(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 106294315 A
(43)申请公布日 2017.01.04

(21)申请号 201610599558.5

(22)申请日 2016.07.27

(71)申请人 昆明理工大学
地址 650093 云南省昆明市五华区学府路253号

(72)发明人 邵玉斌 刘彩 王腾

(51)Int.Cl.
G06F 17/27(2006.01)

权利要求书2页 说明书8页 附图2页

(54)发明名称

基于句法特性与统计融合的自然语言谓语动词识别方法

(57)摘要

本发明涉及一种基于句法特性与统计融合的自然语言谓语动词识别方法，属于自然语言处理技术领域。本发明首先对输入的待测句子进行预处理分析，具体包括：判定语种、词性标注、对应的过滤处理和疑似动词抽取；其次，进行疑似谓语动词的排查，通过句法特性甄别出疑似动词中的谓语动词；接着判断该动词是否以动词搭配组的情况出现，这里利用ϕ^2统计法来判断动词搭配组的真伪；最后根据识别结果输出所测句子的谓语动词或是谓语动词搭配组信息。本发明通过词性标注，过滤处理和疑似动词抽取来提高识别谓语动词的高效性，通过句法特性分析和ϕ^2统计法提高识别谓语动词和谓语动词搭配的精确性。本发明的可行性高并适用于一般自然语言的谓语动词识别。

CN 106294315 A

1.基于句法特性与统计融合的自然语言谓语动词识别方法，其特征在于：首先对输入的待测句子进行预处理分析，具体包括：判定语种、词性标注、对应的过滤处理和疑似动词抽取；其次，进行疑似谓语动词的排查，通过句法特性甄别出疑似动词中的谓语动词；接着判断该动词是否以动词搭配组的情况出现，这里利用Φ^2统计法来判断动词搭配组的真伪；最后根据识别结果输出所测句子的谓语动词或是谓语动词搭配组信息。

2.根据权利要求1所述的基于句法特性与统计融合的自然语言谓语动词识别方法，其特征在于：所述基于句法特性与统计融合的自然语言谓语动词识别方法的具体步骤如下：

Step1、对待测句子进行预处理分析：输入句子，通过文本语种识别工具判定语种，使用词性标注工具对句子中的词逐个进行词性标注，然后对分析谓语动词不相关的词性，如语气词等进行过滤处理，接下来，根据词性标注结果抽取出疑似动词，若无疑似动词，则直接输出句中无谓语动词的提示信息；若有疑似动词，则进行如下步骤Step2；

Step2、疑似谓语动词的排查：通过疑似谓语动词的形态分析和句法规则库得到疑似谓语动词；

Step3、动词搭配组识别：将疑似谓语动词的词找到后，分析该谓语动词是否是以动词搭配组的形式出现，如果不是，则把该疑似谓语动词作为待测句子的谓语动词输出，如果是，则进行动词搭配组的识别，其中，利用Φ^2统计法来判别该动词搭配组的真伪；

Step4、根据上述步骤，输出所识别出待测句子的谓语动词或是谓语动词搭配组信息。

3.根据权利要求1所述的基于句法特性与统计融合的自然语言谓语动词识别方法，其特征在于：所述步骤Step1中，对待测句子进行词性标注、对应的过滤处理和疑似动词抽取，其操作步骤如下：

Step1.1、对输入的待测句子通过文本语种识别工具判定语种，通过分词工具进行分词并对切分出来的单词标注词性；

Step1.2、根据标注的词性判断，若无疑似动词，则不进行下面的一系列分析，直接输出句中无谓语动词的提示信息；若存在疑似动词，则进行步骤Step1.3；

Step1.3、存在疑似动词，则对分析谓语动词不相关的词性，如语气词，部分副词等进行过滤处理，用于减轻句法分析负担，提高识别效率。

4.根据权利要求2所述的基于句法特性与统计融合的自然语言谓语动词识别方法，其特征在于：所述步骤Step2中所述的疑似谓语动词排查，其具体步骤如下：

Step2.1、若疑似谓语动词个数为1，则结合形态分析和句法规则库，对该疑似谓语动词是否在该句中作为谓语成分出现进行甄别；若判断出不是谓语动词，则流程不进行下面的步骤，直接输出句中无谓语动词的提示信息；若判断出是谓语动词，则转入进行动词搭配组识别；

Step2.2、若疑似谓语动词个数超过1个，则逐个对这些词进行形态分析，若可以判定，则转入进行动词搭配组识别；若不能判定，则利用句法规则库进行判定，若判断出不是谓语动词，则流程不进行下面的步骤，直接输出句中无谓语动词的提示信息，若判断出是谓语动词，则转入进行动词搭配组识别。

5.根据权利要求1所述的基于句法特性与统计融合的自然语言谓语动词识别方法，其特征在于：所述步骤Step3中所述的动词搭配组识别，判别该动词是否以动词搭配组的形式在该句子中展现，具体是：

Step3.1、若初步判定是二元动词搭配组，则再通过词语窗口，窗口长度视不同语种而定，判断该二元动词搭配组后面是否有与之搭配的词出现，若有，则通过Φ^2统计法来判别该动词搭配的真伪；若无，则将该二元动词搭配组中的动词作为待测句子最终的谓语动词进行输出；

Step3.2、若初步判定是三元动词搭配组或是更多元的动词搭配组，再通过将其匹配多元动词搭配库的形式进行真伪判别。

基于句法特性与统计融合的自然语言谓语动词识别方法

技术领域

[0001] 本发明涉及一种基于句法特性与统计融合的自然语言谓语动词识别方法，属于自然语言处理技术领域。

背景技术

[0002] 谓语动词的识别在机器翻译、句法分析、信息抽取中扮演着很重要的角色。在句子分析中，主语、谓语、宾语是句子的主干，而谓语是主语和宾语产生联系的关键。故谓语动词可谓是句子的核心所在。例如，依存句法分析中，谓语动词便是放在树根的位置。同时，一个句子的语义主要是由其核心谓语动词所决定的。因此，高效准确的谓语动词识别方法更加凸显它的重要性。

[0003] 语言学家乔姆斯基证明了世界上实际只有一种人类语言。而且，多年前就有人提出世界上所有语言必属于三种类型(SVO、SOV以及VSO)之一，例如汉语、英语语种属于SVO类型，日语语种属于SOV类型，但句子成分都离不开主谓宾的主干成分。那么，世界上肯定有一种通用的识别模型，这种模型在针对一般自然语言上也一定会抓住其共性，把所需要的特征提取出来。

发明内容

[0004] 本发明提供了一种基于句法特性与统计融合的自然语言谓语动词识别方法，以用于提高一般自然语言中的谓语动词识别的高效性和精确度。该方法不仅通过词性标注，过滤处理和疑似动词识别来提高识别谓语动词的高效性，而且通过给定的句法特性和ϕ^2统计法提高识别谓语动词搭配组的精确性。

[0005] 本发明的技术方案是：一种基于句法特性与统计融合的自然语言谓语动词识别方法，首先对输入的待测句子进行预处理分析，具体包括：判定语种、词性标注、对应的过滤处理和疑似动词抽取；其次，进行疑似谓语动词的排查，通过句法特性甄别出疑似动词中的谓语动词；接着判断该动词是否以动词搭配组的情况出现，这里利用ϕ^2统计法来判断动词搭配组的真伪；最后根据识别结果输出所测句子的谓语动词或是谓语动词搭配组信息。

[0006] 所述基于句法特性与统计融合的自然语言谓语动词识别方法的具体步骤如下：

[0007] Step1、对待测句子进行预处理分析：输入句子，通过文本语种识别工具判定语种，使用词性标注工具对句子中的词逐个进行词性标注，然后对分析谓语动词不相关的词性，如语气词等进行过滤处理，接下来，根据词性标注结果抽取出疑似动词，若无疑似动词，则直接输出句中无谓语动词的提示信息；若有疑似动词，则进行如下步骤Step2；

[0008] 通过词性标注，把不同类别的词区别开来，方便后续的疑似动词判别和不相关词性(如语气词)的过滤。

[0009] Step2、疑似谓语动词的排查：通过疑似谓语动词的形态分析和句法规则库得到疑似谓语动词；这部分通过针对谓语动词的词法句法特性分析来达到甄别谓语动词的目的，并为下一步的谓语动词搭配组识别做铺垫。

[0010] Step3、动词搭配组识别:将疑似谓语动词的词找到后,分析该谓语动词是否是以动词搭配组的形式出现,如果不是,则把该疑似谓语动词作为待测句子的谓语动词输出,如果是,则进行动词搭配组的识别,其中,利用Φ^2统计法来判别该动词搭配组的真伪;通过Φ^2统计法来判别动词搭配组的真伪的方法,这样基于统计的方法结合计算机高效的计算能力,从而达到高效的识别出动词搭配组的真伪,避免了基于规则的识别方法带来的繁琐和规则与规则之间相互约束的局限。

[0011] Step4、根据上述步骤,输出所识别出待测句子的谓语动词或是谓语动词搭配组信息。

[0012] 所述步骤Step1中,对待测句子进行词性标注、对应的过滤处理和疑似动词抽取,其操作步骤如下:

[0013] Step1.1、对输入的待测句子通过文本语种识别工具判定语种,通过分词工具进行分词并对切分出来的单词标注词性;

[0014] Step1.2、根据标注的词性判断,若无疑似动词,则不进行下面的一系列分析,直接输出句中无谓语动词的提示信息;若存在疑似动词,则进行步骤Step1.3;

[0015] Step1.3、存在疑似动词,则对分析谓语动词不相关的词性,如语气词,部分副词等进行过滤处理,用于减轻句法分析负担,提高识别效率。

[0016] 所述步骤Step2中所述的疑似谓语动词排查,其具体步骤如下:

[0017] Step2.1、若疑似谓语动词个数为1,则结合形态分析和句法规则库,对该疑似谓语动词是否在该句中作为谓语成分出现进行甄别;若判断出不是谓语动词,则流程不进行下面的步骤,直接输出句中无谓语动词的提示信息;若判断出是谓语动词,则转入进行动词搭配组识别;

[0018] Step2.2、若疑似谓语动词个数超过1个,则逐个对这些词进行形态分析,若可以判定,则转入进行动词搭配组识别;若不能判定,则利用句法规则库进行判定,若判断出不是谓语动词,则流程不进行下面的步骤,直接输出句中无谓语动词的提示信息,若判断出是谓语动词,则转入进行动词搭配组识别。例如,英语中比较句中出现的助动词(如do、will、would等)和这些动词的正向距离大小判定最有可能是谓语动词角色的词,通过判定则转入Step3步骤中的动词搭配组识别。

[0019] 所述步骤Step3中所述的动词搭配组识别,判别该动词是否以动词搭配组的形式在该句子中展现,具体是:

[0020] Step3.1、若初步判定是二元动词搭配组,则再通过词语窗口,窗口长度视不同语种而定,判断该二元动词搭配组后面是否有与之搭配的词出现,若有,则通过Φ^2统计法来判别该动词搭配的真伪;若无,则将该二元动词搭配组中的动词作为待测句子最终的谓语动词进行输出;

[0021] Step3.2、若初步判定是三元动词搭配组或是更多元的动词搭配组,再通过将其匹配多元动词搭配库的形式进行真伪判别。

[0022] 详细的Φ^2统计法用于判定动词搭配组真伪的方法如下:

[0023] 表1对于两个词w_1和w_2,建立关联表如下:

[0024]

	w_2	$!w_2$	Σ

w_1	a	b	a+b
!w_1	c	d	c+d
Σ	a+c	b+d	a+b+c+d

[0025] 上表中，a表示词w_1、w_2出现的次数，b表示不在词w_1、w_2中的w_1的出现次数，c表示不在词w_1、w_2中的w_2的出现次数，d表示既不是w_1又不是w_2的词的次数，a+b是w_1出现的总词数，c+d是非w_1的总词数，a+c是w_2的出现词数，b+d是非w_2的总词数，N＝a+b+c+d表示语料库中的总词数。

[0026] 根据上面的联立表，ϕ^2统计量定义公式如下公式(1)：

[0027] $$\phi^2 = \frac{(a\times d - b\times c)^2}{(a+b)\times(a+c)\times(b+d)\times(c+d)} \qquad (1)$$

[0028] 当a＝0时，ϕ^2近于0，即当w_1和w_2从不共现时，ϕ^2取极小值。当b＝c＝0时，ϕ^2＝1，即当w_1和w_2总是共现时，ϕ^2取极大值。ϕ^2值越大，说明w_1和w_2共现的机会越多，相反，ϕ^2值越小，则说明w_1和w_2共现的机会越小。

[0029] 基于上述ϕ^2统计法思想，借助语料库来统计动词搭配组情况，通过比较设定的门限和统计量ϕ^2的值来判定该动词搭配组真伪。这里需要说明的是：a、a+b、a+c是提前统计好已存放入数据库的；而针对大于两个词组合的动词搭配组(如英语中的take care of)，则通过匹配多元动词搭配库(人工整理)，若匹配成功，则认定是真动词搭配组，否则，系统只输出动词信息作为谓语动词。

[0030] 本发明的有益效果是：

[0031] 本发明基于句法特性与统计相融合的方法，通过词性标注，过滤处理和疑似动词抽取来提高识别谓语动词的高效性，通过句法特性分析和ϕ^2统计法提高识别谓语动词和谓语动词搭配的精确性。本发明的可行性高并适用于一般自然语言的谓语动词识别。

附图说明

[0032] 图1为本发明的整体流程图；

[0033] 图2为本发明的详细流程图。

具体实施方式

[0034] 实施例1：如图1-2所示，一种基于句法特性与统计融合的自然语言谓语动词识别方法，首先对输入的待测句子进行预处理分析，具体包括：判定语种、词性标注、对应的过滤处理和疑似动词抽取；其次，进行疑似谓语动词的排查，通过句法特性甄别出疑似动词中的谓语动词；接着判断该动词是否以动词搭配组的情况出现，这里利用ϕ^2统计法来判断动词搭配组的真伪；最后根据识别结果输出所测句子的谓语动词或是谓语动词搭配组信息。

[0035] 所述基于句法特性与统计融合的自然语言谓语动词识别方法的具体步骤如下：

[0036] Step1、对待测句子进行预处理分析：输入句子，通过文本语种识别工具判定语种，使用词性标注工具对句子中的词逐个进行词性标注，然后对分析谓语动词不相关的词性，如语气词等进行过滤处理，接下来，根据词性标注结果抽取出疑似动词，若无疑似动词，则直接输出句中无谓语动词的提示信息；若有疑似动词，则进行如下步骤Step2；

[0037] Step2、疑似谓语动词的排查：通过疑似谓语动词的形态分析和句法规则库得到疑

似谓语动词；

[0038] Step3、动词搭配组识别：将疑似谓语动词的词找到后，分析该谓语动词是否是以动词搭配组的形式出现，如果不是，则把该疑似谓语动词作为待测句子的谓语动词输出，如果是，则进行动词搭配组的识别，其中，利用ϕ^2统计法来判别该动词搭配组的真伪；

[0039] Step4、根据上述步骤，输出所识别出待测句子的谓语动词或是谓语动词搭配组信息。

[0040] 所述步骤Step1中，对待测句子进行词性标注、对应的过滤处理和疑似动词抽取，其操作步骤如下：

[0041] Step1.1、对输入的待测句子通过文本语种识别工具判定语种，通过分词工具进行分词并对切分出来的单词标注词性；

[0042] Step1.2、根据标注的词性判断，若无疑似动词，则不进行下面的一系列分析，直接输出句中无谓语动词的提示信息；若存在疑似动词，则进行步骤Step1.3；

[0043] Step1.3、存在疑似动词，则对分析谓语动词不相关的词性，如语气词，部分副词等进行过滤处理，用于减轻句法分析负担，提高识别效率。

[0044] 所述步骤Step2中所述的疑似谓语动词排查，其具体步骤如下：

[0045] Step2.1、若疑似谓语动词个数为1，则结合形态分析和句法规则库，对该疑似谓语动词是否在该句中作为谓语成分出现进行甄别；若判断出不是谓语动词，则流程不进行下面的步骤，直接输出句中无谓语动词的提示信息；若判断出是谓语动词，则转入进行动词搭配组识别；

[0046] Step2.2、若疑似谓语动词个数超过1个，则逐个对这些词进行形态分析，若可以判定，则转入进行动词搭配组识别；若不能判定，则利用句法规则库进行判定，若判断出不是谓语动词，则流程不进行下面的步骤，直接输出句中无谓语动词的提示信息，若判断出是谓语动词，则转入进行动词搭配组识别。

[0047] 所述步骤Step3中所述的动词搭配组识别，判别该动词是否以动词搭配组的形式在该句子中展现，具体是：

[0048] Step3.1、若初步判定是二元动词搭配组，则再通过词语窗口，窗口长度视不同语种而定，判断该二元动词搭配组后面是否有与之搭配的词出现，若有，则通过ϕ^2统计法来判别该动词搭配的真伪；若无，则将该二元动词搭配组中的动词作为待测句子最终的谓语动词进行输出；

[0049] Step3.2、若初步判定是三元动词搭配组或是更多元的动词搭配组，再通过将其匹配多元动词搭配库的形式进行真伪判别。

[0050] 实施例2：如图1-2所示，一种基于句法特性与统计融合的自然语言谓语动词识别方法，所述基于句法特性与统计融合的自然语言谓语动词识别方法的具体步骤如下：

[0051] Step1、对待测句子进行预处理分析：输入句子，通过文本语种识别工具判定语种为英语语种，使用词性标注工具(例如stanford大学的词性标注工具)对句子中的词逐个进行词性标注，然后对分析谓语动词不相关的词性，如语气词等进行过滤处理，接下来，根据词性标注结果抽取出疑似动词，若无疑似动词，则直接输出句中无谓语动词的提示信息；若有疑似动词，则进行如下步骤Step2；例如：What an interesting story!中没有疑似动词，则直接输出句中无谓语动词的提示信息；若有疑似动词，可根据词性判断，对分析谓语动词

不相关的词(如语气词,部分副词等)进行过滤处理,也可以进行步骤Step2。

[0052] Step2、疑似谓语动词的排查:通过疑似谓语动词的形态分析(如原形,过去式,过去分词还是动名词形式出现)和句法规则库得到疑似谓语动词;

[0053] Step3、动词搭配组识别:将疑似谓语动词的词找到后,分析该谓语动词是否是以动词搭配组的形式出现,如果不是,则把该疑似谓语动词作为待测句子的谓语动词输出,如果是,则进行动词搭配组的识别,其中,利用Φ^2统计法来判别该动词搭配组的真伪;例如:make up,go on;这里我们利用Φ^2统计法判定其搭配的真伪。

[0054] Step4、根据上述步骤,输出所识别出待测句子的谓语动词或是谓语动词搭配组信息。

[0055] 所述步骤Step1中,对待测句子进行词性标注、对应的过滤处理和疑似动词抽取,其操作步骤如下:

[0056] Step1.1、对输入的待测句子通过文本语种识别工具判定语种,通过分词工具进行分词并对切分出来的单词标注词性;

[0057] Step1.2、根据标注的词性判断,若无疑似动词,则不进行下面的一系列分析,直接输出句中无谓语动词的提示信息;若存在疑似动词,则进行步骤Step1.3;

[0058] Step1.3、存在疑似动词,则对分析谓语动词不相关的词性,如语气词,部分副词等进行过滤处理,用于减轻句法分析负担,提高识别效率。例如:语气词(oh、hi、hello、wow等),大部分副词(wonderfully、quickly、sadly、surprisingly等),如此便会减轻接下来的句法分析负担,提高识别效率;

[0059] 所述步骤Step2中所述的疑似谓语动词排查,其具体步骤如下:

[0060] Step2.1、若疑似谓语动词个数为1,则结合形态分析和句法规则库,对该疑似谓语动词是否在该句中作为谓语成分出现进行甄别;若判断出不是谓语动词,则流程不进行下面的步骤,直接输出句中无谓语动词的提示信息;若判断出是谓语动词,则转入进行动词搭配组识别;例如:What an amazing book!疑似动词有一个,则结合形态分析和句法规则库,对该疑似动词是否在该句中作为谓语成分出现进行甄别;该book前面是一ADJ形容词,说明该book不是谓语动词,则流程不进行下面的步骤,直接输出句中无谓语动词的提示信息;

[0061] Step2.2、若疑似谓语动词个数超过1个,则逐个对这些词进行形态分析,若可以判定,则转入进行动词搭配组识别;若不能判定,则利用句法规则库进行判定,若判断出不是谓语动词,则流程不进行下面的步骤,直接输出句中无谓语动词的提示信息,若判断出是谓语动词,则转入进行动词搭配组识别。例如,英语中比较句中出现的助动词(如do、will、would等)和这些疑似动词的距离大小判定最有可能是谓语动词角色的词,通过判定则转入Step3步骤中的动词搭配组识别。例如:I book some books.有两个疑似动词,而且,两个动词的原形一样,那么,通过上下文分析,第一个book前一个词性是代名词,第二个books前面是形容词词性,则系统自动判定出第一个book便是我们要找的谓语动词。

[0062] 所述步骤Step3中所述的动词搭配组识别,判别该动词是否以动词搭配组的形式在该句子中展现,具体是:

[0063] Step3.1、若初步判定是二元动词搭配组,则再通过词语窗口(人工设定其长度),窗口长度视不同语种而定,判断该二元动词搭配组后面是否有与之搭配的词(介词或是副词)出现,若有,则通过Φ^2统计法来判别该动词搭配的真伪;若无,则将该二元动词搭配组

中的动词作为待测句子最终的谓语动词进行输出；

[0064] Step3.2、若初步判定是三元动词搭配组或是更多元的动词搭配组，再通过将其匹配多元动词搭配库的形式进行真伪判别。

[0065] 详细的Φ^2统计法用于判定动词搭配组真伪的方法如下：

[0066] 对于Φ^2统计法语料库中的动词，依次统计它们在英语语料库中可能出现的搭配组合。并将统计出的各个动词的词频、能够和动词构成搭配的小品词词频以及动词与小品词共同出现的词频存入数据库，以提高后续查询计算的速度。

[0067] 选择大学高年级英语语料库(由开放的CLEC提供)，总计单词量为239387个。如统计动词make的搭配情况，那么，经过统计，可计算出语料库中make(包括make的变形：makes、made、making)和其后所跟的小品词w_2的频次，将其一一存入数据库。

[0068] 表2对于两个单词w_1和w_2，建立关联表，以make(w_1)up(w_2)为例：

[0069]

	up	!up	$\sum$
make(包含其三种变形：makes、made、making)	$a=18$	$b=592$	$a+b=610$
!make(包含其三种变形：makes、made、making)	$c=297$	$d=238480$	$c+d=238777$
$\sum$	$a+c=315$	$b+d=239072$	$a+b+c+d=239387$

[0070] 上表中，a表示单词make、up共同出现的次数，b表示不在单词make、up中的make的出现次数，c表示不在单词make、up中的up的出现次数，d表示既不是make又不是up的词的次数，a+b是make出现的总词数，c+d是非make的总词数，a+c是up的出现词数，b+d是非up的总词数，N＝a+b+c+d表示语料库中的总词数。统计可得到的数据有a、a+b、a+c、a+b+c+d，已在表中展现，表中的其他数据是由上述统计得到的数据计算而来。

[0071] 因此根据上面的联立表，Φ^2统计量计算如下公式(1)：

[0072] $$\phi^2=\frac{(a\times d-b\times c)^2}{(a+b)\times(a+c)\times(b+d)\times(c+d)} \quad (1)$$

[0073] 将表中的相应数据代入公式(1)，则统计量$\Phi^2\approx0.001545$。

[0074] 当统计量Φ^2值越大，说明make(包含其变形)与它后面的小品词w_2共现的机会越多，即它们是动词搭配组的概率越大，通过设置门限T和计算统计量Φ^2，若统计量$\Phi^2>T$，则系统自动将该动词搭配组识别为真动词搭配组，否则，识别为伪动词搭配组。而对于由三个词组合的动词搭配组(如：take care of)，通过其与人工整理的动词搭配库进行匹配，若匹配成功，则系统自动识别为真搭配组，否则为伪搭配组。

[0075] 实施例3：如图1-2所示，一种基于句法特性与统计融合的自然语言谓语动词识别方法，本实施例以壮族语言为背景做谓语动词识别，

[0076] 法定的壮文是拼音文字，由拉丁字母组成，用以拼写壮语标准音的一套书写符号系统，它以北部方言为基础，以武鸣县的语言为标准音组成壮文的书写规范。壮文中的词无词形变化，而次序和虚词是表达语法意义的主要手段；

[0077] 所述基于句法特性与统计融合的自然语言谓语动词识别方法的具体步骤如下：

[0078] Step1、对待测句子进行预处理分析：输入句子，通过文本语种识别工具判定语种为壮族语言，使用词性标注工具对句子中的词逐个进行词性标注，然后对分析谓语动词不相关的词性，如语气词等进行过滤处理，接下来，根据词性标注结果抽取出疑似动词，若无疑似动词，则直接输出句中无谓语动词的提示信息；若有疑似动词，则进行如下步骤Step2；

[0079] Step2、疑似谓语动词的排查：通过疑似谓语动词的形态分析和句法规则库得到疑似谓语动词；

[0080] Step3、动词搭配组识别：将疑似谓语动词的词找到后，分析该谓语动词是否是以动词搭配组的形式出现，如果不是，则把该疑似谓语动词作为待测句子的谓语动词输出，如果是，则进行动词搭配组的识别，其中，利用ϕ^2统计法来判别该动词搭配组的真伪；例如：hau poi(进去)；这里利用ϕ^2统计法进行判定。

[0081] Step4、根据上述步骤，输出所识别出待测句子的谓语动词或是谓语动词搭配组信息。

[0082] 所述步骤Step1中，对待测句子进行词性标注、对应的过滤处理和疑似动词抽取，其操作步骤如下：

[0083] Step1.1、对输入的待测句子通过文本语种识别工具判定语种为壮文，通过分词工具进行分词并对切分出来的单词标注词性；

[0084] Step1.2、根据标注的词性判断，若无疑似动词，则不进行下面的一系列分析，直接输出句中无谓语动词的提示信息；若存在疑似动词，则进行步骤Step1.3；

[0085] Step1.3、存在疑似动词，则对分析谓语动词不相关的词性，如语气词，部分副词等进行过滤处理，用于减轻句法分析负担，提高识别效率。

[0086] 所述步骤Step2中所述的疑似谓语动词排查，其具体步骤如下：

[0087] Step2.1、若疑似谓语动词个数为1，结合壮文动词语法规范做进一步确认，确认完毕，输出该句相应的谓语动词信息；

[0088] Step2.2、若疑似谓语动词个数超过1个，则逐个对这些词进行上下文分析，若可以判定，则转入进行动词搭配组识别；若不能判定，则利用句法规则库进行判定，若判断出不是谓语动词，则流程不进行下面的步骤，直接输出句中无谓语动词的提示信息，若判断出是谓语动词，则转入进行动词搭配组识别。例如谓语动词总是和宾语距离最近，便可排除掉另一个充当副词成分的动词做谓语的误判；

[0089] 例如：put au saə tau.

[0090] 跑拿书来。

[0091] 即：跑去拿书来。

[0092] 那么这个例子中的“put”和“au”都有可能是谓语动词，根据上述句法特性，“au”距离“saə”比较近，故排除“put”，该句的谓语动词为“au”。

[0093] 所述步骤Step3中所述的动词搭配组识别，判别该动词是否以动词搭配组的形式在该句子中展现，具体是：

[0094] Step3.1、若初步判定是二元动词搭配组，则再通过词语窗口(人工设定其长度)，窗口长度视不同语种而定，判断该二元动词搭配组后面是否有与之搭配的词(介词或是副词)出现，若有，则通过Φ^2统计法来判别该动词搭配的真伪；若无，则将该二元动词搭配组中的动词作为待测句子最终的谓语动词进行输出；

[0095] Step3.2、若初步判定是三元动词搭配组或是更多元的动词搭配组，再通过将其匹配多元动词搭配库的形式进行真伪判别。

[0096] 对于壮文，可以借助《武鸣土语》《武鸣壮族民间故事》为语料库，通过Φ^2统计法进行动词搭配组的判别。例如发现hau poi(进去)对应的Φ^2值>设定的门限T，则系统自动判定它们是真动词搭配组，否则判定为假搭配组。

[0097] 而对于三词以上的词所组成的典型动词搭配组，例如:luan lonθam cman(胡说八道)，采用匹配多元动词搭配库(人工整理)的形式进行判别，若匹配成功，则系统判定为真搭配组，否则判定为伪搭配组。

[0098] 最后，结合步骤Step2和步骤Step3的分析结果，将系统的谓语动词或是谓语动词搭配组信息输出。

[0099] 上面结合附图对本发明的具体实施方式作了详细说明，但是本发明并不限于上述实施方式，在本领域普通技术人员所具备的知识范围内，还可以在不脱离本发明宗旨的前提下作出各种变化。

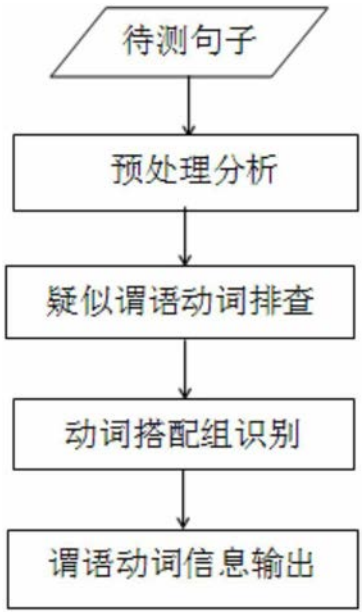

图1

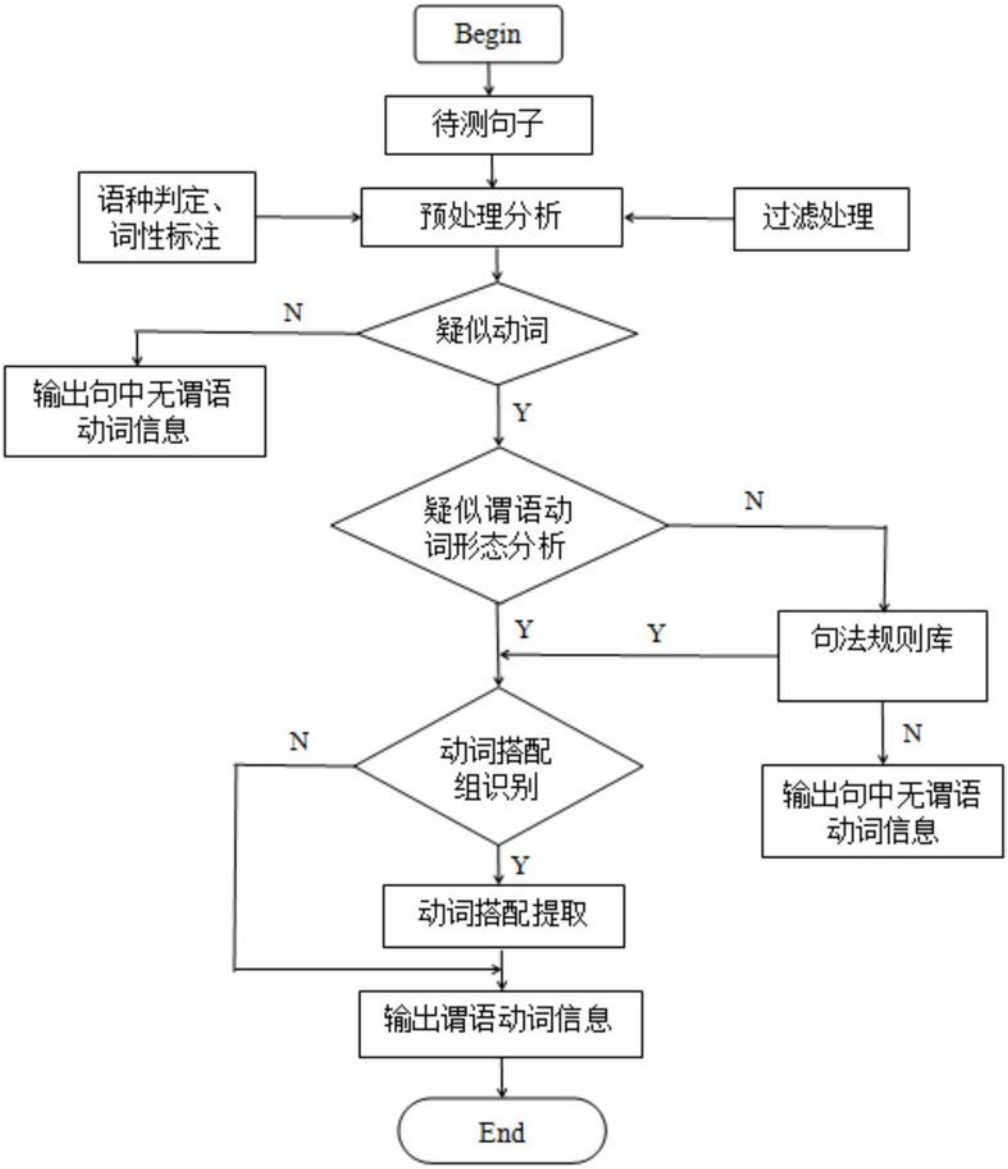
Begin
待测句子
语种判定、
词性标注
预处理分析
过滤处理
疑似动词
N
输出句中无谓语
动词信息
Y
疑似谓语动
词形态分析
N
Y
Y
句法规则库
N
动词搭配
组识别
N
输出句中无谓语
动词信息
Y
动词搭配提取
输出谓语动词信息
End

图2