领域新词提取方法中,可通过专家或从业者会对特定领域的文本进行阅读和分析,识别和标注其中的新词,但这种方法费时费力,依赖于人工经验和专业知识,容易受到主观因素的影响,并且无法处理大规模的文本数据。或者,通过词频统计来提取领域新词,该方法通过分析文本中词语的出现频率和特定领域的背景知识,确定频率较高但在通用语料库中较少出现的词语作为领域新词,然而,该方法可能忽略一些上下文相关的特征,无法准确判断一个词是否是新词,而且对于多义词的判别也存在困难。

[0061] 基于现有领域新词提取方法中存在的问题,本发明通过深度学习模型分析文本,

识别并提取文本中的候选领域新词;然后,利用图算法对这些候选领域新词之间的词关系进行筛选和评估,从而实现领域新词与词关系的准确提取,同时也能提高新词提取的效率。

[0062] 图1为本发明提供的领域新词提取方法的流程示意图,如图1所示,本发明提供了一种领域新词提取方法,包括:

[0063] 步骤101,获取待进行领域新词提取的目标文本数据。

[0064] 在本发明中,可通过各种不同的领域平台获取到目标文本数据,从而通过后续领域新词提取过程,将这些目标文本数据中相关的领域新词提取出来。在一实施例中,目标文本数据可通过以下方式获取得到:1、用户生成内容(User Generated Content,简称UGC),如用户在社交媒体、论坛和博客等平台上发布的文本内容,这些内容通常反映了用户个人的观点、情绪和讨论话题,可能包含一些领域内的新词;2、企业内部文档,包括企业内部的报告、会议记录、项目文档和产品文档等,这些文档可能包含企业特定的术语、缩写词以及行业内的新兴词汇;3、社区讨论内容,包括在线社区、群组或者论坛上用户之间的讨论和互动,这些讨论文本中可能涉及到特定领域的话题和相关术语;4、行业报告内容,包括从专业机构、咨询公司和研究机构等发布的行业报告、调研报告和市场分析报告,这些报告通常包含了对特定领域的深入研究和分析,可能涉及到新兴词汇和行业术语。在一实施例中,可提前对目标文本数据进行文本预处理,如分词、分句、去除停用词和去除无意义符号等,保留后续需要进行领域新词预测的目标词语,从而将原始的目标文本数据转化为更适合深度学习模型处理的输入数据,即得到文本预处理后的目标文本数据。

[0065] 步骤102,将所述目标文本数据输入到领域新词提取模型中,得到由所述领域新词提取模型输出的所述目标文本数据中各个目标词语作为领域新词的新词概率,其中,所述领域新词提取模型是由标记有领域新词标签的样本文本数据,对自然语言处理模型进行训练得到的。

[0066] 在本发明中,在进行领域新词提取之前,前期需要通过对自然语言处理模型进行训练,从而得到领域新词提取模型。具体地,收集并整理标记有领域新词标签的样本文本数据,这些样本文本数据可以来自专家的标注或领域内的专业人士的标注,而领域新词标签可以是二元标签,即表示该词是否为领域新词。进一步地,使用准备好的样本文本数据和特征,对自然语言处理模型进行训练,并在模型完成训练后,使用一部分标记有领域新词标签的测试集对训练好的模型进行评估,评估指标可以包括准确率、召回率和F1值等,以根据评估结果,对模型进行调优和优化,如调整模型参数和增加训练样本等。

[0067] 在确定模型训练好之后,通过将目标文本数据输入到训练好的领域新词提取模型中,可以获得各个目标词语作为新词的概率,即目标文本数据中各个位置的词语被预测为一个领域新词的概率,这个概率表示模型认为这个位置的词语是新词的可能性有多大。在一实施例中,可根据设定的阈值,筛选出具有较高新词概率的词语。

[0068] 步骤103,根据各个所述目标词语的所述新词概率,从多个所述目标词语中确定候选领域新词。

[0069] 在本发明中,可确定一个概率阈值,该阈值用于从目标文本数据中筛选出具有较高新词概率的词语作为候选领域新词,阈值的选择可以根据实际需求和领域特点进行调整。在一实施例中,也可根据每个目标词语的新词概率,将所有目标词语按照新词概率从高到低进行排序;然后,对排序后的目标词语列表开始逐个遍历,依次比较每个词语的新词概

率与设定的阈值,如果某个词语的新词概率高于阈值,且该词语的长度也在预设新词长度(如设置最大新词长度)以内,则将其选为候选领域新词。

[0070] 步骤104,根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型。

[0071] 在本发明中,通过前面的步骤,得到一组候选领域新词,这些词语将作为节点构成新词图模型。同时,在目标文本数据中,统计候选领域新词之间的共现关系,在本发明中,共现关系可以基于词语的上下文窗口内出现频率进行计算,反映了候选领域新词之间的联系程度。

[0072] 进一步地,根据共现关系,将候选领域新词和目标文本中其它词语作为节点,共现关系作为边,构建得到一个由新词和文本中所有词语构成的图模型,该模型详细地描述了目标文本数据中词语之间的共现关系,将用于后续步骤中基于图算法的新词及其关联词的分析。

[0073] 步骤105,根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词。

[0074] 在本发明中,根据预设需求,选择适当的图算法,对领域新词图模型进行评估和分析。进一步地,图算法会根据节点在图中的位置以及与其它节点的连接关系,计算出每个节点的重要性得分,其中,对于属于候选领域新词的节点,若与更多其它节点相连接,或者与重要性更高的节点相连接,通常会获得更高的得分。然后,从排名较高且属于候选领域新词的节点中选取几个作为目标领域新词,这些词语被认为在领域中具有较高的重要性和代表性,同时,选取的目标领域新词具有较高的准确性和实用性。

[0075] 本发明提供的领域新词提取方法,通过由自然语言处理模型训练得到的领域新词提取模型,对文本数据中的词语进行新词概率预测,再根据预测得到的新词概率,选取候选领域新词进行领域新词图模型构建,最终通过领域新词图模型获取文本数据中的领域新词,从而更为准确地提取文本中的领域新词。

[0076] 在上述实施例的基础上,所述领域新词提取模型通过以下步骤训练得到:

[0077] 获取多种行业领域对应的样本文本数据,并对所述样本文本数据进行文本预处理,得到预处理后的样本文本数据,其中,文本预处理至少包括分句处理、分词处理、去除停用词处理以及去除标点符号处理。

[0078] 在本发明中,在获取到多种行业领域的样本文本数据之后,需要对原始的样本文本数据进行清洗和整理,包括分句、分词、去除停用词和去除无意义的符号。具体地,通过分句处理,将长篇幅的样本文本数据分解成独立的句子,在一实施例中,使用SpaCy自然语言处理库,根据标点符号和其他语言规则将文本分割成句子;然后,通过分词处理,将每个句子进一步被分解成单独的词语或标记(token),这个过程也称为标记化(tokenization),在本实施例中,可使用Jieba库完成分词功能;进一步地,对上述分句分词后的样本文本数据进行去除停用词处理,其中,停用词是指在文本中频繁出现但并不含有太多信息的词,如,“的”、“是”以及“和”等,本实施例通过预先定义的停用词列表,将样本文本数据中的这些停用词去除;最后,将样本文件数据中的一些无意义的符号去除,例如,标点符号、数字以及其它非字母字符,这些符号对于新词的提取通常没有太大的帮助。上述这些文本预处理操作,都是为了将原始的样本文本数据转化为适合深度学习模型处理的输入数据。需要说明的

是,在实际领域新词提取场景中,也可对获取到的原始文本数据进行上述文本预处理过程,使得输入至领域新词提取模型的文本数据更适合模型输入。

[0079] 进一步地,通过领域新词标签,标记出各个所述文本预处理后的样本文本数据中的样本领域新词,构建得到各种所述行业领域的样本文本数据集。

[0080] 在本发明中,对文本预处理后的样本文本数据的词语进行标记,标记对象为在通用语料库中不常见但在特定领域中常见的词汇。通过对样本文本数据进行分析,可以识别出这些样本领域新词(可以理解的是,这些样本领域新词为历史领域新词,已经在历史时刻的样本文本数据中确定为领域新词),并利用领域新词标签将它们进行标记。当构建样本文本数据集时,能够清楚地知道哪些词汇是领域新词,从而更好地处理和分析这些数据。

[0081] 通过所述样本文本数据集,对预训练的BERT模型进行训练,若训练结果满足预设条件,得到所述领域新词提取模型。

[0082] 在本发明中,首先可根据实际需求(如基于数据量、计算资源和任务需求),确定用于新词提取的BERT(Bidirectional Encoder Representations from Transformers)模型版本,包括BERT-Base、BERT-Large和BERT-Mini等,每个版本的模型参数和结构略有不同,本发明采用BERT-Base,参数数量在1.1亿,适合几十万到几百万量级的文档数据集,从而满足领域新词抽取的任务要求。优选地,在一实施例中,可使用在大规模语料库上预训练过的BERT模型进行初始化,从而利用预训练模型学习到的语言知识,加速模型训练,并提高模型性能。

[0083] 进一步地,在训练过程中,将预处理后的样本文本数据输入到BERT模型之前,通过三个嵌入层,即Token Embeddings、Segment Embeddings和Position Embeddings,将输入转换为嵌入向量,进而再将转换后的嵌入向量书让人至BERT模型,其中,Token Embeddings是将每个词转换为对应的词向量;Segment Embeddings用于区分不同的句子;Position Embeddings用于表示词在句子中的位置。

[0084] 然后,BERT模型通过训练数据,学习上下文信息,预测可能出现新词的位置,模型训练的目标是最小化模型预测和实际结果之间的差距。在本发明中,训练过程可使用梯度下降,梯度下降方法与关键指标推荐值如下:学习率(推荐值为0.001),批次大小(推荐值为64),迭代次数(推荐值为100,配合早停策略),优化器(推荐使用Adam),损失函数(推荐均方误差)。

[0085] 进一步地,为了防止模型过拟合,本发明在训练过程中使用了dropout与weight decay正则化策略,以及学习率衰减策略,结合历史新词词库对训练后的BERT模型进行滚动微调,使模型能更好的识别和提取新词,并更有效的更新新词词库。同时,在本发明中,通过在训练过程中定期使用验证集测试模型的性能,以监控模型训练的进度,并通过模型在验证集上的表现来调整和优化模型的参数。

[0086] 本发明利用BERT模型构建领域新词提取模型,从而对大规模未标注的文本数据进行动态的新词提取,相较于传统的基于词频和信息熵的新词提取方法,BERT模型可以更好地理解词的上下文信息,从而更准确地提取新词;同时,利用历史新词词库对BERT模型进行滚动微调,使模型能更好的识别和提取新词,并更有效的更新新词词库。

[0087] 在上述实施例的基础上,所述根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型,包括:

[0088] 将各个所述候选领域新词和所述目标文本数据中其它目标词语作为节点,并根据所述目标文本数据的上下文信息,判断所述目标文本数据中的各个所述节点之间在是否存在共现关系;

[0089] 若判断获知存在所述共现关系,将所述共现关系作为所述节点之间的边,并根据各个所述边两端的所述节点在所述目标文本数据中的共现频率,确定各个所述边的权重,构建得到所述领域新词图模型。

[0090] 在本发明中,将所有候选领域新词和目标文本数据中的其它词语(即非候选领域新词)定义为待构建的领域新词图模型中的节点,每个节点代表一个唯一的词语。然后,定义节点之间的连接关系,也就是边的构建,在本发明中,将词语(包括候选领域新词和非候选领域新词)在目标文本数据中的共现关系作为边的依据,即如果两个词语在同一个上下文环境(如同一句、同一段或同一篇文章)中出现,就在这两个节点之间建立一条边,从而获取到目标文本数据中的词语共现信息,反映出词语之间的相关性。

[0091] 进一步地,还需要为每条边赋予一个权重,权重的大小通常取决于两个词语在目标文本数据中的共现频率,即共现频率越高,权重越大,表示两个词语的关联性更强。在一实施例中,也可使用词频-逆文本频率指数(Term Frequency-Inverse Document Frequency,简称TF-IDF)作为权重,衡量词在文档中的重要程度,从而降低一些常见词语的权重,同时提高一些稀有但重要的词语的权重。

[0092] 最后,基于上述步骤定义的节点和边,对于目标文本数据,构建得到一个由候选领域新词和目标文本数据中所有其它词语构成的图模型,即领域新词图模型,该模型详细地描述了词语之间的共现关系,以用于后续基于图算法的领域新词及其关联词的分析。

[0093] 在上述实施例的基础上,所述根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词,包括:

[0094] 基于PageRank算法,根据各个所述节点在所述领域新词图模型中的位置,以及各个所述节点之间的连接关系,对所述领域新词图模型中的各个所述节点进行重要性评估,得到所述领域新词图模型中各个所述节点的重要性得分;

[0095] 将所述领域新词图模型中所述重要性得分满足预设新词重要性阈值的所述候选领域新词确定为所述目标领域新词。

[0096] 在本发明中,可通过图算法对目标文本数据对应的领域新词图模型中的每个节点进行评估,图算法可选取网页排名(PageRank)算法或基于超链接的主题搜索(Hyperlink Induced Topic Search,简称HITS)。优选地,在本发明中,采用PageRank算法可以更有效地评估节点的重要性,相比于HITS算法,PageRank算法更适合候选领域新词图模型的评估过程。

[0097] 进一步地,通过PageRank算法对候选领域新词图模型中的每个节点进行评估,候选领域新词图模型会根据节点在候选领域新词图模型中的位置以及与其它节点的连接关系,计算出每个节点的重要性得分。在这个过程中,若某个节点与更多其它节点相连接,或者与重要性更高的节点相连接,该节点通常会获得更高的得分。

[0098] 在一实施例中,设置一个预设新词重要性阈值,只有重要性得分超过这个阈值的词语才会被认为是新词,其中,预设新词重要性阈值的设定通常需要根据实际应用的需求和数据的特性来确定,例如,如果希望提取更多的领域新词,可以设定较低的阈值;如果更

关心新词的精确性,可以设定较高的阈值。在一实施例中,将所有候选领域新词的重要性得分的第75百分位数作为预设新词重要性阈值。

[0099] 在上述实施例的基础上,所述获取待进行领域新词提取的目标文本数据,包括:

[0100] 在预设文本数据采集时段内,获取多个所述目标文本数据,且各个所述目标文本数据之间的文本内容是不同的,以通过各个所述目标文本数据,构建得到对应的所述领域新词图模型;

[0101] 在所述基于PageRank算法,根据各个所述节点在所述领域新词图模型中的位置,以及各个所述节点之间的连接关系,对所述领域新词图模型中的各个所述节点进行重要性评估,得到所述领域新词图模型中各个所述节点的重要性得分之后,所述方法还包括:

[0102] 获取各个所述目标文本数据对应的所述领域新词图模型;

[0103] 通过图对齐技术和图嵌入技术,将所有所述领域新词图模型进行融合处理,得到领域新词融合图;

[0104] 获取所述领域新词融合图中各个所述候选领域新词对应的词关联关系对,其中,所述词关联关系对是由所述领域新词融合图中存在所述边的所述候选领域新词形成的;

[0105] 根据所述领域新词融合图中各个所述节点的所述重要性得分,获取各个所述词关联关系对的词关联得分;

[0106] 将所述词关联得分满足预设词关联阈值的所述词关联关系对确定为目标词关联关系对,并根据所述目标词关联关系对,生成对应的词云;

[0107] 在所述根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词之后,所述方法还包括:

[0108] 将所述目标领域新词和所述词云存储至领域新词库。

[0109] 在本发明中,若在预设文本数据采集时段内获取到多个目标文本数据(这些目标文本数据可根据实际需求进行选取,如尽量选取领域较为接近的文本数据),进而可通过上述步骤,获取到各个目标文本数据对应的领域新词图模型。

[0110] 进一步地,在完成每个目标文本数据对应的领域新词图模型的重要性分析后,进而通过使用图对齐、图嵌入的技术,对各个领域新词图模型进行节点对齐、边融合和新增点边的操作,从而将所有领域新词图模型进行整合,融合成一个统一的图模型,即领域新词融合图。

[0111] 在本发明中,基于图对齐技术和图嵌入技术,将各个领域新词图模型中的相似节点进行匹配和对齐,从而将不同领域新词图模型中具有相似语义的节点进行匹配,找到它们之间的对应关系;然后,将各个领域新词图模型中的边进行合并,其中,边融合可以基于边的权重、共现频率或其他度量指标来进行,以保留各个领域中边的信息,并在整合后的图模型中形成更全面的关系网络;最后,可以根据需要对整合后的图模型进行新增点边的操作,如根据新的目标文本数据预测得到的新候选领域新词或新的共现关系,从而可以通过添加额外的节点或边来丰富图的信息。通过以上步骤,本发明可以将各个领域新词图模型进行整合,融合成一个统一的领域新词融合图,整合后的领域新词融合图将提供更全面、丰富的语义信息和知识支持。

[0112] 进一步地,遍历领域新词融合图中的每条边,判断边两端的两个节点是否包含候选领域新词,如果包含,将这两个节点作为词关联关系对中的一对词汇,形成一个关系对,

重复该步骤,直到遍历完所有边,从而获取领域新词融合图中各个候选领域新词对应的词关联关系对。这些关系对由在融合图中存在边的候选领域新词形成,可以用于进一步分析和处理词汇间的关联关系。

[0113] 进一步地,基于上述实施例中获取到的各个节点的重要性得分来计算词关联得分。在一实施例中,可将连接到该词关联关系对的两个节点的重要性得分进行加权平均,从而获取词关联关系对的词关联得分,这些得分可以用于量化词关联关系对之间的相关程度。

[0114] 在获取领域新词融合图中各个所述词关联关系对的词关联得分之后,根据预设词关联阈值,筛选出满足条件的词关联关系对作为目标词关联关系对,即选择词关联得分高于或等于预设阈值的关系对,只有得分超过这个阈值的词关联关系对,才会被认为是有显著关联的词语对。具体地,这个预设词关联阈值的设定同样取决于实际应用的需求和数据的特性,例如,如果希望提取出更强烈的词语关系,可以设定较高的阈值;如果更关注词语关系的广度,可以设定较低的阈值。在一实施例中,可使用Louvain社区检测算法找到社区值中位数,作为预设词关联阈值。

[0115] 在本发明中,通过目标词关联关系对,可以生成对应的词云。词云是一种可视化工具,用来展示文本数据中词汇的频率或重要性。在生成词云时,可以使用目标词关联关系对中的词汇作为输入,根据词关联得分来决定词汇的大小、颜色等属性。最后,将生成的词云进行可视化展示,这可以是一个静态图像,也可以是一个交互式的词云,可以根据用户的操作进行缩放、旋转等操作,从而基于每个节点的重要性得分,形成一张多个新词与关联词相连的相关性词云。

[0116] 图2为本发明提供的基于图算法的领域新词的关系构建过程示意图,可参考图2所示,本发明将图算法应用于领域新词提取过程中,通过构建领域新词共现网络并应用图算法,使用图对齐、图嵌入的技术,能够更全面、更准确地评估和确定领域新词,能通过灵活地调整新词提取的数量和精度,从而满足不同的应用需求。

[0117] 最后,将阈值排除后的领域新词与词云结果进行人工审核筛选,这样,这些新词不仅在文本中频繁出现,而且与其它领域新词有着密切的共现关系,因此可被认为是具有代表性的领域新词。进一步地,将得到的领域新词与词相关关系(即词云)记录进领域新词库,这个领域新词库将作为自然语言处理任务的重要资源,为后续的新词提取、词法分析、语义检索等任务提供支持;同时,领域新词的入库也意味着可不断地更新和完善领域知识,实时跟踪和掌握领域的最新动态和发展趋势。

[0118] 在上述实施例的基础上,所述根据各个所述目标词语的所述新词概率,从多个所述目标词语中确定候选领域新词,包括:

[0119] 基于所述新词概率从大到小的顺序,将所有所述目标词语进行排序,并根据排序结果,选取前N个所述目标词语作为待定词语;

[0120] 获取各个所述待定词语在所述目标文本数据中的出现次数,并判断所述出现次数是否大于预设频次阈值,若大于,则将所述待定词语确定为所述候选领域新词,其中,所述预设频次阈值是根据所述领域新词库中所有领域新词在所有历史文本数据中的出现次数生成的;所述历史文本数据为所述领域新词库中在历史时刻新增领域新词时,获取到的新增的领域新词对应的文本数据;

[0121] 在本发明中,将预处理后的目标文本数据输入到领域新词提取模型模型中,领域新词提取模型模型会为目标文本中的每个位置的词预测该词为一个新词的概率;然后,对于模型预测的每个位置,将其对应的词汇抽取出来,按照新词概率从大到小排序,并选取排序靠前的N个目标词语作为待定词语。在一实施例中,还需要对待定新词的长度进行限制,以避免抽取出过长的领域新词,例如,取当前领域新词词库中最长的词字符长度加一为长度限制上限,下限设定为2,通常范围是2~7,具体长度可以根据实际需求进行调整。

[0122] 进一步地,统计每个待定新词在目标文本数据中的出现次数,并设定一个频次阈值,即预设频次阈值,只有出现频次超过这个阈值的待定新词才会被作为候选领域新词保留。在一实施例中,预设频次阈值的设定,可基于历史新词词库中所有词在所在的历史文本数据中出现次数的平均数的一半,作为预设频次阈值。

[0123] 在所述将所述目标领域新词和所述词云存储至领域新词库之后,所述方法还包括:

[0124] 将所述目标领域新词对应的目标文本数据存储至所述领域新词库,得到更新后的领域新词库;

[0125] 基于所述更新后的领域新词库,对所述预设频次阈值进行更新,得到更新后的预设频次阈值。

[0126] 在本发明中,预设频次阈值随着领域新词词库的更新是动态变化的,一个真正的领域新词通常会在文本中多次出现,而不是偶尔出现一两次。本发明通过设置动态频次阈值,可以过滤掉大量的偶然出现的、无意义的候选领域新词,提高领域新词的提取准确性。

[0127] 在上述实施例的基础上,在所述获取各个所述待定词语在所述目标文本数据中的出现次数,并判断所述出现次数是否大于预设频次阈值,若大于,则将所述待定词语确定为所述候选领域新词之后,所述方法还包括:

[0128] 将获取到的多个所述候选领域新词构建为候选领域新词集合,并对所述候选领域新词集合进行筛选处理,所述筛选处理包括:

[0129] 判断所述候选领域新词集合中任意一个候选领域新词是否属于其它候选领域新词的子串,若属于,则将属于所述子串的候选领域新词从所述候选领域新词集合中删除,得到筛选处理后的候选领域新词集合;

[0130] 所述根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型,包括:

[0131] 根据所述筛选处理后的候选领域新词集合中的各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型。

[0132] 在本发明中,在候选领域新词中,可能存在一些候选领域新词是另一个候选领域新词的子串,例如,如果“深度学习”和“学习”都是候选领域新词,那么“学习”就是“深度学习”的子串。在这种情况下,需要保留更长的候选领域新词,即“深度学习”,并过滤掉更短的候选领域新词,即“学习”。

[0133] 在一实施例中,对于在当前领域新词库中已经出现的候选领域新词,可对这些候选领域新词进行标记,这些词虽然已入库,但新的文本可能会给提供新的词关系信息,对于后续的领域新词提取过程来说仍有价值。通过上述筛选处理操作,可以进一步提高候选领域新词的质量,为后续的新词词关系构建和入库做好准备。

[0134] 图3为本发明提供的基于深度学习模型的候选领域新词的提取过程示意图,可参考图3所示,本发明利用BERT深度学习模型,可以更好地理解词的上下文信息,从而更准确地提取领域新词。在一实施例中,通过将本发明提供的领域新词提取方法与现有领域新词提取方法进行实验评估,其评估结果显示本方法在新词提取任务上的准确率比现有词频统计方法提高了约9.5%。计算方式为:(本方法准确率92%-词频统计方法准确率84%)/84%*100%。

[0135] 而在模型训练策略中,本发明通过历史新词词库对深度学习模型进行滚动微调,可以有效地更新新词词库,使模型的新词提取能力随时间持续提升。在实验中,滚动微调策略使模型的新词提取性能相比于固定训练策略提升了约16%。计算方式为:(滚动微调策略预测准确率87%-固定训练策略预测准确率75%)/75%*100%。

[0136] 并且,本发明通过设定动态的频次阈值对候选领域新词进行过滤,可以有效过滤掉偶然出现的、无意义的候选领域新词。在实验中,动态频次过滤方法比固定阈值过滤方法减少了约10%的无效新词,并避免了5%左右的有效新词被过滤。

[0137] 除此之外,在后续的候选领域新词的分析过程中,本发明创新性地引入图模型和图算法,不仅可用于新词的提取,而且可以分析新词与其他词之间的关系,形成一张词相关性云图,进一步丰富新词的语义信息。

[0138] 下面对本发明提供的领域新词提取装置进行描述,下文描述的领域新词提取装置与上文描述的领域新词提取方法可相互对应参照。

[0139] 图4为本发明提供的领域新词提取装置的结构示意图,如图4所示,本发明提供了一种领域新词提取装置,包括文本数据输入模块401、预测模块402、候选新词筛选模块403、图模型构建模块404和图模型处理模块405,其中,文本数据输入模块401用于获取待进行领域新词提取的目标文本数据;预测模块402用于将所述目标文本数据输入到领域新词提取模型中,得到由所述领域新词提取模型输出的所述目标文本数据中各个目标词语作为领域新词的新词概率,其中,所述领域新词提取模型是由标记有领域新词标签的样本文本数据,对自然语言处理模型进行训练得到的;候选新词筛选模块403用于根据各个所述目标词语的所述新词概率,从多个所述目标词语中确定候选领域新词;图模型构建模块404用于根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型;图模型处理模块405用于根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词。

[0140] 本发明提供的领域新词提取装置,通过由自然语言处理模型训练得到的领域新词提取模型,对文本数据中的词语进行新词概率预测,再根据预测得到的新词概率,选取候选领域新词进行领域新词图模型构建,最终通过领域新词图模型获取文本数据中的领域新词,从而更为准确地提取文本中的领域新词。

[0141] 本发明提供的装置是用于执行上述各方法实施例的,具体流程和详细内容请参照上述实施例,此处不再赘述。

[0142] 图5为本发明提供的电子设备的结构示意图,如图5所示,该电子设备可以包括:处理器(Processor)501、通信接口(Communications Interface)502、存储器(Memory)503和通信总线504,其中,处理器501,通信接口502,存储器503通过通信总线504完成相互间的通信。处理器501可以调用存储器503中的逻辑指令,以执行领域新词提取方法,该方法包括:

获取待进行领域新词提取的目标文本数据;将所述目标文本数据输入到领域新词提取模型中,得到由所述领域新词提取模型输出的所述目标文本数据中各个目标词语作为领域新词的新词概率,其中,所述领域新词提取模型是由标记有领域新词标签的样本文本数据,对自然语言处理模型进行训练得到的;根据各个所述目标词语的所述新词概率,从多个所述目标词语中确定候选领域新词;根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型;根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词。

[0143] 此外,上述的存储器503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。

[0144] 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方法所提供的领域新词提取方法,该方法包括:获取待进行领域新词提取的目标文本数据;将所述目标文本数据输入到领域新词提取模型中,得到由所述领域新词提取模型输出的所述目标文本数据中各个目标词语作为领域新词的新词概率,其中,所述领域新词提取模型是由标记有领域新词标签的样本文本数据,对自然语言处理模型进行训练得到的;根据各个所述目标词语的所述新词概率,从多个所述目标词语中确定候选领域新词;根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型;根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词。

[0145] 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领域新词提取方法,该方法包括:获取待进行领域新词提取的目标文本数据;将所述目标文本数据输入到领域新词提取模型中,得到由所述领域新词提取模型输出的所述目标文本数据中各个目标词语作为领域新词的新词概率,其中,所述领域新词提取模型是由标记有领域新词标签的样本文本数据,对自然语言处理模型进行训练得到的;根据各个所述目标词语的所述新词概率,从多个所述目标词语中确定候选领域新词;根据各个所述候选领域新词在所述目标文本数据中的共现关系,构建领域新词图模型;根据所述领域新词图模型,从多个所述候选领域新词中确定目标领域新词。

[0146] 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。

[0147] 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。

[0148] 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。

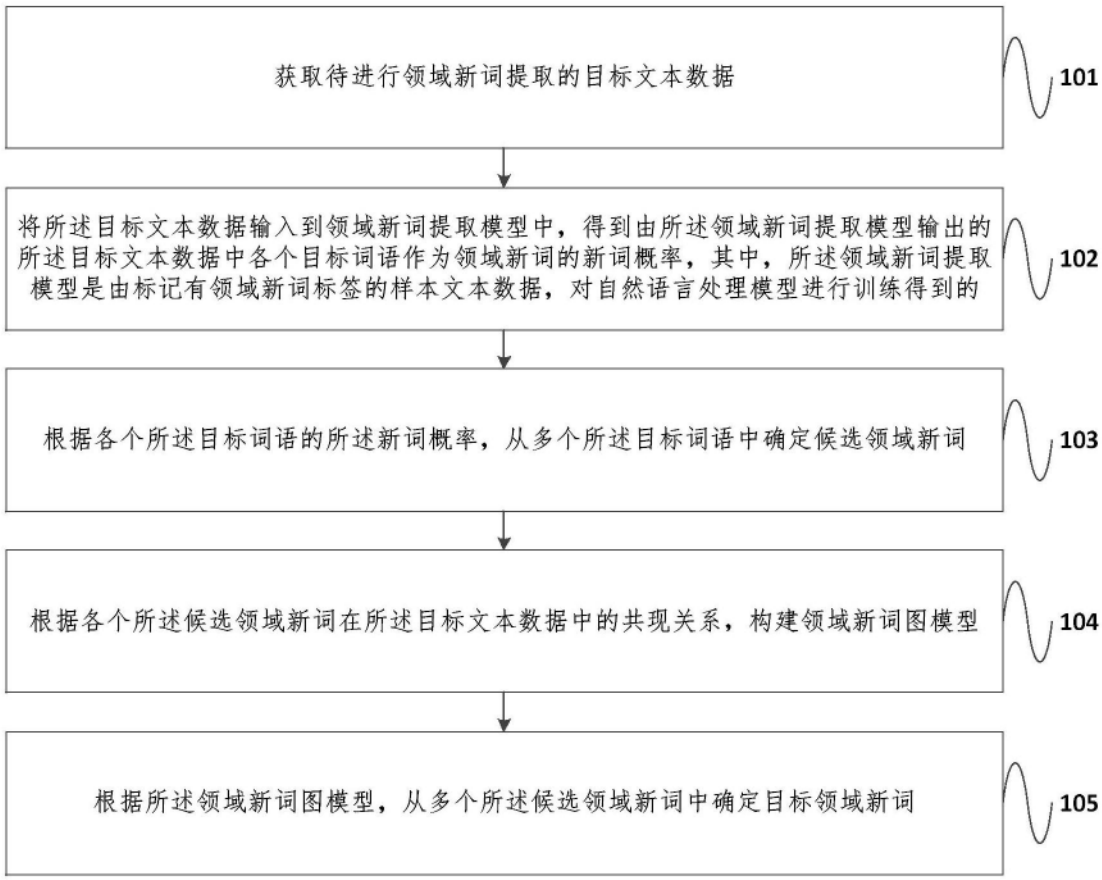

获取待进行领域新词提取的目标文本数据
101
将所述目标文本数据输入到领域新词提取模型中，得到由所述领域新词提取模型输出的所述目标文本数据中各个目标词语作为领域新词的新词概率，其中，所述领域新词提取模型是由标记有领域新词标签的样本文本数据，对自然语言处理模型进行训练得到的
102
根据各个所述目标词语的所述新词概率，从多个所述目标词语中确定候选领域新词
103
根据各个所述候选领域新词在所述目标文本数据中的共现关系，构建领域新词图模型
104
根据所述领域新词图模型，从多个所述候选领域新词中确定目标领域新词
105

图1

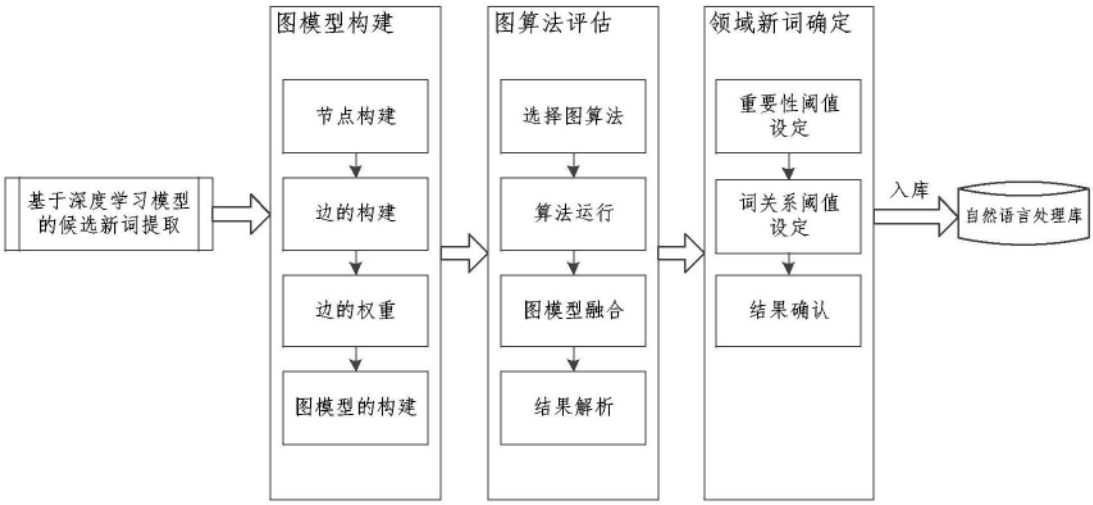

基于深度学习模型的候选新词提取
图模型构建
节点构建
边的构建
边的权重
图模型的构建
图算法评估
选择图算法
算法运行
图模型融合
结果解析
领域新词确定
重要性阈值设定
词关系阈值设定
结果确认
入库
自然语言处理库

图2

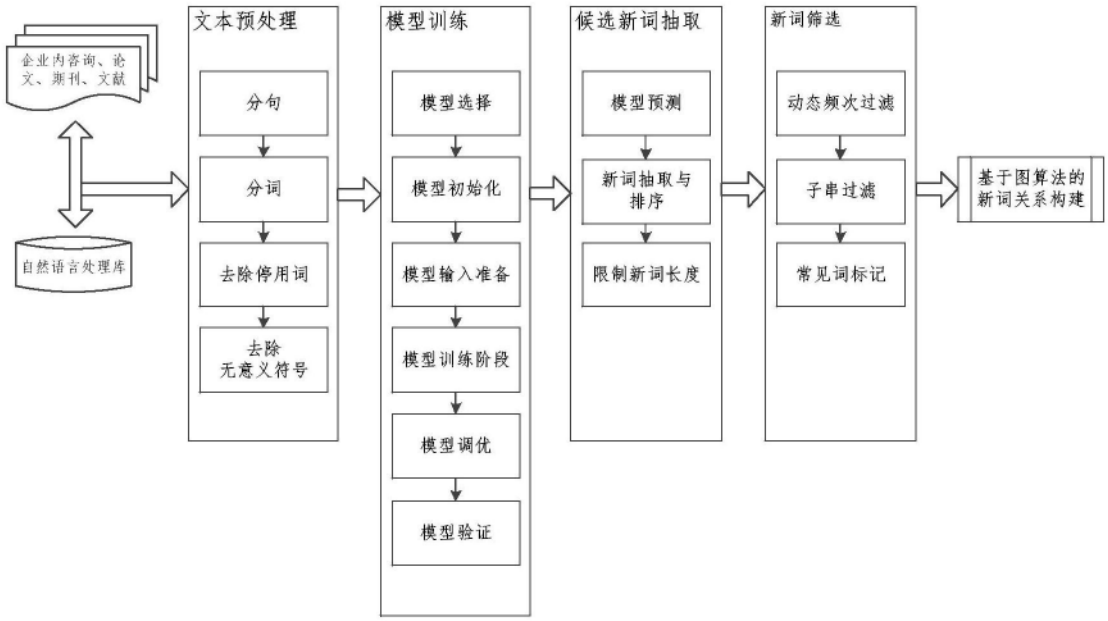

企业内咨询、论文、期刊、文献
自然语言处理库
文本预处理
分句
分词
去除停用词
去除无意义符号
模型训练
模型选择
模型初始化
模型输入准备
模型训练阶段
模型调优
模型验证
候选新词抽取
模型预测
新词抽取与排序
限制新词长度
新词筛选
动态频次过滤
子串过滤
常见词标记
基于图算法的新词关系构建

图3

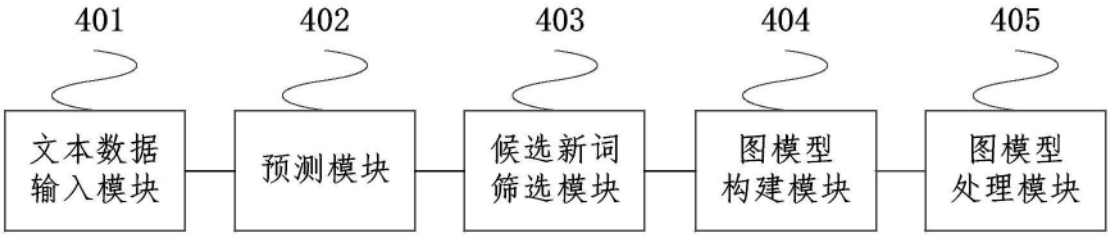

401
402
403
404
405
文本数据输入模块
预测模块
候选新词筛选模块
图模型构建模块
图模型处理模块

图4

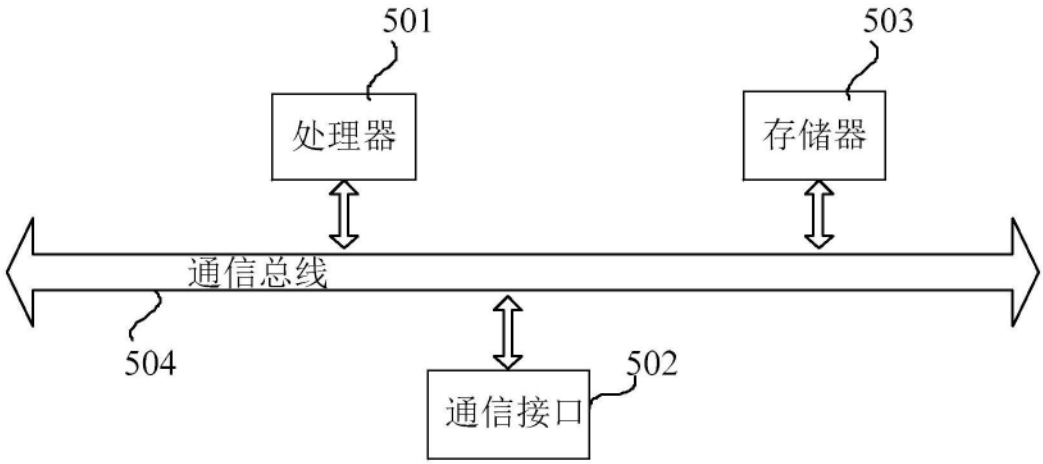

501
503
处理器
存储器
通信总线
504
502
通信接口

图5