(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108255857 B
(45)授权公告日 2021.10.15

(21)申请号 201611246151.0
(22)申请日 2016.12.29
(65)同一申请的已公布的文献号
申请公布号 CN 108255857 A
(43)申请公布日 2018.07.06
(73)专利权人 北京国双科技有限公司
地址 100083 北京市海淀区北四环中路229号海泰大厦4层南401号
(72)发明人 王天祎
(74)专利代理机构 北京鼎佳达知识产权代理事务所(普通合伙) 11348
代理人 王伟锋 刘铁生
(51)Int.Cl.
G06F 16/9535(2019.01)
G06F 40/30(2020.01)

审查员 安飞

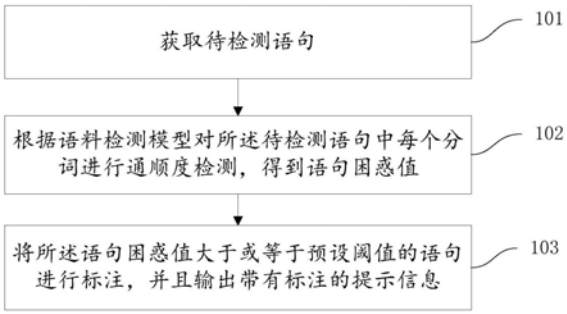
权利要求书2页 说明书9页 附图4页

(54)发明名称

一种语句检测方法及装置

(57)摘要

本发明公开了一种语句检测方法及装置,涉及互联网技术领域,主要目的是提高广告语句检测结果的准确性。所述方法包括:获取待检测语句,根据语料检测模型对所述待检测语句中每个分词进行通顺度检测,得到语句困惑值,所述语料检测模型记录有不同领域自然文本中的分词、各个分词对应的频率以及各个分词在不同语句中的组合概率,将所述语句困惑值大于或等于预设阈值的语句进行标注,并且输出带有标注的提示信息。本发明主要用于广告语句的检测。



CN 108255857 B

1.一种语句检测方法,其特征在于,包括:

获取待检测语句;

根据语料检测模型对所述待检测语句中每个分词进行通顺度检测,得到语句困惑值,所述语料检测模型记录有不同领域自然文本中的分词、各个分词对应的频率以及各个分词在不同语句中的组合概率,具体包括:

将所述待检测语句进行划分,得到所述待检测语句的各个分词;根据所述语料检测模型获取所述待检测语句中各个分词在所述待检测语句中的组合概率;根据所述组合概率对所述待检测语句进行评估,得到所述待检测语句的语句困惑值;

其中,所述根据所述组合概率对所述待检测语句进行评估,得到所述待检测语句的语句困惑值包括:

将所述各个分词在所述待检测语句中的组合概率带入困惑值计算公式,得到所述待检测语句的语句困惑值;

所述困惑值计算公式为:

$$PP(W)=\sqrt[N]{\prod_{i=1}^{N}\frac{1}{P(w_i|w_{i-1})}}$$

其中,PP(W)为待检测语句的困惑值,$P(w_i|w_{i-1})$为待检测语句中第i个分词依赖于第i-1个分词的组合概率;

将所述语句困惑值大于或等于预设阈值的语句进行标注,并且输出带有标注的提示信息。

2.根据权利要求1所述的方法,其特征在于,在所述根据语料检测模型对所述待检测语句中每个分词进行通顺度检测,得到语句困惑值之前,所述方法还包括:

获取不同领域的自然文本作为样本数据;

对所述样本数据中的每条语句进行分词,得到样本分词集合;

采用语料训练模型对所述样本分词集合进行训练,得到语料检测模型。

3.根据权利要求1或2所述的方法,其特征在于,所述方法还包括:

将所述语句困惑值小于预设阈值的语句对应的广告进行展示。

4.一种语句检测装置,其特征在于,包括:

第一获取单元,用于获取待检测语句;

检测单元,用于根据语料检测模型对所述待检测语句中每个分词进行通顺度检测,得到语句困惑值,所述语料检测模型记录有不同领域自然文本中的分词、各个分词对应的频率以及各个分词在不同语句中的组合概率;

所述检测单元包括:划分模块,用于将所述待检测语句进行划分,得到所述待检测语句的各个分词;获取模块,用于根据所述语料检测模型获取所述待检测语句中各个分词在所述待检测语句中的组合概率;估计模块,用于根据所述组合概率对所述待检测语句进行评估,得到所述待检测语句的语句困惑值;

估计模块,具体用于将所述各个分词在所述待检测语句中的组合概率带入困惑值计算公式,得到所述待检测语句的语句困惑值;

所述困惑值计算公式为:

$$PP(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_{i-1})}}$$

其中，PP(W)为待检测语句的困惑值，$P(w_i|w_{i-1})$为待检测语句中第i个分词依赖于第i-1个分词的组合概率；

输出单元，用于将所述语句困惑值大于或等于预设阈值的语句进行标注，并且输出带有标注的提示信息。

5.根据权利要求4所述的装置，其特征在于，所述装置还包括：

第二获取单元，用于获取不同领域的自然文本作为样本数据；

分词单元，用于对所述样本数据中的每条语句进行分词，得到样本分词集合；

训练单元，用于采用语料训练模型对所述样本分词集合进行训练，得到语料检测模型。

6.根据权利要求4或5所述的装置，其特征在于，所述装置还包括：

展示单元，用于将所述语句困惑值小于预设阈值的语句对应的广告进行展示。

7.一种存储介质，其特征在于，所述存储介质包括存储的程序，其中，在所述程序运行时控制所述存储介质所在设备执行权利要求1至权利要求3中任意一项所述的语句检测方法。

8.一种处理器，其特征在于，所述处理器用于运行程序，其中，所述程序运行时执行权利要求1至权利要求3中任意一项所述的语句检测方法。

一种语句检测方法及装置

技术领域

[0001] 本发明涉及互联网技术领域，尤其是一种语句检测方法及装置。

背景技术

[0002] 搜索引擎营销(Search Engine Marketing，SEM)是根据用户使用搜索引擎的方式利用用户检索信息的机会尽可能将营销信息传递给用户。广告主网站通常会在搜索引擎平台上投放SEM广告，当用户触发广告关键词时会在网页中展示多个广告商的网站信息，用户可以根据显示的网站信息选择感兴趣广告语句点击进入网站，进一步了解所需要的信息，因此，创意好的广告语句会吸引更多的用户点击，从而能够为广告商实现更好的推广。

[0003] 由于在SEM广告在投放至搜索引擎平台之前，搜索引擎平台需要对广告语句的通顺度进行检测，若广告语句不通顺则搜索引擎平台不予展示该SEM广告。目前，通常使用人工操作对每条待投放SEM广告的广告语句进行通顺度检测，进而保证SEM广告能够顺利展示。

[0004] 然而，在通过人工操作进行通顺度检查的过程中，由于每个人的先验知识不同，并且对广告语句通顺度的理解也缺乏统一的标准，使得检测结果不准确，导致部分能够顺利展示的广告无法在页面中展示，给广告商带来了经济损失。

发明内容

[0005] 鉴于上述问题，提出了本发明以便提供一种克服上述问题或者至少部分地解决上述问题的一种语句检测方法及装置，能够提高广告语句检测结果的准确性。

[0006] 一方面，本发明提供了一种语句检测方法，包括：

[0007] 获取待检测语句；

[0008] 根据语料检测模型对所述待检测语句中每个分词进行通顺度检测，得到语句困惑值，所述语料检测模型记录有不同领域自然文本中的分词、各个分词对应的频率以及各个分词在不同语句中的组合概率；

[0009] 将所述语句困惑值大于或等于预设阈值的语句进行标注，并且输出带有标注的提示信息。

[0010] 另一方面，本发明提供一种语句检测装置，包括：

[0011] 第一获取单元，用于获取待检测语句；

[0012] 检测单元，用于根据语料检测模型对所述待检测语句中每个分词进行通顺度检测，得到语句困惑值，所述语料检测模型记录有不同领域自然文本中的分词、各个分词对应的频率以及各个分词在不同语句中的组合概率；

[0013] 输出单元，用于将所述语句困惑值大于或等于预设阈值的语句进行标注，并且输出带有标注的提示信息。

[0014] 借由上述技术方案，本发明提供的一种语句检测方法及装置，首先获取待检测语句，这里的待检测语句为SEM广告中的广告语句，然后根据语料检测模型对待检测语句中每

个分词进行通顺度检测，得到语句困惑值，进而对广告语句的通顺度进行评估，将语句困惑值大于或等于预设阈值的广告语句对应的SEM广告进行标注，并且输出带有标注的提示信息，提示该广告语句对应的SEM广告不符合展示条件。相对于现有的通过人工操作对每个待展示SEM广告的广告语句进行检测的方法相比，本发明通过从大数据积累的语料库建立的语料检测模型对所述待检测语句进行通顺度检测，无需人工操作，能够批量快速检测语句的通顺度，提高了广告语句检测结果的准确性，进一步通过将不符合展示条件的SEM广告进行预警，避免了由于广告上线不通顺给广告主带来的经济损失。

[0015] 上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

[0016] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0017] 图1示出了本发明实施例提供的一种语句检测方法流程示意图；

[0018] 图2示出了本发明实施例提供的另一种语句检测方法流程示意图；

[0019] 图3示出了本发明实施例提供的一种语句检测装置结构示意图；

[0020] 图4示出了本发明实施例提供的另一种语句检测装置结构示意图。

具体实施方式

[0021] 下面将参照附图更详细地描述本公开的示例性实施例。虽然附图中显示了本公开的示例性实施例，然而应当理解，可以以各种形式实现本公开而不应被这里阐述的实施例所限制。相反，提供这些实施例是为了能够更透彻地理解本公开，并且能够将本公开的范围完整的传达给本领域的技术人员。

[0022] 本发明实施例提供了一种语句检测方法，如图1所示，该方法主要用于对广告语句进行通顺度检测，具体步骤包括：

[0023] 101、获取待检测语句。

[0024] 其中，待检测语句为SEM广告中的广告语句，当用户在搜索栏中输入检索词后，搜索引擎平台会将用户检索相关的信息展示在网页中，以供用户选择，这里的SEM广告利用用户对搜索引擎的依赖和使用习惯，在用户检索信息的时候将广告信息传递给目标用户。

[0025] 本发明实施例对获取待检测语句的方法不做限定，本步骤中可以在展示SEM广告中之前通过在网页内部署脚本来获取待检测语句，还可以通过存储有SEM广告的数据库中获取待检测语句。

[0026] 另外，由于SEM广告涉及多个领域，如金融领域、电子领域、医疗领域等等，对于本发明实施例可以在获取到待检测语句后，对涉及不同领域的待检测语句进行归类，以方便对不同领域的广告语句进行检测。

[0027] 102、根据语料检测模型对所述待检测语句中每个分词进行通顺度检测，得到语句困惑值。

[0028] 其中，语料检测模型为以从数据库中抽取不同领域自然语言文本作为基础训练语料库，进而对基础语料库中语句的分词进行训练得到的语料检测模型，因此该语料检测模型记录不同领域自然文本中的分词、各个分词的频率以及各个分词在不同语句中的组合概率，当然也包括有待检测语句中的分词、各个分词的频率以及各个分词在待检测语句中出现的组合概率，这里各个分词的频率可以是单个词或字在语料库中的频率值，这里各个分词在不同语句中的组合概率可以为当前分词与语句中前一分词或者语句中后一分词的组合的概率，两个分词的组合概率越大说明两个分词组合的通顺度越高。

[0029] 本发明实施例中的语句困惑值为经过量化的数值，具体通过从语料检测模型中获取与待检测语句中各个分词在待检测语句中的组合概率，结合各个分词在待检测语句中的组合概率对待检测语句进行评估，得到待检测语句的语句困惑值，这里的困惑值用于表示该条语句的通顺度程度，其中，语句困惑值越大代表语句越不通顺。

[0030] 103、将所述语句困惑值大于或等于预设阈值的语句进行标注，并且输出带有标注的提示信息。

[0031] 由于SEM广告中包括有广告语句，该广告语句的通顺程度涉及到该SEM广告是否能顺利展示，因此本发明实施例通过设置预设阈值来对SEM广告进行筛选，将不符合展示条件的SEM广告进行预警。对于本发明实施例，对语句困惑值大于或等于预设阈值的广告语句对应的SEM广告提示不符合展示条件，并且进行广告预警，这里的预警方式可以为将该广告语句对应的SEM广告进行标注后，以反馈至广告商进行修正，还可以为将该广告语句对应的SEM广告进行标注后，发送至工作人员进行复查，能够批量快速的对SEM广告语句进行检测，提高了语句检测的准确性。

[0032] 本发明实施例对预设阈值不做限定，具体根据搜索引擎平台对SEM广告的展示条件进行设置，若搜索引擎平台对SEM广告的展示要求较高，则相应将预设阈值设置为较低数值，当语句的困惑值在较低的范围时，说明该SEM广告的通顺程度较高。

[0033] 结合上述的实现方式可以看出，本发明实施例提供的一种语句检测方法，首先获取待检测语句，这里的待检测语句为SEM广告中的广告语句，然后根据语料检测模型对待检测语句中每个分词进行通顺度检测，得到语句困惑值，进而对广告语句的通顺度进行评估，将语句困惑值大于或等于预设阈值的广告语句对应的SEM广告进行标注，并且输出带有标注的提示信息，提示该广告语句对应的SEM广告不符合展示条件。相对于现有的通过人工操作对每个待展示SEM广告的广告语句进行检测的方法相比，本发明通过从大数据积累的语料库建立的语料检测模型对所述待检测语句进行通顺度检测，无需人工操作，能够批量快速检测语句的通顺度，提高了广告语句检测结果的准确性，进一步通过将不符合展示条件的SEM广告进行预警，避免了由于广告上线不通顺给广告主带来的经济损失。

[0034] 以下为了更加详细地说明本发明提出的一种语句检测方法，特别是根据语料检测模型如何对待检测语句的通顺度检测的过程，本发明实施例还提供了另一种语句检测方法，如图2所示，该方法的具体步骤包括：

[0035] 201、获取待检测语句。

[0036] 这里的待检测语句为SEM广告中的广告语句，创意好的SEM广告除了可以更好的提升广告商企业的形象之外，还能带来更多优质的潜在客户，因此，SEM广告语句的通顺度对SEM广告至关重要。

[0037] 示例性的，当用户检索“买意外险”关键词后，在搜索页面展示出多条关于买意外险的SEM广告标题以及广告语句，如中国太平洋人寿保险的广告标题，并且标题下面附有描述广告的语句。

[0038] 通常情况下搜索引擎平台在SEM广告投放之前会获取待检测广告语句，并对待检测广告语句进行通顺度检测，以保证符合展示要求的SEM广告能够顺利展示，进而保证用户在浏览页面过程中能够更加清晰的寻找到所需要的内容，同时也对筛选出的不符合展示条件的SEM广告进行预警。

[0039] 202、获取不同领域的自然文本作为样本数据。

[0040] 对于本发明实施例，可以通过从数据库中抽取不同领域自然文本作为基础训练的样本数据，还可以通过预先存储的不同领域的广告文本作为基础训练的样本数据，本发明实施例不进行限定。

[0041] 203、对所述样本数据中的每条语句进行分词，得到样本分词集合。

[0042] 由于样本数据中包括不同领域的语句，并且该语句通常为比较通顺的语句，通过对样本数据中每条语句进行分词后可以得到样本分词集合，例如样本数据中的语句为“温馨母亲节”，对该语句进行分词得到“温馨”、“母亲节”。

[0043] 204、采用语料训练模型对所述样本分词集合进行训练，得到语料检测模型。

[0044] 其中，语料检测模型中记录有不同领域自然文本中的分词、各个分词的频率以及各个分词在不同语句中出现的组合概率，这里各个分词的频率可以是单个词或字在语料库中的频率值，这里各个分词在不同语句中的组合概率可以为当前分词与语句中前一分词或者语句中后一分词的组合的概率，两个分词的组合概率越大说明两个分词组合的通顺度越高。

[0045] 对于本发明实施例，当获取到样本数据后，可以为采用语言模型工具(Stanford Research Institute Language Modeling，Srilm)对样本数据进行训练，进而搭建语料检测模型。该模型通过采用最大似然估计及相应的平滑算法对样本数据进行处理，得到语料检测模型，并统计出样本数据中语句的分词、各个分词的频率以及各个分词的组合的概率，

[0046] 对于步骤202至步骤204，示例性的，通过从数据库中抽取目标领域大量的自然语言文本作为基础语料训练裤C，对基础语料训练库中每条句子进行分词，得到每个句子的样本分词集合S＝{分词1，分词2，分词3…}，将所有的句子的样本分词集合组成训练数据trainset，设定最大N-gram数N，N表示最长的关联词个数，利用Srilm语言模型工具对训练数据trainset进行训练，得到语料检测模型M，该语料检测模型中记录了出现过的分词，以及N个数以内分词的组合，例如对于“支持向量机”，得到的语料检测模型中记录有分词支持、向量、机、支持向量、向量机、支持向量机，以及每个分词对应的组合概率，本发明实施例对上述的语言模型工具的类型不做限定，还可以使用其他语言模型工具对样本数据进行训练。

[0047] 205、将所述待检测语句进行划分，得到所述待检测语句的各个分词。

[0048] 通常情况下，待检测语句中包括多个分词，由于各个分词的组合概率大小从一定程度上反应待检测语句的通顺度，因此本步骤中将待检测语句进行划分，以便于后续对该语句进行通顺度检测。

[0049] 示例性的，对于用户检索关键词为整容医院，则在网页中对应展示有SEM广告语句

“xxx整形医院，知名专家会诊，口碑良好”将该广告语句进行划分，得到多个分词包括整形、医院、知名、专家、会诊、口碑、良好。

[0050] 206、根据所述语料检测模型获取所述待检测语句中各个分词在所述待检测语句中的组合概率。

[0051] 由于在语料检测模型中记录有不同领域的语言文本中的分词、各个分词的频率以及各个分词的组合概率，因此，该语料检测模型中记录有待检测语句中各个分词的组合概率，每个分词与该语句中前一分词组合的概率会有所不同，概率越高则说明这两个分词组合后的通顺度越高，进一步从语料检测模型中获取与待检测语句中的各个分词在待检测语句中出现的组合概率。

[0052] 207、根据所述组合概率对所述待检测语句进行评估，得到所述待检测语句的语句困惑值。

[0053] 由于待检测语句中包括有多个分词，不同分词的组合概率有所不同，使得待检测语句的通顺度有所不同，进一步根据各个分词的组合概率对待检测语句进行评估，对于本发明实施例中可以通过将待检测语句中分词的组合概率带入困惑值计算公式，得到待检测语句的语句困惑值的方式来对待检测语句进行评估，，这里的语句困惑值计算公式可以为：

[0054] $$PP(W)=\sqrt[N]{\prod_{i=1}^{N}\frac{1}{P(w_i|w_{i-1})}}$$

[0055] 其中，PP(W)为待检测语句的困惑值，$P(w_i|w_{i-1})$为待检测语句中第i个分词依赖于第i-1个分词的组合概率，共有N个分词，这里困惑值能够反应语句的通顺度，若得到的语句困惑值越低，说明该语句越通容易被理解，通顺度越高，符合展示条件，反之，说明该语句不够通顺，不符合广告展示条件。

[0056] 这里的各个分词的组合概率大小从一定程度上反应待检测语句的通顺度，所以不同的分词组合后语句的困惑值会有所不同，例如，对于分词“好好”、“学习”，由于两个分词的组合的概率大小不同，所以组合成“好好学习”评估得到的语句困惑值相比组合成“学习好好”评估得到的语句困惑值要低，困惑值越低说明该语句通顺度越好，更易于用户了解该语句表达的意思。

[0057] 208、将所述语句困惑值大于或等于预设阈值的语句进行标注，并且输出带有标注的提示信息。

[0058] 对于广告语句的通顺程度涉及到该SEM广告是否能顺利展示，因此本发明实施例通过设置预设阈值来对SEM广告进行筛选，当语句困惑值大于预设阈值的广告语句对应的SEM广告不符合展示条件，无法在网页中进行展示，需要对该广告进行标注，并且输出提示信息，提示该广告不符合展示条件。本发明实施例对预设阈值的大小不做限定，具体可以通过搜索引擎平台进行设置，若对广告通顺度要求较高的平台，可以将预设阈值设置相对较小，以保证投放至搜索平台的广告语句的通顺度符合展示条件。

[0059] 209、将所述语句困惑值小于预设阈值的语句对应的广告进行展示。

[0060] 由于现有技术在SEM广告在投放至搜索引擎平台之前需要通过人工操作对每条待投放SEM广告的广告语句进行通顺度检测，进而保证SEM广告能够顺利展示，本发明实施例在通过步骤205中得到的语句困惑值后，该数值能够表明该条SEM广告中广告语句的通顺度，进一步将语句困惑值小于预设阈值的广告语句对应的SEM广告作为符合展示条件的广

告，进而在网页中展示。

[0061] 本发明实施例的具体步骤可以包括但不限制于下述实现方式：首先通过在网页部署脚本获取到SEM广告中的广告语句，当用户在搜索引擎中输入检索关键词“乐视手机”后，在网页展示有标题为xx网乐视手机官网，附带有广告语句“xx网提供乐视手机全面服务信息”，然后通过预先对不同领域的分词数据进行训练，建立语料检测模型C，该模型C中包括有不同领域语句S，即C＝{S1,S2,S3…}，以及语句S中的各个分词n，即S＝{n1,n2,n3…}，以及各个分词的频率、以及各个分词的组合概率，进一步将获取的SEM广告中的广告语句进行划分，得到多个分词有xx网、提供、乐视、手机、全面、服务、信息，并且从语料训练模型C中获取各个分词的组合概率，如对于乐视的分词来说，分词组合为乐视手机的组合概率，进而从语料检测模型C中获取待检测语句中各个分词的组合概率，并通过计算困惑度的公式得到待检测语句的语句困惑值，进一步查找到分词组合的语句为xx网提供乐视手机全面服务信息的语句困惑度是否大于与预设阈值，若大于或等于预设阈值，则说明该广告语句对应的SEM广告不符合展示条件，无法进行展示，并输出提示广告预警，若小于预设阈值，则说明该广告语句对应的SEM广告符合展示条件，将该SEM广告进行展示。

[0062] 由于在SEM广告投放至搜索引擎之前需要对广告语句进行通顺度检测，而现有技术通常采用人工操作对每条待投放SEM广告进行通顺度检测，从而保证SEM广告能够顺利展示，然而由于工作人员的先验知识不同，对于通顺度的理解标准也不一致，使得检测结果不够准确，导致部分本能够展示的SEM广告无法顺利展示。

[0063] 本发明实施例提供的另一种语句检测方法，通过从大数据积累的语料中总结经验，建立语料检测模型，批量快速的判断每条SEM广告的通顺度，，从而提高广告语句的检测效率，通过设定预设阈值，将语句困惑值大于预设阈值的广告语句对应的SEM广告进行预警，能够更高效、更客观的评估每条SEM广告，避免了由于误判使得广告无法上线导致广告商的经济损失。

[0064] 进一步地，作为图1所示方法的具体实现，本发明实施例提供一种语句检测装置，该装置实施例与前述方法实施例对应，为便于阅读，本装置不在对前述方法实施例中的细节内容进行逐一赘述，但应当明确，本实施例中的装置能够对应实现前述方法实施例中的全部内容，如图3所示，所述装置包括：

[0065] 第一获取单元31，用于获取待检测语句；

[0066] 检测单元32，用于根据语料检测模型对所述待检测语句中每个分词进行通顺度检测，得到语句困惑值，所述语料检测模型记录有不同领域自然文本中的分词、各个分词对应的频率以及各个分词在不同语句中的组合概率；

[0067] 输出单元33，用于将所述语句困惑值大于或等于预设阈值的语句进行标注，并且输出带有标注的提示信息。

[0068] 本发明实施例提供的一种语句检测装置，首先获取待检测语句，这里的待检测语句为SEM广告中的广告语句，然后根据语料检测模型对待检测语句中每个分词进行通顺度检测，得到语句困惑值，进而对广告语句的通顺度进行评估，将语句困惑值大于或等于预设阈值的广告语句对应的SEM广告进行标注，并且输出带有标注的提示信息，提示该广告语句对应的SEM广告不符合展示条件。相对于现有的通过人工操作对每个待展示SEM广告的广告语句进行检测的方法相比，本发明通过从大数据积累的语料库建立的语料检测模型对所述

待检测语句进行通顺度检测，无需人工操作，能够批量快速检测语句的通顺度，提高了广告语句检测结果的准确性，进一步通过将不符合展示条件的SEM广告进行预警，避免了由于广告上线不通顺给广告主带来的经济损失。

[0069] 进一步地，作为图2所示方法的具体实现，本发明实施例提供了另一种语句检测装置，该装置实施例与前述方法实施例对应，为便于阅读，本装置不在对前述方法实施例中的细节内容进行逐一赘述，但应当明确，本实施例中的装置能够对应实现前述方法实施例中的全部内容，如图4所示，所述装置包括：

[0070] 第一获取单元41，用于获取待检测语句；

[0071] 第二获取单元42，用于获取不同领域的自然文本作为样本数据；

[0072] 分词单元43，用于对所述样本数据中的每条语句进行分词，得到样本分词集合；

[0073] 训练单元44，用于采用语料训练模型对所述样本分词集合进行训练，得到语料检测模型；

[0074] 检测单元45，用于根据语料检测模型对所述待检测语句中每个分词进行通顺度检测，得到语句困惑值，所述语料检测模型记录有不同领域自然文本中的分词、各个分词对应的频率以及各个分词在不同语句中的组合概率；

[0075] 输出单元46，用于将所述语句困惑值大于或等于预设阈值的语句进行标注，并且输出带有标注的提示信息；

[0076] 展示单元47，用于将所述语句困惑值小于预设阈值的语句对应的广告进行展示。

[0077] 进一步地，检测单元45包括：

[0078] 划分模块451，用于将所述待检测语句进行划分，得到所述待检测语句的各个分词；

[0079] 获取模块452，用于根据所述语料检测模型获取所述待检测语句中各个分词在所述待检测语句中的组合概率；

[0080] 估计模块453，用于根据所述组合概率对所述待检测语句进行评估，得到所述待检测语句的语句困惑值。

[0081] 进一步地，估计模块453，具体用于根据所述组合概率对所述待检测语句进行评估，得到所述待检测语句的语句困惑值。

[0082] 本发明实施例提供的另一种语句检测装置，通过从大数据积累的语料中总结经验，建立语料检测模型，批量快速的判断每条SEM广告的通顺度，从而提高广告语句的检测效率，通过设定预设阈值，将语句困惑值大于预设阈值的广告语句对应的SEM广告进行预警，能够更高效、更客观的评估每条SEM广告，避免了由于误判使得广告无法上线导致广告商的经济损失。

[0083] 所述语句检测装置包括处理器和存储器，上述第一获取单元31、检测单元32和输出单元33等均作为程序单元存储在存储器中，由处理器执行存储在存储器中的上述程序单元来实现相应的功能。

[0084] 处理器中包含内核，由内核去存储器中调取相应的程序单元。内核可以设置一个或以上，通过调整内核参数来节省人力，能够提高广告语句检测结果的准确性。

[0085] 存储器可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM)和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)，存储器包括至少一个存

储芯片。

[0086] 本申请还提供了一种计算机程序产品，当在数据处理设备上执行时，适于执行初始化有如下方法步骤的程序代码：获取待检测语句，根据语料检测模型对所述待检测语句中每个分词进行通顺度检测，得到语句困惑值，所述语料检测模型记录有不同领域自然文本中的分词、各个分词对应的频率以及各个分词在不同语句中的组合概率，将所述语句困惑值大于或等于预设阈值的语句进行标注，并且输出带有标注的提示信息。

[0087] 本领域内的技术人员应明白，本申请的实施例可提供为方法、系统、或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0088] 本申请是参照根据本申请实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0089] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0090] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0091] 在一个典型的配置中，计算设备包括一个或多个处理器(CPU)、输入/输出接口、网络接口和内存。

[0092] 存储器可能包括计算机可读介质中的非永久性存储器，随机存取存储器(RAM)和/或非易失性内存等形式，如只读存储器(ROM)或闪存(flash RAM)。存储器是计算机可读介质的示例。

[0093] 计算机可读介质包括永久性和非永久性、可移动和非可移动媒体可以由任何方法或技术来实现信息存储。信息可以是计算机可读指令、数据结构、程序的模块或其他数据。计算机的存储介质的例子包括，但不限于相变内存(PRAM)、静态随机存取存储器(SRAM)、动态随机存取存储器(DRAM)、其他类型的随机存取存储器(RAM)、只读存储器(ROM)、电可擦除可编程只读存储器(EEPROM)、快闪记忆体或其他内存技术、只读光盘只读存储器(CD-ROM)、数字多功能光盘(DVD)或其他光学存储、磁盒式磁带，磁带磁磁盘存储或其他磁性存储设备或任何其他非传输介质，可用于存储可以被计算设备访问的信息。按照本文中的界定，计算机可读介质不包括暂存电脑可读媒体(transitory media)，如调制的数据信号和载波。

[0094] 以上仅为本申请的实施例而已，并不用于限制本申请。对于本领域技术人员来说，

本申请可以有各种更改和变化。凡在本申请的精神和原理之内所作的任何修改、等同替换、改进等,均应包含在本申请的权利要求范围之内。

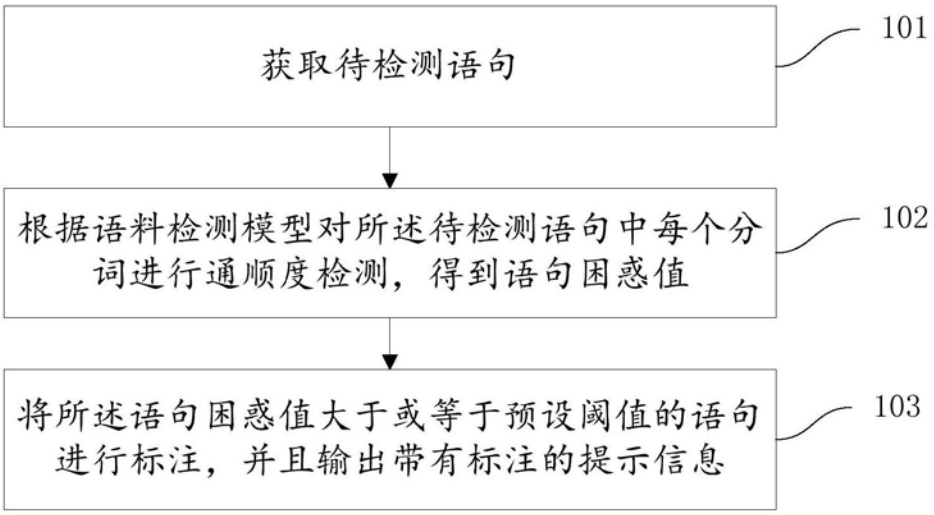

获取待检测语句
101
根据语料检测模型对所述待检测语句中每个分词进行通顺度检测，得到语句困惑值
102
将所述语句困惑值大于或等于预设阈值的语句进行标注，并且输出带有标注的提示信息
103

图1

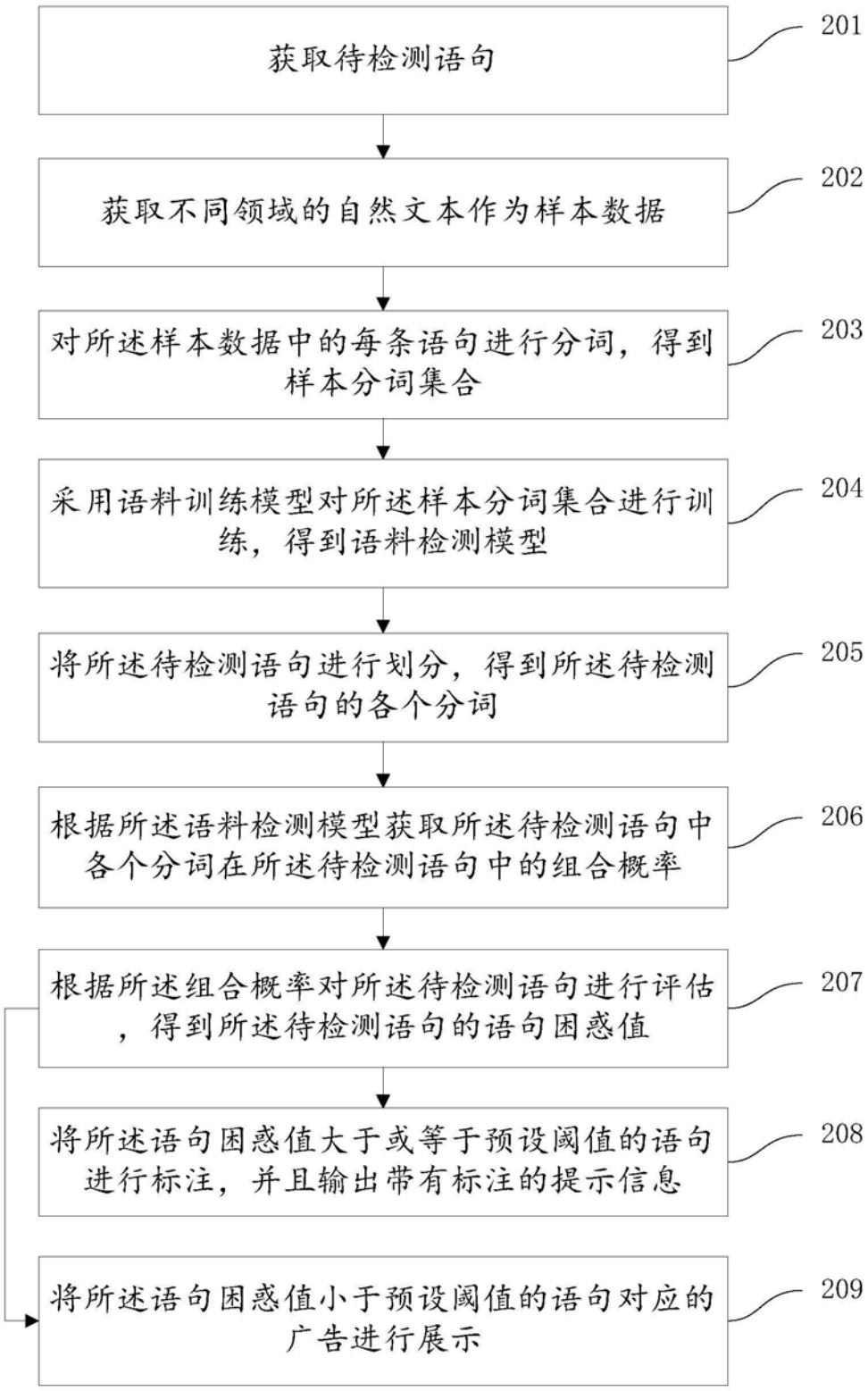

获取待检测语句
201
获取不同领域的自然文本作为样本数据
202
对所述样本数据中的每条语句进行分词，得到样本分词集合
203
采用语料训练模型对所述样本分词集合进行训练，得到语料检测模型
204
将所述待检测语句进行划分，得到所述待检测语句的各个分词
205
根据所述语料检测模型获取所述待检测语句中各个分词在所述待检测语句中的组合概率
206
根据所述组合概率对所述待检测语句进行评估，得到所述待检测语句的语句困惑值
207
将所述语句困惑值大于或等于预设阈值的语句进行标注，并且输出带有标注的提示信息
208
将所述语句困惑值小于预设阈值的语句对应的广告进行展示
209

图2

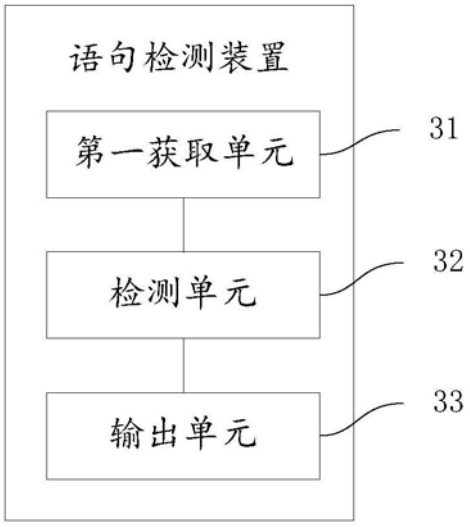


图3

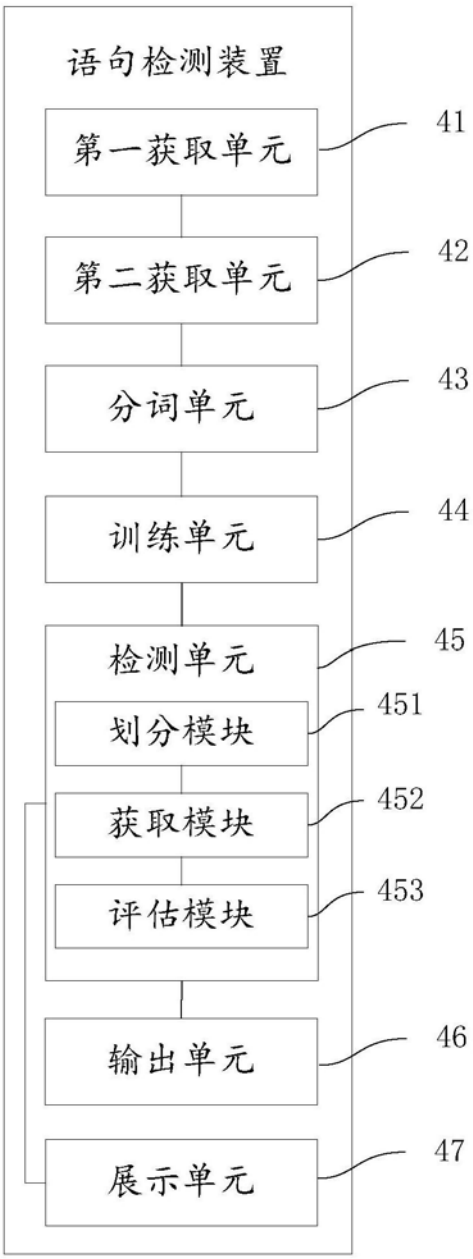

语句检测装置
第一获取单元
41
第二获取单元
42
分词单元
43
训练单元
44
检测单元
45
划分模块
451
获取模块
452
评估模块
453
输出单元
46
展示单元
47

图4