



(12) 发明专利申请

(10) 申请公布号 CN 102207947 A

(43) 申请公布日 2011. 10. 05

(21) 申请号 201010212267. 9

(22) 申请日 2010. 06. 29

(71) 申请人 天津海量信息技术有限公司

地址 300384 天津市华苑产业区榕苑路 1 号
B 北 322-323 室

(72) 发明人 宋传宝 张旭成

(74) 专利代理机构 北京汲智翼成知识产权代理
事务所（普通合伙） 11381

代理人 陈曦

(51) Int. Cl.

G06F 17/30(2006. 01)

G06F 17/27(2006. 01)

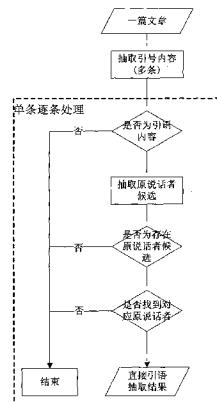
权利要求书 2 页 说明书 5 页 附图 2 页

(54) 发明名称

一种直接引语素材库的生成方法

(57) 摘要

本发明公开了一种直接引语素材库的生成方法，包括如下的步骤：(1) 对于一篇文章，首先抽取引号内的直接引语内容；(2) 以直接引语内容所在位置为中心，考察前一句和后一句的内容，通过词性标注识别出前一句和后一句中的人名和 / 或机构名；(3) 对于识别出来的存在于前一句或者后一句内容中的所有人名和 / 或机构名，作为直接引语陈述者的候选集合，从该候选集合中挑选出真正的直接引语陈述者；(4) 将挑选出来的直接引语陈述者和直接引语内容添加到直接引语素材库中。利用本方法生成的直接引语素材库，可以在互联网中提供更新、搜索、查询等服务，也可以在媒体资讯领域为写作、编辑、专题制作等提供技术支持。



1. 一种直接引语素材库的生成方法，其特征在于包括如下的步骤：

(1) 对于一篇文章，首先从左至右扫描，当扫描到的字符为左引号时，从下一个字符开始记录，一直记录到当前字符为右引号为止，在这个过程中记录的内容为直接引语内容；

(2) 以直接引语内容所在位置为中心，考察前一句和后一句的内容，通过词性标注识别出前一句和后一句中的人名和 / 或机构名；

(3) 对于识别出来的存在于前一句或者后一句内容中的所有人名和 / 或机构名，作为直接引语陈述者的候选集合，从该候选集合中挑选出真正的直接引语陈述者；

(4) 将挑选出来的直接引语陈述者和直接引语内容添加到直接引语素材库中。

2. 如权利要求 1 所述的直接引语素材库的生成方法，其特征在于：

所述步骤 (1) 中，所述左引号为全角左引号、半角左引号、单引号、双引号中的任意一种。

3. 如权利要求 1 所述的直接引语素材库的生成方法，其特征在于：

所述步骤 (1) 中，所述右引号为全角右引号、半角右引号、单引号、双引号中的任意一种。

4. 如权利要求 1 所述的直接引语素材库的生成方法，其特征在于：

所述步骤 (2) 中，以人工收集的机构名后缀词和常用构成词作为识别触发条件，根据隐马尔科夫模型或者最大熵模型进行识别机构名的计算。

5. 如权利要求 1 所述的直接引语素材库的生成方法，其特征在于：

所述步骤 (3) 按照如下情况分别进行处理：

(31) 如果候选集合中不存在人名和 / 或机构名，则丢弃掉该直接引语内容，不进行处理；

(32) 如果候选集合中只存在一个人名或者机构名，则将识别出来的人名或者机构名作为直接引语陈述者；

(33) 如果候选集合中存在多个人名或者机构名，则对候选集合中的人名或者机构名进行选择，选出一个作为直接引语陈述者。

6. 如权利要求 5 所述的直接引语素材库的生成方法，其特征在于：

所述步骤 (33) 中，根据以下因素对于候选人名或者机构名进行打分：1. 字符距离，考察候选人名或者机构名与直接引语的字符距离，通过计算文本中两者间隔的字符数来获得字符距离；2. 语义距离，考察句子的词性架构，使用浅层语义分析，分析出候选人名或者候选机构名与直接引语的结构距离；将字符距离和语义距离相加，找出距离最小的候选人名或者候选机构名作为直接引语陈述者。

7. 如权利要求 6 所述的直接引语素材库的生成方法，其特征在于：

在进行所述浅层语义分析时，首先统计出主语为人名或者机构名与引语谓词近距离搭配的文字片段，进而统计表达模式的数量以获得所有常见的表达模式，最终根据模式统计值、谓词统计值，综合计算信度值。

8. 如权利要求 1 所述的直接引语素材库的生成方法，其特征在于：

所述步骤 (4) 中，以数据库方式保存直接引语陈述者和直接引语内容，即在数据库中设立两个字段，一个字段保存直接引语陈述者，另一个字段保存直接引语内容。

9. 如权利要求 1 所述的直接引语素材库的生成方法，其特征在于：

所述步骤(4)中,以文本方式保存直接引语陈述者和直接引语内容,即将直接引语按照“直接引语陈述者直接引语内容”的方式保存为文本文件,其中直接引语陈述者与直接引语内容之间用间隔符号分开。

一种直接引语素材库的生成方法

技术领域

[0001] 本发明涉及一种语言素材库的生成方法,尤其涉及一种针对直接引语的句子级素材库生成方法,属于计算语言学技术领域。

背景技术

[0002] 素材库也称语料库 (corpus),是存储于计算机中并可利用计算机进行检索、查询、分析的语言素材的总体。素材库具有“大规模”和“真实性”这两个特点,因此是最理想的语言知识资源。

[0003] 文本是最基本、最常用的信息载体。在计算机语言处理工作中,文本的加工与处理技术显得尤为重要。句子作为能够表达完整的意思最小语言单位,在信息处理与应用中,具有多种表现形式和使用价值,尤其是在媒体资讯的检索、写作、整理等过程中更是如此。但在当前存在的各种语言处理技术中,以句子作为处理颗粒的尚不多见。

[0004] 在 2003 年举行的全国第七届计算语言学联合学术会议中,苗传江、刘智颖合作发表了论文《现代汉语语料的句子级语义标注》。在该论文中,讨论了一种标注现代汉语语料的方案。它有两个特点:一是采取自下而上的标注方式,即先标注大的语言单位,再标注小的语言单位;二是对句子进行语义标注,标注了句子及句内子句的语义类型和它们的下一级语义构成成分。按此方案建立的语料库是现代汉语句子语义研究和处理的重要资源。

[0005] 另外,在申请号为 200810065527.7 的中国发明专利申请中,公开了一种用电子装置对文章句子进行快速分类及检索的方法。在该技术方案中,电子装置按特定的分类方法生成文章句子的分类目录表。在检索时:用户打开的电子书内容,处理器逐个提取每个句子,并查找到该句子所在的分类目录,将分类目录名称作为分类标记附注添加到该句子上,带分类标记的句子被用户选中后,句子读取指针定位到分类目录,对其中的句子进行输出。电子装置可对文章句子进行分类贮存,其步骤如下:1) 显示屏上显示由若干条句子组成的文章内容;2) 通过编辑器为其中的任意句子做上特定分类标记;3) 贮存器中建立与上述每个分类标记对应的类别目录,如果目录已经存在,则不建立;4) 处理器对有分类标记的句子进行检测及识别,自动将每个有分类标记的句子保存到对应的上述类别目录中。但是,该专利申请中对句子的挖掘与整理在很大程度上依赖于人工,工作效率并不高,根本无法满足海量中文文本数据的处理要求。

发明内容

[0006] 本发明所要解决的技术问题在于提供一种针对直接引语的句子级素材库生成方法。该方法以句子为颗粒抽取文本中的直接引语信息,从而将原始的文章数据库转换为所需的直接引语素材库。

[0007] 为实现上述的发明目的,本发明采用下述的技术方案:

[0008] 一种直接引语素材库的生成方法,其特征在于包括如下的步骤:

[0009] (1) 对于一篇文章,首先从左至右扫描,当扫描到的字符为左引号时,从下一个字

符开始记录,一直记录到当前字符为右引号为止,在这个过程中记录的内容为直接引语内容;

[0010] (2) 以直接引语内容所在位置为中心,考察前一句和后一句的内容,通过词性标注识别出前一句和后一句中的人名和 / 或机构名;

[0011] (3) 对于识别出来的存在于前一句或者后一句内容中的所有人名和 / 或机构名,作为直接引语陈述者的候选集合,从该候选集合中挑选出真正的直接引语陈述者;

[0012] (4) 将挑选出来的直接引语陈述者和直接引语内容添加到直接引语素材库中。

[0013] 其中,在所述步骤(1)中,所述左引号为全角左引号、半角左引号、单引号、双引号中的任意一种。所述右引号为全角右引号、半角右引号、单引号、双引号中的任意一种。

[0014] 所述步骤(2)中,以人工收集的机构名后缀词和常用构成词作为识别触发条件,根据隐马尔科夫模型或者最大熵模型进行识别机构名的计算。

[0015] 所述步骤(3)按照如下情况分别进行处理:

[0016] (31) 如果候选集合中不存在人名和 / 或机构名,则丢弃掉该直接引语内容,不进行处理;

[0017] (32) 如果候选集合中只存在一个人名或者机构名,则将识别出来的人名或者机构名作为直接引语陈述者;

[0018] (33) 如果候选集合中存在多个人名或者机构名,则对候选集合中的人名或者机构名进行选择,选出一个作为直接引语陈述者。

[0019] 所述步骤(33)中,根据以下因素对于候选人名或者机构名进行打分:1. 字符距离,考察候选人名或者机构名与直接引语的字符距离,通过计算文本中两者间隔的字符数来获得字符距离;2. 语义距离,考察句子的词性架构,使用浅层语义分析,分析出候选人名或者候选机构名与直接引语的结构距离;将字符距离和语义距离相加,找出距离最小的候选人名或者候选机构名作为直接引语陈述者。

[0020] 在进行所述浅层语义分析时,首先统计出主语为人名或者机构名与引语谓词近距离搭配的文字片段,进而统计表达模式的数量以获得所有常见的表达模式,最终根据模式统计值、谓词统计值,综合计算信度值。

[0021] 所述步骤(4)中,以数据库方式保存直接引语陈述者和直接引语内容,即在数据库中设立两个字段,一个字段保存直接引语陈述者,另一个字段保存直接引语内容。或者,以文本方式保存直接引语陈述者和直接引语内容,即将直接引语按照“直接引语陈述者直接引语内容”的方式保存为文本文件,其中直接引语陈述者与直接引语内容之间用间隔符号分开。

[0022] 本发明所提供的直接引语素材库生成方法具有如下的优点:

[0023] 1. 可以实现直接引语的准确识别,实现了对于有引号但非引语表达内容的准确判别;

[0024] 2. 可以实现对原说话者的准确识别,根据直接引语位置,从其附近位置准确识别原说话者候选集;

[0025] 3. 可以实现直接引语与原说话者的准确对应,两者的准确映射结果就形成了完整的直接引语结果数据。

[0026] 利用本方法生成的直接引语素材库,可以在互联网中提供更新、搜索、查询等服

务,也可以在媒体资讯领域为写作、编辑、专题制作等提供技术支持。

附图说明

[0027] 下面结合附图和具体实施方式对本发明作进一步的详细说明。

[0028] 图 1 为从文章库转换为句子级素材库的基本操作流程示意图;

[0029] 图 2 为直接引语素材库的生成过程示意图。

具体实施方式

[0030] 图 1 为句子级素材库生成方法中,从文章库转换为句子级素材库的基本流程示意图。从图 1 可以看出,对于文章库中的每一篇中文文章,可以通过句子级素材抽取操作获得各种类型的句子素材,例如“直接引语”句子、“大事记”句子等。这些“直接引语”句子、“大事记”句子等可以分别放入相应的直接引语素材库或者大事记素材库中进行保存。需要说明的是,对于文本中的诸多句子,并非每一个句子都可以形成有价值、有意义的素材。只有那些确定类型,并进行结构化处理后的句子类型,才可以形成对应的句子级素材。根据网络编辑工作的实际需要,句子级素材库中的一个子集 - 直接引语素材库十分有用。下面对它的生成过程进行详细的说明。

[0031] 直接引语是指作者在文本中直接引用别人的话,即采用直接引述别人原话内容的方式并且把原话内容放入引号中,这些原话内容即为直接引语。直接引语作为一种常见的写作表达方式,在文章中非常普遍,尤其是在媒体资讯的报道性、评论性文章中,更是大量存在。与直接引语相对应的是间接引语,间接引语是不用引号而引述别人讲话内容的一种表述方式。

[0032] 在文本中,直接引语是大量存在的。如何通过计算机技术自动识别并结构化抽取直接引语是我们要着力解决的问题。抽取的直接引语,其结构化结果至少包含两部分:直接引语和原说话者。直接引语处理技术通过识别引语内容,识别原说话者,进而建立两者之间的对应关系,实现信息的抽取与结构化。

[0033] 作为观点的引述,一般直接引语的原始说话者会被清晰明确地描述出来,因此,“原说话者 + 说话内容”就成为可以独立出来、且其语义内容相对完整的数据记录。将多个这样的数据记录,依据字段将其内容存入数据库,则形成了直接引语的句子级素材库,可供后续查询、检索、统计等使用。

[0034] 对于文本中直接引语的抽取过程,主要通过如图 2 所示的步骤予以实现:

[0035] 1. 对于一篇文章,首先从左至右扫描,当扫描到的字符为左引号(包括全角左引号、半角左引号、单引号、双引号)时,从下一个字符开始记录,一直记录到当前字符为右引号(包括全角右引号、半角右引号、单引号、双引号)为止。在这个过程中记录的内容为发现的直接引语内容。

[0036] 2. 以扫描到的直接引语内容所在位置为中心,考察前一句和后一句的内容。借助于现有分词软件中的词性标注功能,可以识别出前一句和后一句中的人名和 / 或机构名。此处用到的分词软件可以是 SCWS、ICTCLAS、HTTPCWS 等中文开源分词软件。这些分词软件大部分都支持词性标注,可以直接识别出人名。对于机构名的识别可以借助于基于隐马尔科夫模型、最大熵模型等统计模型实现。具体而言,识别机构名的主要方法为根据人工收集

的机构名后缀词（如：“公司”“公安局”等）和常用构成词（如：“有限责任”“国际”，“物流”等）作为识别触发条件，然后根据隐马尔科夫模型或者最大熵模型进行识别计算。此处使用的隐马尔科夫模型、最大熵模型等都是常用的自然语言处理统计模型，关于这些模型的更多信息可以参考宗成庆编写的《统计自然语言处理》一书（清华大学出版社 2008 年 5 月版，ISBN :978-7-302-16598-9），在此就不详细赘述了。

[0037] 3. 对于识别出来的存在于前一句或者后一句内容中的所有人名和 / 或机构名，作为该引语陈述者的候选集合，从这个集合中挑选出真正的直接引语陈述者。

[0038] 具体的挑选方法阐述如下：

[0039] 3.1 如果候选集合中不存在人名和 / 或机构名，即在该句子前后句中没有人名和 / 或机构名，说明上下文中并没有明确的陈述者出现，则丢弃掉该引语内容，不进行处理。

[0040] 3.2 如果候选集合中只存在一个人名或者机构名，则不需要进行挑选，识别出来的人名或者机构名即为句子的陈述者。此时，将直接引语陈述者和引语内容添加到直接引语素材库中。

[0041] 3.3 如果候选集合中存在多个人名或者机构名，则使用浅层语义分析方法对候选集合中的人名或者机构名进行选择，选出一个作为引语的陈述者。

[0042] 具体而言，根据以下因素对于候选人名或机构名进行打分：1. 字符距离，考察候选人名或者机构名与引语的字符距离，通过计算文本中两者间隔的字符数来获得字符距离。2. 语义距离，考察句子的词性架构，使用下述的浅层语义分析，分析出候选人名或者候选机构名与引语的结构距离。将两个距离相加，找出距离最小的候选人名或者候选机构名作为该引语的陈述者，将挑选出来的人名或者机构名和引语内容添加到直接引语素材库中。

[0043] 上面提到的浅层语义分析方法是这样的：首先对直接引语内容前后的句子进行句法分析。该句法分析可以使用现有的成熟技术实现，例如哈工大信息检索中心开发的 LTP 平台系统，就提供了对句子进行句法分析的功能。通过句法分析，可以标出句子中的主语、谓语、宾语、修饰语及其对应原句子中的词。然后判断主语所对应的词，如果是人名或者机构名，则判断谓语是否为引语功能的谓语词汇，如“说”“表示”“称”等等。如果满足上面两个条件，就可以简单认为主语对应的人名或者机构名为直接引语的陈述者，将挑选出来的人名或者机构名和引语内容添加到直接引语素材库中。

[0044] 当然，主语谓语的搭配模式有很多种。本发明人根据机器统计的方法，统计出主语为人名或机构名与引语谓词（如：“说”“表示”等）近距离搭配的文字片段，这些文字片段就是引语的表达模式（如：“人名 + 强调说”，“人名 + 发表谈话说”等形式），进而统计表达模式的数量，即可获得所有常见的表达模式；最终，根据模式统计值（即统计出的其使用频度）、谓词统计值（即统计出的其使用频度）等因素，综合计算信度值（如：将统计值归一化为 0~1 的浮点数，加权后相加累计获得信度值），当有多个候选主语时，优选出信度最高的作为原说话者的对应结果。

[0045] 对于上面直接引语抽取方法所得到的直接引语素材库，可以按照两种方式保存：

1. 通过数据库保存。在数据库中设立两个字段，一个字段保存直接引语陈述者，另一个字段保存直接引语内容。2. 通过文本方式保存，即直接将识别出来的直接引语按照“直接引语陈述者直接引语内容”保存为文本文件。其中直接引语陈述者与直接引语内容之间用间隔

符号分开，间隔符号可以为空格、Tab 键或是用户自己定义的任意一个符号。

[0046] 另外，对于上述获得的直接引语素材库，相关的检索工作分为按直接引语陈述者检索和按直接引语内容进行检索两种。

[0047] 在检索之前，需要对直接引语素材库建立索引。对于存储在数据库中的，直接对于两个字段内容进行索引；对于存储在文本中的，可以借助于文本索引软件如开源软件 Lucene 等进行索引。同样地，索引时也是分别按引语陈述者和直接引语内容分别索引。

[0048] 在建立了索引之后，对于按直接引语陈述者检索的检索需求，可以在索引中的直接引语陈述者字段内容中检索，返回匹配的直接引语陈述者和直接引语内容。对于按直接引语内容检索的检索需求，可以在索引中的直接引语内容字段内容中检索，返回匹配的直接引语陈述者和直接引语内容。

[0049] 为了实现直接引语素材库的可运营化，即动态地向语料库中添加新内容，删除过时和不正确的内容，本专利申请进一步提出更新直接引语素材库的方法，具体说明如下：

[0050] 添加操作：对于待添加内容，可以按照两种方法添加到直接引语素材库中。1. 对于待添加内容，在索引中查找是否已经存在相同的直接引语条目，如果不存在，将内容添加进去，同时更新索引，将新加的内容添加到索引中。2. 直接将待添加内容加入直接引语素材库中，然后进行消重操作，重新生成索引。

[0051] 删除操作：对于待删除的内容，在索引中查找到相应的直接引语条目，然后从索引中删除。

[0052] 修改操作：对于修改的内容，在索引中找到相应的直接引语条目，删除该条目并将修改的内容加入索引。在这些基础之上，进行修改处理。

[0053] 本发明所生成的直接引语素材库可以在互联网检索和传媒领域得到广泛的使用。其中对于互联网而言，互联网中存在大量的文本信息，尤其是媒体资讯信息，且每天都在不断地增长；针对互联网上的文本进行直接引语抽取后，我们就能获得一个庞大的直接引语素材库，这一直接引语素材库，可以按说话者或按说话内容进行检索，其可能的用户描述如下：

[0054] 1) 对于普通网民而言，非常方便他们了解自己关心的名人所说过的内容，加入文本时间维度，则还能按时间进行过滤筛选；同时还能搜索某个关键词，看哪些人发表过相关看法；还可以说话者和说话内容关键词同时为条件检索等。

[0055] 2) 对于写作者或媒体从业者，尤其是记者，可以很方便地组织写作素材，形成稿件；对于网站编辑进行专题制作，也可以针对专题中的人物、机构，直接展示列举其言论观点，或者针对专题主体内容，列举所有内容相关的直接引语和说话者等等。

[0056] 另外，在政府机关或传统媒体行业中，均存在大量的行业文本数据，也会存在文章中直接引语包含密集的情况。在这种情况下，通过对行业数据的再处理，可以将这些行业数据盘活，产生新的检索查阅和生产价值。

[0057] 以上对本发明所提供的直接引语素材库生成方法进行了详细的说明。对本领域的技术人员而言，在不背离本发明实质精神的前提下对它所做的任何显而易见的改动，都将构成对本发明专利权的侵犯，将承担相应的法律责任。

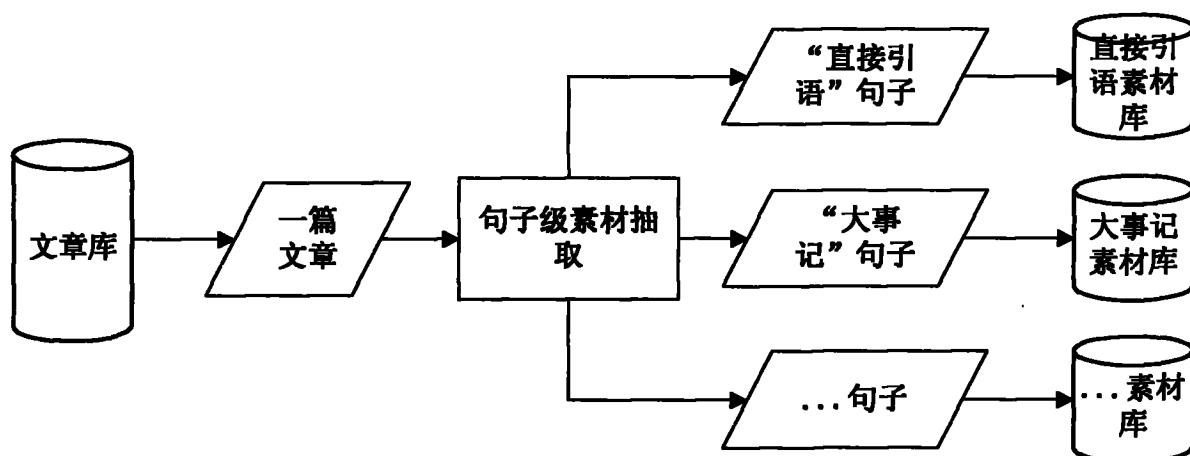


图 1

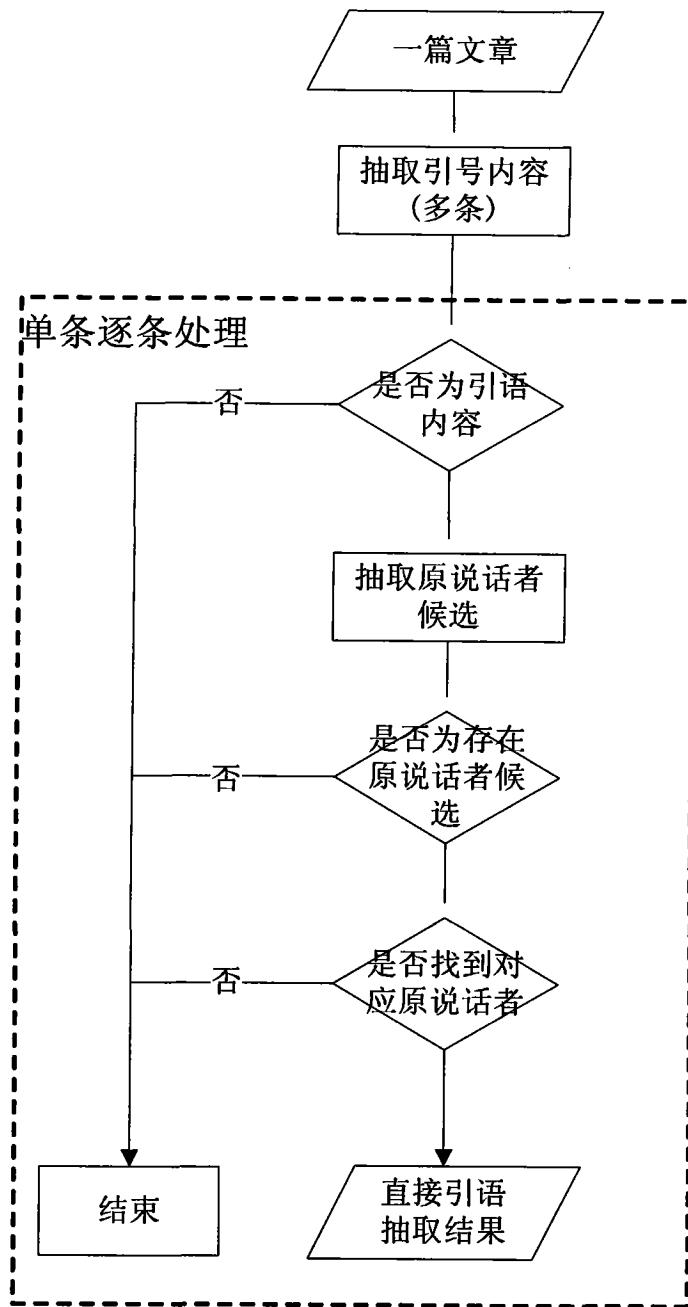


图 2