(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 101075228 B
(45)授权公告日 2012.05.23

(21)申请号 200610079890.5
(22)申请日 2006.05.15
(73)专利权人 松下电器产业株式会社
地址 日本大阪府
(72)发明人 燕鹏举 孙羽菲 续木贵史
(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 王玮
(51)Int.Cl.
G06F 17/27(2006.01)
(56)对比文件
CN 1352774 A,2002.06.05, 全文.
US 20030208354 A1,2003.11.06, 权利要求1-3.

审查员 吴敏

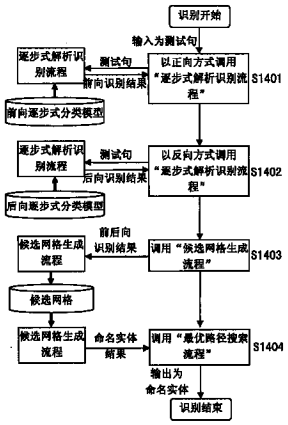
权利要求书 3 页 说明书 16 页 附图 11 页

(54)发明名称

识别自然语言中的命名实体的方法和装置

(57)摘要

本发明提供了一种识别自然语言中的命名实体的方法，包括步骤：对自然语言执行逐步式解析模型训练，以获得分类模型；基于得到的所述分类模型对自然语言执行逐步式解析识别，以得到侯选命名实体的位置和类型信息；利用拒识器对侯选命名实体进行拒识处理；和对经过拒识处理的侯选命名实体生成侯选命名实体网络，并执行最优路径搜索。本发明使用候选命名实体的全局特征，在得到仅使用局部特征的前向解析识别结果和后向解析识别结果的基础上，使用一个单类分类器对这些结果进行打分或评判，来得到最为可靠的命名实体起始和终止边界。



CN 101075228 B

1. 一种识别自然语言中的命名实体的方法，包括步骤：

利用逐步式识别器识别出候选命名实体；

利用拒识器抽取识别出的侯选命名实体基于字的全局特征；

使用所述全局特征来测试所述侯选命名实体；

如果测试得分超过一个事先给定的阈值，则接受所述候选命名实体，否则被拒识；和

将拒识器识别出的候选命名实体组成网格，在此网格上搜索拒识得分最大的路径。

2. 根据权利要求 1 所述的方法，其中使用全局特征测试候选命名实体的步骤包括将全局特征作为输入，使用单类支撑向量机对候选命名实体进行测试的步骤。

3. 根据权利要求 1 所述的方法，其中所述逐步式识别步骤包括前向逐步式解析训练步骤和后向逐步式解析训练步骤，以分别生成前向逐步式分类模型和后向逐步式分类模型。

4. 根据权利要求 1 所述的方法，其中所述拒识得分等于单类支撑向量机的测试得分减去所述阈值。

5. 一种识别自然语言中的命名实体的方法，包括步骤：

使用一个特征窗口，对窗口中心包含的词或字进行局部特征抽取；

基于对自然语言执行逐步式解析模型训练后所得到的分类模型，对自然语言执行逐步式解析识别，以得到侯选命名实体的位置和类型信息；

使用一个特征窗口，对窗口中心包含的候选命名实体进行全局特征抽取；

利用拒识器对侯选命名实体进行拒识处理；和

对经过拒识处理的侯选命名实体生成侯选命名实体网络，并执行最优路径搜索。

6. 根据权利要求 5 所述的方法，其中所述逐步式解析模型训练步骤包括使用执行前向逐步式解析模型训练得到的前向逐步式分类模型，和执行后向逐步式解析模型训练得到的后向逐步式分类模型。

7. 根据权利要求 5 所述的方法，其中所述逐步式解析识别步骤包括对读取的词进行局部特征抽取，并基于这些局部特征进行解析的步骤。

8. 根据权利要求 7 所述的方法，进一步包括特征抽取模块得到表示所述侯选命名实体的所有特征的多维向量，使用逐步式分类模型对得到的多维向量进行分类的步骤。

9. 根据权利要求 5 所述的方法，进行拒识处理的步骤包括对候选命名实体中的字进行全局特征抽取，针对得到的全局特征在单类分类器上使用学习得到的模型进行拒识处理的步骤。

10. 根据权利要求 5 所述的方法，进行拒识处理的步骤包括，由拒识器抽出候选命名实体，计算出表示这些候选命名实体的准确性得分，根据算出的得分，对候选命名实体进行接受或拒绝的处理。

11. 根据权利要求 5 所述的方法，进行拒识处理的步骤使用的是对于不同类别的命名实体使用不同的训练集而得到的拒识器。

12. 根据权利要求 5 所述的方法，其中所述拒识处理步骤使用单类支撑向量机测试该侯选命名实体，如果测试得分超过预定的阈值，则接受所述候选命名实体，否则拒识所述候选命名实体。

13. 根据权利要求 5 所述的方法，其中在所述拒识处理步骤中，如果前向和后向解析得到了同一个候选命名实体，则使用单类支撑向量机和相同侯选阈值对该候选命名实体进行

评价。

14. 根据权利要求 5 所述的方法，其中在所述拒识处理步骤中，如果一个单遍解析得到一个候选命名实体，而另一个单遍解析未得到与所述候选在位置上相交的其它候选命名实体，则使用单类支撑向量机和自由阈值对候选命名实体进行评价。

15. 根据权利要求 5 所述的方法，其中在所述拒识处理步骤中，如果前向解析得到一个候选命名实体，后向解析得到另一个候选命名实体，且两个候选在位置上相交，而且具有相同类型，则根据位置重叠情况至多生成两个同类型的新候选，使用单类支撑向量机和重叠侯选阈值对至多 4 个候选进行评价。

16. 根据权利要求 5 所述的方法，其中在所述拒识处理步骤中，如果两遍解析得到 2 个在位置上相交的、且类型不同的候选命名实体，则使用单类支撑向量机和前后向冲突侯选阈值对这 2 个候选进行评价。

17. 一种识别自然语言中的命名实体的离线训练方法，包括步骤：

对自然语句进行前向逐步式解析模型训练，以得到前向逐步式分类模型；

对所述自然语句进行后向逐步式解析模型训练，以得到后向逐步式分类模型；和

根据得到的前向逐步式分类模型和后向逐步式分类模型对侯选命名实体进行拒识模型训练，以得到拒识分类模型。

18. 一种识别自然语言中的命名实体的在线识别方法，包括步骤：

使用前向逐步式分类模型对自然语言进行识别，得到前向识别结果；

使用后向逐步式分类模型对自然语言进行识别，得到后向识别结果；

根据所述前向识别结果和所述后向识别结果生成侯选网格；和

使用生成的所述侯选网格来计算最优路径，并输出命名实体。

19. 一种识别自然语言中的命名实体的离线训练系统，包括：

局部特征抽取装置，用于使提供的训练文本生成一个以特征向量和样本标记表示的命名实体训练样本；

多类支撑向量机训练装置，用于对训练文本进行训练，生成逐步式分类模型；

全局特征抽取装置，用于使命名实体训练样本生成一个基于字的以特征向量和样本标记表示的拒识训练样本；

单类支撑向量机训练装置，用于对得到的拒识训练样本进行拒识训练，以生成拒识分类模型；

训练样本存储器，用于存储训练过程中使用的训练文本。

20. 一种识别自然语言中的命名实体的在线识别系统，包括：

局部特征抽取装置，用于使提供的测试样本生成局部特征向量；

多类支撑向量机识别装置，用于根据样本的局部特征向量对输入的样本进行识别，以得到侯选命名实体；

全局特征抽取装置，用于对候选命名实体及其上下文抽取全局特征向量；和

单类支撑向量机识别装置，用于根据样本的全局特征向量对输入的候选命名实体进行识别；

其中所述多类支撑向量机识别装置，利用多类分类模型，对输入的局部特征向量进行测试以得到其类别标记，并根据属于同一种类型的命名实体的一串起始和继续标记，形成

一个候选命名实体，所述单类支撑向量机识别装置，利用单类分类模型，对输入的全局特征向量进行测试以得到其测试得分，从得到的测试得分减去不同的阈值得到拒识得分，根据拒识得分进行最优路径搜索，和接受最优路径上的候选命名实体。

21. 根据权利要求 20 所述的系统，还包括：

拒识打分装置，用于根据所述单类支撑向量机识别装置得到的候选命名实体识别结果，以及候选命名实体的位置关系，确定不同的阈值，以计算拒识得分，并根据计算出的拒识得分，来接受或拒绝候选命名实体。

22. 根据权利要求 20 所述的系统，还包括：

最优路径搜索装置，用于根据候选命名实体的位置和拒识得分，搜索拒识得分之和最大的最优路径。

识别自然语言中的命名实体的方法和装置

技术领域

[0001] 本发明涉及语言处理方法和系统，特别是涉及识别自然语言中的命名实体的方法和系统，从而能够提取语言信息，进行相应的处理。

背景技术

[0002] 命名实体是指包括人名、地名、机构名、时间、数量等特定种类词的集合。命名实体识别在信息提取、信息检索方面有着广泛的应用。

[0003] 近年来，逐步式的命名实体 (named entity，NE) 识别或语块 (chunk) 识别方法表现了比较高的性能。Taku Kudo，Yuji Matsumoto 在 2001 年的 NAACL 上发表的题为 Chunking with Support Vector Machines 的文章对此做了说明。这些方法的主要特征是将识别分成若干前后相继的步骤，每一步扫描输入句子中的一个词，通过观察当前词的上下文 (context) 特征 (feature)，使用预定或者统计 (stochastic) 的方法预测当前词的标记 (token)。不同的方法使用不同的标记集合，但基本上包括 B、I、E 和 O 四种，分别表示命名实体的起始 (B)、中间 (I)、结束 (E) 位置和不属于命名实体 (O)。在输入句子中所有词的标记确定之后，所有 B、I、E 标记串就直接组成了命名实体。在识别中的每一步，识别器使用的特征是包含在以当前词为中心的一个特征窗口内的局部特征。

[0004] 表 1 是一个从句子的开始位置解析 (parsing) 到句子结束位置的方法示例，下文称为前向解析。

[0005] 表 1

[0006]

	→ 解析方向 →				
词	L_2	L_1	C	R_1	R_2
特征	F_L^2	F_L^1	F_C	F_R^1	F_R^2
标记	T_L^2	T_L^1	T_C	N/A*	N/A*
	前续词		当前词	后续词	

[0007] 在表 1 中 C 表示当前词，L_1 和 L_2 是当前词的左上下文，而 R_1 和 R_2 是右上下文。特征窗口的大小是 5，F_C、F_L^1、F_L^2、F_R^1和F_R^2是特征窗口内每个词对应的特征，而T_L^1、T_L^2是前续词的识别标记。N/A 表示当前时刻该特征还无法得到。

[0008] 所谓特征是指所有一切在上下文中可以观察得到的信息。例如，这个词是什么，词的长度，词性是什么，前面决定的该词对应的标记是什么，等等，如下面的表 2 所示。具体使用什么样的特征，由系统设计人员根据应用的特点来进行选定，目标是使系统达到最高识别性能。在表 2 所示的前向解析中，当系统观察到所有这些特征时，它就可能对当前词“邓”作出“B-PER”的标记预测。

[0009] 表 2

[0010]

	→ 解析方向 →				
词	决心	继承	邓	小平	同志
特征	{词=决心, 词长=2, 词性=副词, 标记=O}	{词=继承, 词长=2, 词性=动词, 标记=O}	{词=邓, 词长=1,词性=人名词}	{词=小平, 词长=2, 词性=人名词}	{词=小平, 词长=2, 词性=动词}
标记	O	O	B-PER*	N/A	N/A
	前续词		当前词	后续词	

[0011] 其中 B-PER 标记表示当前词是一个人名的开始。

[0012] 在表 2 给出的示例中，以“继承”为例，在第三行中给出了该词的特征为：词的内容是“继承”，词的长度是 2，词性为动词，标记为 O（说明其不是命名实体）。

[0013] 从上面的说明可以看到，逐步式的识别方法有一个缺点，就是只能使用一个固定大小的特征窗口内的局部特征。由于长程 (long distance) 特征没有得到使用，会造成起始边界 B 标记的误警 (false alarm)，即不是命名实体起始边界的地方有可能被识别器认为是一个起始边界。ManabuSassano，Takehito Utsuro 在 COLING2000 :705-711 中发表的题为“NamedEntity Chunking Techniques in Supervised Learning for Japanese NamedEntity Recognition”的文章提出一个可变长度模型 (Variable Length Model) 的方法。其中特征窗口的大小可以在一个预先确定的范围内变化，可以看出，该方法仍然不能处理任意长度范围内的特征。

[0014] 一些基于概率 (probabilistic) 模型的方法可以使用全局特征。例如，2000 年 2 月 17 日提交的题为“System for Chinese tokenization and namedentity recognition”的美国专利申请 No.09/403,069。然而，概率模型方法受数据稀疏 (data sparseness) 问题的影响比较大，而且需要使用复杂的解码 (decoding) 方法在庞大的候选 (candidate) 网格 (lattice) 空间中进行搜索。当训练 (training) 数据不够，或者计算资源不够的情况下（比如嵌入式设备），概率模型不具备可行性。

[0015] 另外，当前的命名实体识别方法受切分词 (word segmentation) 错误的影响很大。在基于分词结果之上进行的命名实体识别，没有办法恢复分词过程中被错分的边界，从而影响命名实体识别的正确性。如表 3 给出的例子所示，由于“北京市 | 公安 | 局长 | 江 | 金福”被错误地切分成“北京市 | 公安局 | 长江 | 金 | 福”，这直接导致“北京市 | 公安局”片断被错误地识别成了一个类型为 ORG（机构名）的命名实体。而实际上，这个句子的“……北京市公安局……”这个部分中并没有命名实体，而是在句子后部存在一个真正的 PER（人名）类型的命名实体，即“江金福”。此时，使用基于字 (character) 的模型会避免分词错误引起的这种后果。

[0016] 表 3

[0017]

字	[句首]	北	京	市	公	安	局	长	江	金	福
正确分词		北京市			公安		局长		江	金福	
预测分词		北京市			公安局			长江		金	福
预测词性		ns			n			ns		Ng	n
基于词的命名实体识别		ORG									

[0018] 上面提到的Kudo等人使用投票(voting)方法对正向和反向识别结果作出选择以决定最终标记，但投票结果是针对每个步骤的标记识别结果而言的，所以使用的仍是局部特征。此外其它文献中也披露了很多其它分类器(classifier)结合的方法，然而，这些方法都没有使用全局特征。

发明内容

[0019] 鉴于上述问题，本发明的目的是提供一种识别自然语言中的命名实体的方法和系统，使用候选命名实体的全局特征，在得到仅使用局部特征的前向解析识别结果和后向解析识别结果（即候选命名实体）的基础上，使用一个单类分类器对这些结果进行打分或评判，来得到最为可靠的命名实体起始和终止边界。

[0020] 根据本发明的一个方面，提供一种识别自然语言中的命名实体的方法，包括步骤：利用逐步式识别器识别出候选命名实体；利用拒识器抽取识别出的侯选命名实体基于字的全局特征；使用所述全局特征来测试所述侯选命名实体；和如果测试得分超过一个事先给定的阈值，则接受所述候选命名实体，否则被拒识。

[0021] 根据本发明的另一个方面，提供一种识别自然语言中的命名实体的方法，包括步骤：使用一个特征窗口，对窗口中心包含的词或字进行局部特征抽取；基于对自然语言执行逐步式解析模型训练后所得到的分类模型，对自然语言执行逐步式解析识别，以得到侯选命名实体的位置和类型信息；使用一个特征窗口，对窗口中心包含的候选命名实体进行全局特征抽取；利用拒识器对侯选命名实体进行拒识处理；和对经过拒识处理的侯选命名实体生成侯选命名实体网络，并执行最优路径搜索。

[0022] 根据本发明的再一个方面，提供一种识别自然语言中的命名实体的离线训练方法，包括步骤：对自然语句进行前向逐步式解析模型训练，以得到前向逐步式分类模型；对所述自然语句进行后向逐步式解析模型训练，以得到后向逐步式分类模型；和根据得到的前向逐步式分类模型和后向逐步式分类模型对侯选命名实体进行拒识模型训练，以得到拒识分类模型。

[0023] 根据本发明的再一个方面，提供一种识别自然语言中的命名实体的在线识别方

法，包括步骤：使用前向逐步式分类模型对自然语言进行识别，得到前向识别结果；使用后向逐步式分类模型对自然语言进行识别，得到后向识别结果；根据所述前识别结果和所述后向识别结果生成侯选网格；和使用生成的所述侯选网格来计算最优路径，并输出命名实体。

[0024] 根据本发明的再一个方面，提供一种识别自然语言中的命名实体的离线训练系统，包括：局部特征抽取装置，用于使提供的训练文本生成一个以特征向量和样本标记表示的命名实体训练样本；多类支撑向量机训练装置，用于对训练文本进行训练，生成逐步式分类模型；全局特征抽取装置，用于使命名实体训练样本生成一个基于字的以特征向量和样本标记表示的拒识训练样本；单类支撑向量机训练装置，用于对得到的拒识训练样本进行拒识训练，以生成拒识分类模型；训练样本存储器，用于存储训练过程中使用的训练文本。

[0025] 根据本发明的再一个方面，提供一种识别自然语言中的命名实体的在线识别系统，包括：局部特征抽取装置，用于使提供的测试样本生成局部特征向量；多类支撑向量机识别装置，用于根据样本的局部特征向量对输入的样本进行识别，以得到侯选命名实体；全局特征抽取装置，用于对候选命名实体及其上下文抽取全局特征向量；和单类支撑向量机识别装置，用于根据样本的全局特征向量对输入的候选命名实体进行识别；其中所述多类支撑向量机识别装置，利用多类分类模型，对输入的局部特征向量进行测试以得到其类别标记，并根据属于同一种类型的命名实体的一串起始和继续标记，形成一个候选命名实体，所述单类支撑向量机识别装置，利用单类分类模型，对输入的全局特征向量进行测试以得到其测试得分，从得到的测试得分减去不同的阈值得到拒识得分，根据拒识得分进行最优路径搜索，和接受最优路径上的候选命名实体。

[0026] 根据本发明，使用全局特征的命名实体识别方法。可以拒识逐步式命名实体识别方法产生的不可靠候选命名实体（具有不可靠起始边界或不可能结束边界）。另外，基于字的特征抽取避免了分词错误带来的影响。通过结合前向和后向两遍解析结果，使命名实体的识别性能得到提高。

附图说明

[0027] 通过阅读和理解下面参考附图对本发明优选实施例所做的详细描述，将使本发明的这些和其它目的、特征、和优点变得显而易见。其中：

[0028] 图 1 是表示在命名实体识别中采用两类分类器进行分类的示意图；

[0029] 图 2 是表示在命名实体识别中采用单类分类器进行分类的示意图；

[0030] 图 3 示出了调整阈值时精确度、召回率以及 F-measure 之间的关系示意图；

[0031] 图 4 示出了调整阈值时精确度、召回率以及 F-measure 之间的关系示意图；

[0032] 图 5 示出了调整阈值时精确度、召回率以及 F-measure 之间的关系示意图；

[0033] 图 6 示出了调整阈值时精确度、召回率以及 F-measure 之间的关系示意图；

[0034] 图 7 是表示根据本发明实施例的命名实体识别过程中的逐步式解析模型的训练流程图；

[0035] 图 8 是表示根据本发明实施例的命名实体识别过程中的逐步式解析识别的流程图；

[0036] 图 9 是表示根据本发明实施例的命名实体识别过程中的拒识模型训练的流程图；

[0037] 图 10 是表示根据本发明实施例的命名实体识别过程中的拒识打分的流程图；

[0038] 图 11 是表示根据本发明实施例的命名实体识别过程中的侯选网格生成的流程图；

[0039] 图 12 是表示根据本发明实施例的命名实体识别过程中的最优路径搜索示意图；

[0040] 图 13 是表示根据本发明实施例的命名实体识别过程中的离线训练的总流程图；

[0041] 图 14 是表示根据本发明实施例的命名实体识别过程中的在线训练的总流程图；

[0042] 图 15 是表示根据本发明实施例的命名实体识别装置的离线训练系统的方框图；和

[0043] 图 16 是表示根据本发明实施例的命名实体识别装置的在线训练系统的方框图。

具体实施方式

[0044] 下面参照附图对本发明的实施例进行详细说明，在描述过程中省略了对于本发明来说是不必要的细节和功能，以防止对本发明的理解造成混淆。

[0045] 下面首先对命名实体全局建模的方式进行描述，以便更好地理解本发明。

[0046] 如果将命名实体作为一个整体，一个命名实体的特征可以用其左上下文、右上下文和其内部特征表示。如下面表 4 给出的示例所示，目标命名实体的特征由F_L^2、F_L^1、F_{In}^1、F_{In}^m、F_R^1、F_R^2，以及诸如命名实体长度 m 等组成。

[0047] 表 4

[0048]

词	L_2	L_1	$In_1, In_2, \cdots, In_m$	R_1	R_2
特征	F_L^2	F_L^1	$F_{In}^1, F_{In}^2, \cdots, F_{In}^m$	F_R^1	F_R^2
	左上下文		命名实体内部	右上下文	

[0049] 与表 1 所示示例的（词的）局部特征选取方法相比，表 1 所示示例所关注的是当前单个词的上下文特征，而表 4 所示方法关注的是一个命名实体整体的特征。这样，无论命名实体的长度有多大，总能观察到该命名实体的右上下文特征。因此，本发明把这种特征选取方法称为命名实体的全局特征。

[0050] 表 5 给出了一个具体的全局特征示例。除了前文所述基于词的特征之外（例如，这个词是什么，词有多长，词性是什么，等等），还可以包括命名实体的长度、类型等等。

[0051] 表 5

[0052]

词	决心	继承	邓 | 小平	同志	的
特征	{词=决心，词长=2，词性=副词}	{词=继承，词长=2，词性=动词}	{包含词=(邓,小平)，长度=2,类型=人名}	{词=同志，词长=2，词性=动词}	{词=的，词长=2，词性=助词}
	左上下文		命名实体内部	右上下文	

[0053] 训练集中的命名实体样本(sample)用于训练命名实体的全局模型。可以采用两种建模方法,一是两类分类器的建模方法,另一类是单类分类器的建模方法。

[0054] 现有技术中已经揭示了上述有关的分类器的详细内容。鉴于分类器本身并不是本发明的内容所在,在此省略对分类器的具体描述。

[0055] 下面简单说明这两种建模的实现方法。使用两类分类器时,需要收集足够的正样本(在本发明中是指命名实体)和负样本(在本发明中是指"非"命名实体)。将正、负样本的特征表示成高维空间中的向量(或点),训练过程就是选用一个两类分类器学习以得到这两类样本的分类面。训练完成后,当需要测试一个新样本时,只需检测该样本相对于分类面的位置,即可作出该样本是正样本还是负样本的预测。而测试样本与分类面的距离也代表着分类器对该样本所作预测可靠性,距离分类面越远,则可靠性越高,反之,可靠性越低。

[0056] 如图 1 所示,以圆圈表示正样本,以交叉表示负样本,虚线表示的是分类面,分类面内侧的测试样本将被预测为正样本。反之,将被预测为负样本。对图中以方框表示的新样本,本分类器将认为该样本是一个正样本。

[0057] 无论使用什么分类器,总会有错误分类的情况。例如,图 1 中的分类面就使得一些原来正样本可以出现的区域(本例中为分类面外侧),被认为只能出现负样本,反之亦然。这种分类器错误在所难免,而且当正样本和负样本数目不太平衡时,分类面的确定将会更加困难,导致分类错误加大。而命名实体识别正是这样一种应用,因为命名实体所占文本的百分比只有不到 10%,又考虑到由于使用的是全局特征,不同的起始、结束边界组合形成的负样本,其数量将远远大于正样本的数量,因此最后的分类面将会严重倾向于负样本,导致正样本被识别成负样本的错误机会大大增加。

[0058] 使用单类分类器时,只需要收集足够的正样本,训练过程就是选用一个单类分类器学习得到单类的"分类面",至于该分类面的形式和定义,依赖于选用的不同分类器而定。此时由于避免了负样本的选择,简化了系统的设计,并能减少系统识别误差。图 2 示出了使用单类分类器的分类示意图。在图 2 中,只收集正样本,并由此减小了识别误差。

[0059] 基于上述原因,本发明使用单类模型。作为一种实现,本发明提出以单类支撑向量机(One-Class SVM:单类SVM)作为拒识模型。这是基于单类SVM的高推广能力、有效处理高维和非线性空间的能力、以及少量训练数据即可达到较高性能的能力。简单地说,单类 SVM 算法试图去寻找能够分离训练数据和坐标原点的最佳超平面。B. Sch ¨ olkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson 在题为"Estimating thesupport of a high-dimensional distribution"的文章(见 Neural Computation, 13(7):1443-1471, 2001),和 Chih-Chung Chang and Chih-Jen Lin 在题为"LIBSVM:a library for support vector machines"的文章(见 2001. Softwareavailable at http://www.csie.ntu.edu.tw/ ～ cjlin/libsvm) 中对单类 SVM 进行了详细描述,在此省略对其的说明。

[0060] 通过这种方式,候选命名实体的全局上下文都可以被系统所利用,无论命名实体的长度有多大。

[0061] 作为本发明的一个实施例,采用了基于字的建模。下面对基于字的建模进行描述。如前文所述,一些命名实体识别错误是由分词错误所导致的。因此,用基于字的模型来替代基于词的模型在性能上会对这类识别错误有一定的修正。表 6 给出了一个基于字的全局特征建模示例,其中每个特征都是关于左上下文、右上下文和命名实体内部的字。

[0062] 表 6

[0063]

字	ChL_2	ChL_1	$ChIn_1, ChIn_2, \cdots, ChIn_m$	ChR_1	ChR_2
特征	F^2_{ChL}	F^1_{ChL}	$F^1_{ChIn}, F^2_{ChIn}, \cdots, F^m_{ChIn}$	F^1_{ChR}	F^2_{ChR}
	左上下文		命名实体内部	右上下文	

[0064] 同样，表 7 给出一个具体的基于字的建模示例。作为实例，字的特征可以包括，是否可单独成词，作出词首字、中字、末字出现的概率，等等。

[0065] 表 7

[0066]

字	继	承	邓 | 小 | 平	同	志
特征	{字=继，单独成词=0，词首=0.7，词中=0.1，词末=0.2}	{字=承，单独成词=1，词首=0.3，词中=0.2，词末=0.4}	{包含字=(邓,小,平)，长度=3,类型=人名}	{字=同，单独成词=1，词首=0.3，词中=0.1，词末=0.6}	{字=承，单独成词=1，词首=0.5，词中=0.3，词末=0.2}
	左上下文		命名实体内部	右上下文	

[0067] 例如，以表 7 中的“承”字为例，其特征行中给出了“承”字，即字为“承”，单独成词为 1，作为词首的概率是 0.3，处在词中的概率是 0.2，位于词末的概率是 0.4。作为另一个实例，作为侯选命名实体的“邓小平”的特征行中给出了字的内容，和长度，以及命名实体的类型。

[0068] 当候选命名实体被第一阶段的逐步式识别器识别出来之后，第二阶段的拒识器抽取 (extract) 该侯选命名实体基于字的全局特征。然后，将全局特征作为输入，使用单类 SVM 来测试该侯选命名实体。如果测试得分超过一个事先给定的阈值，则接受该候选命名实体，否则拒识该候选命名实体。测试得分越高说明该候选越可靠。

[0069] 这样，拥有不可靠起始边界（通常来源于前向解析结果）或不可靠结束边界（通常来源于后向解析结果）的命名实体的候选命名实体就可以被拒识。

[0070] 仅仅使用拒识方法并不一定能提高以 F-measure 为指标的系统性能 (F-measure 是精确率 precision 和召回率 recall 的折衷)，但是精确率会得到提高。然而，根据本发明，在结合了前向和后向两遍解析结果和拒识方法之后，系统性能会得到明显的提高。拒识处理过程可以描述如下：

[0071] 1. 如果前向和后向解析得到了同一个候选命名实体，则使用单类 SVM 和阈值 th_{ident} 对该候选进行评价。

[0072] 2. 如果一个单遍解析得到一个候选命名实体，而另一个单遍解析没有得到与这个候选在位置上相交的其他命名实体，则使用单类 SVM 和阈值 th_{free} 对该候选进行评价。

[0073] 3. 如果前向解析得到一个位置为 (B_{fwd}, E_{fwd}) 的候选命名实体，后向解析得到另一个位置为 (B_{bwd}, E_{bwd}) 的候选命名实体，这两个候选在位置上相交，而且它们的类型相同（例

如，都是 PER)，则根据位置重叠情况至多生成两个同类型的新候选，位置分别是 (B_{fwd}, E_{bwd}) 和 (B_{bwd}, E_{fwd})，然后使用单类 SVM 和重叠候选阈值 $th_{difbndry}$ 对这至多 4 个候选进行评价。

[0074] 4. 如果两遍解析得到 2 个在位置上相交的候选命名实体，而且它们的类型不相同，则使用单类 SVM 和前后向冲突候选阈值 th_{cnflct} 对这 2 个候选进行评价。

[0075] 5. 对于每个输入句子，所有在第一阶段得到的候选命名实体组成一个网格，网格上的每一个候选命名实体附带一个得分信息，即（拒识得分＝单类 SVM 测试得分—阈值）。在此网格上采用动态规划 (dynamicprogramming) 的方法去搜索得分之和最大的路径，这条最佳路径上的候选命名实体即予以接受并作为最终结果输出。

[0076] 图 3-6 示出了调整上面所述的各种阈值时，精确度、召回率以及 F-measure 的关系示意图。调整上述各种阈值会有不同的效果。

[0077] 图 3 中的曲线示出了调整相同候选阈值 th_{ident} 的情况。当阈值 th_{ident} 增大时，精确率会有少量提升。但当 th_{ident} 变得足够大时，召回率和 F-measure 会急剧下降。

[0078] 图 4 中的曲线示出了调整自由阈值 th_{free} 的情况。当阈值 th_{free} 增大时，精确率会稳步上升，而召回率会稳步下降。但当 th_{free} 超过一定值时，精确率和召回率趋于稳定。F-measure 会有少量上升，然后再少量下降，但基本保持在一个较小的范围内。

[0079] 图 5 中的曲线示出了调整阈值 $th_{difbndry}$ 的情况。当阈值 $th_{difbndry}$ 增大时，精确率会稳步上升，而召回率会稳步下降。但当 $th_{difbndr}$ 超过一定值，或小于一定值时，精确率和召回率趋于稳定。F-measure 会保持少量上升趋势，但基本保持在一个较小的范围内。

[0080] 图 6 中的曲线示出了调整阈值 th_{cnflct} 的情况。当阈值 th_{cnflct} 增大时，精确率会稳步上升，而召回率会先上升，后下降。F-measure 的表现和召回率类似，即先上升，后下降。

[0081] 如果使用一个集中的阈值来代替上述分立的各个阈值，以方便调整系统性能，则总的趋势是：随着阈值的增大，精确率会上升，召回率会下降，而 F-measure 会先上升，后下降。

[0082] 通过实验表明，调整上述分立的各个阈值所获得了精确度、召回率以及 F-measure 的关系变化是基于本发明的命名实体识别方法所特有的，并且可由此判断对本发明的使用。

[0083] 在一个实际系统上的实验数据显示，相对于单遍解析结果，本发明的方法可以达到 12.14％的错误下降率 (error reduction rate)。

[0084] 表 7 给出的实验中所用的是一个中文数据集，训练集包括 25,616 个命名实体，测试集包括 27,615 个命名实体，分别包含人名、地名和机构名 3 种类型的命名实体。

[0085] 这里给出召回率 (recall)、精确率 (precision) 和 F-measure 的定义：

[0086]

$$召回率=\frac{系统识别出的正确的命名实体数目}{测试集中真正的命名实体数目}*100\%,$$

[0087]

$$精确率=\frac{系统识别出的正确的命名实体数目}{系统识别出的命名实体数目}*100\%,$$

[0088]

$$F - measure = \frac{2 \cdot 召回率 \cdot 精确率}{召回率 + 精确率}。$$

[0089] 表 7 实验结果

[0090]

	召回率（%）	精确率（%）	F-measure
前向解析	91.03	90.88	90.96
后向解析	91.61	91.09	91.35
本发明方法	91.92	92.89	92.40

[0091] 利用上面给出的数据可以计算出根据本发明的方法获得的错误下降率为((92.40-91.35)/(100-91.35) = 12.14%)。

[0092] 以上对本发明的命名实体的总体方法进行了描述，下面参考附图对该方法中各个过程进行详细的描述。

[0093] 首先描述逐步式解析模型的训练过程。逐步式解析模型使用基于词的局部特征，使用多类分类器进行模型学习，其流程如图 7 所示。在开始时输入训练文本。在步骤 S701，解析模型的训练过程读取输入文本中的下一个句子。然后，在步骤 S702 使用特征窗口，对特征窗口中包含的所读取的语句进行切分词，以找出可能的命名实体。特征窗口的大小可以固定，也可以是可变的。对当前的词切分完成后，在步骤 S703 读取下一个词。此后，流程进行步骤 S704，对特征窗口中包含的读取的词或字进行局部特征抽取以提取出该词或字的特征，例如，词或字的内容，词或字的长度，词性等。接下来，在步骤 S705，把样本与其类标记一起加入到训练集中。在步骤 S706，判断读取的语句中是否还有未识别的词，如果还有未识别的词，流程返回步骤 S704，对仍未被识别的词重复执行步骤 S703 至 S706，直到识别完该语句中的所有词。如果在步骤 S706 的判断结果是该语句中的词已经识别完成，流程则进行到步骤 S707，判断文本中是否还有下一个语句。如果判断结果为肯定，即还有下一个语句，流程返回步骤 S701，读取下一个语句，然后重复步骤 S701 至 S707，识别下一个语句中的命名实体。如果步骤 S707 的判断结果为否定，流程进行到步骤 S708，对自然语言执行逐步式解析模型训练，利用形成的训练集，使用学习器根据训练样本进行分类器学习。最后输出分类模型。

[0094] 局部特征抽取可以包括词性标注模块，以得到每个词对应的词性。对于每个词样本，特征抽取模块得到的是表示了该词所有特征的一个高维向量（或点）。

[0095] 样本特征的向量化表示是非常通用和普遍的技术，而且每种应用可以有各种各样的表示方法，没有一个统一的定义或者方法，在此仅以一种表示方法为例简单说明样本特征的向量化。参见前述的表 2，此时需要表示的是以“邓”为中心，特征窗口大小为 5 的样本。可以假设系统词表大小为 50000（即含有 50,000 个词），词性表大小为 40（即含有 40 种词性），类标记集大小为 10（即含有 10 种类标记），词长为 1 维，则对于特征窗口中的每个位置，预留有 50,000+50+1 = 50,041 维，对于总共 5 个位置，则特征总空间有 50,051*5 = 250,255 维。可以假设“决心”、“继承”、“邓”、“小平”和“同志”的词号（即在词表中的序号，从 0 到 49,999）分别为 99、199、299、399 和 499，副词、动词、人名词、动词的词性号（即在词性表中的序号，从 0 到 39）分别为 0、1、2 和 3，类标记“O”的标记号（即在类标记表中的序号，从 0 到 9）为 0，则该样本的特征向量如下：

[0096] 第 100 维的值为 1（代表第 1 个位置的词为“决心”）；

[0097] 第 50,001 维的值为 1（代表第 1 个位置的词性为副词）；

[0098] 第 50,041 维的值为 1（代表第 1 个位置的类标记为“0”）；

[0099] 第 50,051 维的值为 2（代表第 1 个位置的词长为 2）；

[0100] 第 50,051+200 = 50,251 维的值为 1（代表第 2 个位置的词为“继承”）；

[0101] 第 50,051+50,002 = 100,043 维的值为 1（代表第 2 个位置的词性为动词）；

[0102] 第 50,051+50,041 = 100,092 维的值为 1（代表第 2 个位置的类标记为“0”）；

[0103] 第 50,051+50,051 = 100,102 维的值为 2（代表第 2 个位置的词长为 2）；

[0104] 第 100,102+300 = 100,402 维的值为 1（代表第 3 个位置的词为“邓”）；

[0105] 第 100,102+50,003 = 150,105 维的值为 1（代表第 3 个位置的词性为人名词）；

[0106] 第 100,102+50,051 = 150,153 维的值为 1（代表第 3 个位置的词长为 1）；

[0107] 第 150,153+400 = 150,553 维的值为 1（代表第 4 个位置的词为“小平”）；

[0108] 第 150,153+50,003 = 200,156 维的值为 1（代表第 4 个位置的词性为人名词）；

[0109] 第 150,153+50,051 = 200,204 维的值为 2（代表第 4 个位置的词长为 2）；

[0110] 第 200,204+500 = 200,704 维的值为 1（代表第 5 个位置的词为“同志”）；

[0111] 第 200,204+50,004 = 250,208 维的值为 1（代表第 5 个位置的词性为动词）；

[0112] 第 200,204+50,051 = 250,255 维的值为 2（代表第 5 个位置的词长为 2）；

[0113] 其他维的值都为 0。

[0114] 应该指出的是，在流程中，对于前向解析所需要的分类模型训练，“下一个词”指的是当前词的右边一个词，而对于后向解析，“下一个词”指的是当前词的左边一个词。

[0115] 不限定使用何种多类分类器，但作为一种实现，可以采取 SVM 来实现。两类 SVM 问题的训练和识别公式如下：

[0116] 给定训练集$\{\mathbf{x}_i, y_i\}_{i=1}^{l}$，$x_i \in R^n$，$y_i \in \{-1,1\}$，$i = 1,\ldots,l$（其中 x_i 表示训练样本的特征向量，y_i 表示训练样本的类标记），SVM 对新样本 X 作出的类标记预测公式可以用下面的公式 (1) 表示。

[0117] $$y = \mathrm{Sgn}\{\langle w, x\rangle - b\} \qquad (1)$$

[0118] 其中 w 由求解下列二次规划得到：

[0119] $$\min \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

[0120] $$\text{s.t.} \quad y_i[(w \cdot x_i)+b]-1+\xi_i \geqslant 0, \; i = 1, \cdots, n。$$

[0121] 如果该分类问题不是线性可分的 (linear inseparable)，则 SVM 使用一个隐含的映射 $x \rightarrow \Phi(x)$ 将问题映射到另外一个更高维的空间，期待在该空间下问题的可分性会更好。实际上映射函数 Φ 并不单独出现，而是体现在优化过程中的内积计算中，即用下式表示。

[0122] $$k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2)\rangle$$

[0123] 此时的 $k(x_1, x_2)$ 称为核函数 (kernel function)，以替代所有公式中出现的内积。

[0124] 由于 SVM 是处理两类分类问题的，则在处理多类（如 k）问题时，需要构建 k(k-1) 个两类 SVM 分类器，测试时使用投票方法来决定新样本的类标记。一个简单的投票策略就是多数决策，即得到最多投票的类标记被赋以新样本。

[0125] 图 8 示出了根据本发明实施例的逐步式解析识别过程的流程图。下面结合图 8 描

述逐步式解析识别过程。在开始时输入测试语句。在步骤 S801，使用特征窗口，对窗口中包含的输入的测试语句进行切分词，以找出可能的命名实体。特征窗口的大小可以固定，也可以是可变的。对当前的词切分完成后，在步骤 S802 读取下一个词。此后，流程进行到步骤 S803，对特征窗口中包含的读取的词进行局部特征抽取，并基于这些局部特征解析该词的特征，例如，词的内容，词的长度，词性等。接下来，在步骤 S804，根据参考图 7 的过程得到的分类模型，对当前的词进行类标记预测。此后，在步骤 S805，判断读取的语句中是否还有未识别的词，如果还有未识别的词，流程返回到步骤 S802，对读取仍未被识别的词重复执行步骤 S802 至 S805，直到识别完该测试语句中的所有词。如果在步骤 S805 的判断结果是该测试语句中的词已经识别完成，流程则进行到步骤 S806，将对命名实体给出的 B、I、和 E 标记组成命名实体串。此后，输出命名实体的位置和类型。

[0126] 应该指出的是，在逐步式解析识别过程中的特征抽取模块与逐步式解析模型训练过程中的特征抽取模块是一致的。另外，还要指出的是，对于前向解析流程，“下一个词”指的是当前词的右边一个词，而对于后向解析流程，“下一个词”指的是当前词的左边一个词。

[0127] 有关类标记的预测公式，可以参见前面针对逐步式解析模型训练过程的描述。

[0128] 在得到 B、I、E 及 O 标记后，连续的 B、I、I、……、I、E 标记串则被组装成命名实体。

[0129] 得到侯选命名实体后，根据本发明，需要利用拒识器对侯选命名实体进行拒识处理。下面参考图 9 描述拒识模型训练流程。

[0130] 在开始时输入训练文本。在步骤 S901，拒识模型的训练过程读取输入文本中的下一个句子。然后，在步骤 S902 读取当前语句中的侯选命名实体。此后，在步骤 S903，使用特征窗口，对特征窗口中包含的读取的侯选命名实体进行全局特征抽取，例如，词的内容，词的长度，词性等。接下来，流程进行到步骤 S904，把处理后的样本加到拒识训练集中。特征窗口的大小可以固定，也可以是可变的。在拒识训练集中，针对得到的全局特征在单类分类器上使用学习得到的模型进行拒识处理。在拒识处理中，由拒识器抽取侯选命名实体，计算出表示这些侯选命名实体的准确性得分，根据计算的得分，对侯选命名实体进行接受或拒绝处理。此后，在步骤 S905，判断读取的语句中是否还有未经过拒识处理的侯选命名实体，如果还有未处理的侯选命名实体，流程返回到步骤 S902，读取下一个侯选命名实体，并对所读取的侯选命名实体重复执行步骤 S902 至 S905，直到对该语句中的所有侯选命名实体进行了拒识处理。如果在步骤 S905 的判断结果是该语句中的侯选命名实体已经被处理完毕，流程则进行到步骤 S906，判断输入训练文本中是否还有下一个语句。如果判断结果为肯定，即还有下一个语句，流程返回步骤 S901，读取下一个语句，然后重复步骤 S901 至 S906，对下一个语句中的侯选命名实体进行拒识处理。如果步骤 S906 的判断结果为否定，流程进行到步骤 S907，利用形成的拒识训练集，使用学习器根据训练样本进行分类器学习。最后输出分类模型。

[0131] 拒识模型使用基于字的命名实体的全局特征，使用单类分类器进行模型学习。对于每个命名实体的样本，特征抽取模块得到的是表示了该命名实体所有特征的一个高维向量（或点）。使用逐步式分类模型对得到的多维向量进行分类。此处使用的特征向量化表示方法与前面的描述中使用的特征向量化表示方法类似，在此省略对其的说明。

[0132] 由于单类分类器是描述单一种类样本的可靠性的，所以对于不同类别的命名实体

（如人名、地名、机构名），要使用不同的训练集，并且训练得到不同的拒识模型。

[0133] 在本发明中，不限定使用何种单类分类器，但作为一种实现，可以采取单类 SVM 来实现。Chih-Chung Chang and Chih-Jen Lin 发表的题为“LIBSVM:a library for support vector machines”的文章（参见 2001. Software available at http://www.csie.ntu.edu.tw/～cjlin/libsvm）描述了单类 SVM 问题的训练和识别公式如下：

[0134] 对于给定的训练集 $x_i \in R^n$，SVM 对新样本 x 作出的可靠性打分公式为

[0135] $$\sum_{i=1}^{l} \alpha_i k(\mathbf{x}_i, \mathbf{x}) - \rho \ ,$$

[0136] 其中各 α_i 值由求解下列二次规划得到：

[0137] $$\min_{\alpha} \quad \frac{1}{2}\alpha^T \mathbf{Q} \alpha$$

[0138] s.t. $0 \leqslant \alpha_i \leqslant 1/(vl), i = 1, \ldots, l,,$

[0139] $e^T \alpha = 1,$

[0140] 其中 $Q_{ij} = k(x_i, x_j) \equiv \langle \Phi(x_i), \Phi(x_j) \rangle$。

[0141] 得到拒识训练集后，需要根据拒识训练集对候选命名实体进行打分。图 10 示出了对候选命名实体进行打分的流程图。下面参考图 10 描述对候选命名实体的打分过程。

[0142] 首先，输入候选命名实体，候选命名实体的位置以及有关的阈值。接下来，在步骤 S1001，对候选命名实体进行全局特征抽取。此后，在步骤 S1002，根据前面结合图 9 描述的过程得到的拒识分类模型，对该候选命名实体进行可靠性打分。如前所述，拒识得分＝可靠性得分－阈值。此后，在步骤 S1003，判断拒识得分是否大于 0。如果判断拒识得分大于 0，流程则进行到步骤 S1004，将该候选命名实体、其位置以及其拒识得分信息加入到候选命名实体网格，并输出更新的候选网格。如果在步骤 S1003 判断拒识得分不大于 0，则直接输出更新的候选网格。

[0143] 打分过程中的特征抽取模块与“拒识模型训练过程”中的特征抽取模块是一致的。

[0144] 可靠性打分的预测公式可以参见前面的描述。

[0145] 接下来，参考图 11 描述候选网格的生成过程。首先，输入经过前向解析和后向解析得到的所有候选命名实体。在步骤 S1101，判断经过前、后向解析得到的候选是否是前后向相同候选。如果是前后向相同候选，则在步骤 S1102 使用前后向相同候选阈值 th_{ident} 调用拒识打分流程，并向拒识打分流程提供候选命名实体、位置及相同候选阈值 th_{ident} 信息，以便执行拒识打分流程。此后，流程进行到步骤 S1103。需要说明的是，如果在步骤 S1101 的判断结果为否定，处理流程也转到步骤 S1103，判断经过前、后向解析得到的候选是否是自由候选。如果是自由候选，则在步骤 S1104 使用自由阈值 th_{free} 调用拒识打分流程，并向拒识打分流程提供候选命名实体、位置及自由阈值 th_{free} 信息，以便执行拒识打分流程。此后，流程进行到步骤 S1105。需要说明的是，如果在步骤 S1103 的判断结果为否定，处理流程也转到步骤 S1105，判断经过前、后向解析得到的候选是否是前后向重叠候选。如果在步骤 S1105 判断是前后向重叠候选，则在步骤 S1106 计入新边界候选，并使用前后向重叠候选阈值 $th_{difbndry}$ 调用拒识打分流程，并向拒识打分流程提供候选命名实体、位置及前后向重叠候选阈值 $th_{difbndry}$ 信息，以便执行拒识打分流程。此后，流程进行到步骤 S1107。需要说明的是，如果在步骤 S1105 的判断结果为否定，处理流程也转到步骤 S1107，判断经过前、后向解

析得到的侯选是否是前后向冲突侯选。如果在步骤 S1107 判断是前后向冲突侯选，则在步骤 S1108 使用前后向冲突侯选阈值 th_{cnflct} 调用拒识打分流程，并向拒识打分流程提供侯选命名实体、位置及前后向冲突侯选阈值 th_{cnflct} 信息，以便执行拒识打分流程。此后，流程进行到步骤 S1109。需要说明的是，如果在步骤 S1107 的判断结果为否定，处理流程也转到步骤 S1109。在步骤 S1109，判断是否还有未处理的侯选命名实体，如果判断结果表明还有未处理的侯选命名实体，流程则返回步骤 S1101，重复步骤 S1101 至 S1109。如果在步骤 S1109 判断已经处理了所有侯选命名实体，则输出侯选网格。

[0146] 得到侯选网格后，需要执行最优路径搜索过程。图 12 示出了最优路径搜索的示意图。最优路径搜索的核心算法是使用动态规划的方法在候选网格中搜索出一条累计得分最高的路径，其中每条路径上的节点在位置上不能重叠。输出是将该最优路径上的所有命名实体。

[0147] 下面描述有关的动态规划算法所执行的处理。

[0148] 1. 操作对象是由节点组成的网格，每个节点附带有得分信息，以及每个节点所处的开始和结束位置信息。如果节点 A 的结束位置小于节点 B 的开始位置，则称 A 是 B 的前驱节点，而 B 是 A 的后续节点。网格中有一个特殊的开始节点和一个特殊的终止节点，起始节点是所有其它节点的前驱节点，终止节点是所有其他节点的后续节点。起始和终止节点的分数都是 0。

[0149] 2. 初始状态为 :当前节点是开始节点，当前节点的累计分数设为 0，将该节点的来源指针设置为空。

[0150] 3. 在网格中寻找下一个开始位置最小，并且是当前节点后续节点的节点，并将其设置为当前节点。

[0151] 4. 针对当前节点，在网格中循环查找该当前节点的所有前驱节点。其处理过程还执行下列处理子过程。

[0152] 4.1. 对当前节点的任一前驱节点，创建一条临时路径，该临时路径的得分为该前驱节点的累计分数与当前节点的分数之和。

[0153] 4.2. 对所有这些临时路径的得分求其最大值，将最大临时路径的得分设置为当前节点的累计得分，当前节点的来源指针设置为该最大得分临时路径所对应的前驱节点。

[0154] 4.3. 删除所有临时路径。

[0155] 5. 如果网格中还有未处理的节点，则转到处理 3，否则转到处理 6。

[0156] 6. 从结束节点开始，使用每个节点的来源指针进行回溯，将该路径上的所有节点输出。

[0157] 根据本发明，识别系统需要逐步式分类模型和拒识分类模型两种模型，训练过程是可以是离线处理。图 13 示出了它们的离线训练的总流程。在训练开始后，首先，在步骤 S1301，以正向方式调用“逐步式解析模型训练流程”，通过前面所述的相应处理得到前向逐步式分类模型。此后，在步骤 S1302，以后向方式调用“逐步式解析模型训练流程”，通过前面所述的相应处理得到后向逐步式分类模型。接下来，在步骤 S1303，调用“拒识模型训练流程”，通过前面所述的相应处理得到拒识分类模型。得到相应的分类模型后结束训练。在调用各训练流程的过程中，系统向各个流程提供训练文本。

[0158] 得到逐步式分类模式和拒识分类模型这两种模型后，在线系统使用这两种模型对

输入的语句进行命名实体识别。图 14 示出了在线系统进行命名实体识别的总流程图。下面对该过程进行描述。

[0159] 在识别开始后，首先输入待测试的语句。然后，在步骤 S1401，以前向方式调用“逐步式解析识别流程”进行前向识别。在该过程中，根据前向逐步式分类模型对测试句进行识别，得到前向识别结果。此后，在步骤 S1402，以后向方式调用“逐步式解析识别流程”进行后向识别。在该过程中，根据后向逐步式分类模型对测试句进行识别，得到后向识别结果。在步骤 S1403，系统调用“侯选网格生成流程”以生成侯选网格。在该过程中，根据前、后向识别结果生成侯选网格。接下来，在步骤 S1404，系统调用“最优路径搜索流程”，根据生成的侯选网格来计算最优路径。最后，输出命名实体，该处理过程结束。

[0160] 接下来，描述根据本发明实施例的命名实体识别系统。根据本发明，进行命名实体识别可以包括进行离线训练的离线训练系统，以及进行在线测试及识别的在线识别系统。

[0161] 图 15 示出了根据本发明一个实施例的命名实体离线训练系统。如图 15 所示，本发明的命名实体离线训练系统包括：前向逐步式模型存储器 1501，后向逐步式模型存储器 1502，多类 SVM 训练器 1503，逐步式训练样本存储器 1504，逐步式训练引擎 1505，局部特征抽取器 1506，训练文本存储器 1507，拒识训练引擎 1508，全局特征抽取器 1509，拒识模型存储器 1510，单类 SVM 训练器，和拒识训练样本存储器 1512。

[0162] 下面描述命名实体离线训练系统的操作。逐步式训练样本存储器 1504 保存系统所使用的训练文本。逐步式训练引擎 1505 在需要下一句训练文本时，向训练文本存储器 1507 请求训练文本。逐步式训练引擎 1505 对每一个训练语句均触发局部特征抽取器 1506 的操作，并将该语句的训练文本传递给局部特征抽取器 1506。每当局部特征抽取器 1506 生成一个以特征向量和样本标记表示的训练样本时，将其传递给逐步式训练样本存储器 1504 存储。无论是前向解析还是后向解析，均使用同一个逐步式训练样本存储器 1504，这是因为前向训练和后向训练是顺序发生的。在训练文本的特征抽取操作的处理全部结束后，逐步式训练引擎 1505 触发多类 SVM 训练器 1503 操作。多类 SVM 训练器 1503 向逐步式训练样本存储器 1504 请求得到所有训练样本，进行训练。当多类 SVM 训练器 1503 生成前向逐步式分类模型时，将其传递给前向逐步式模型存储器 1501 并存储在其中。同样，当多类 SVM 训练器 1503 生成后向逐步式分类模型时，将其传递给后向逐步式模型存储器存储 1502，并存储在其中。

[0163] 拒识训练引擎 1508 在需要下一句训练文本时，向训练文本存储器 1507 请求得到该训练语句。拒识训练引擎 1508 对每一个语句的训练文本均触发全局特征抽取器 1509 的操作，并将该语句的训练文本传递给全局特征抽取器 1509。每当全局特征抽取器 1509 生成一个以特征向量和样本标记表示的训练样本时，将其提供给拒识训练样本存储器 1512 并存储在其中。在训练文本的特征抽取工作全部技术之后，拒识训练引擎 1508 触发单类 SVM 训练器 1511 的操作。单类 SVM 训练器 1511 向拒识训练样本存储器 1512 请求得到所有训练样本，并进行训练。当单类 SVM 训练器 1511 生成拒识分类模型时，将其传递给拒识模型存储器 1510，并存储在其中。

[0164] 经过离线训练后，可以利用在线系统对输入的语句进行测试和识别。

[0165] 图 16 示出了根据本发明一个实施例的命名实体在线识别系统。如图 16 所示，本发明的命名实体在线识别系统包括：前向逐步式模型存储器 1601，后向逐步式模型存储器

1602，多类 SVM 识别器 1603，逐步式识别引擎 1604，局部特征抽取器 1605，最优路径搜索器 1606，拒识打分引擎 1607，全局特征抽取器 1608，拒识模型存储器 1609，和单类 SVM 识别器。

[0166] 下面描述命名实体在线识别系统的操作。逐步式识别引擎 1604 在对测试输入语句的一个样本进行识别后，触发局部特征抽取器 1605 的操作，并将该语句的测试文本提供给触发局部特征抽取器 1605。触发局部特征抽取器 1605 将逐步（分前向后向两种工作模式）抽取到的下一个样本的局部特征向量传回给逐步式识别引擎 1604。逐步式识别引擎 1604 在得到测试语句的下一个样本时，触发多类 SVM 识别器 1603 的操作，将该样本的特征向量递给多类 SVM 识别器 1603。在前向工作模式下，多类 SVM 识别器 1603 向前向逐步式模型存储器 1601 请求得到前向逐步式分类模型，对输入的样本进行识别，然后将识别结果传回给逐步式识别引擎 1604。多类支撑向量机识别装置，利用多类分类模型，对输入的局部特征向量进行测试得到其类别标记，属于同一种类别的命名实体的一串起始和继续标记形成一个候选命名实体。

[0167] 在得到一个样本的识别结果后，逐步式识别引擎 1604 再次触发局部特征抽取器 1605 的操作。此后，局部特征抽取器 1605 执行如前所述的操作。后向工作模式的操作过程与此相同。

[0168] 在得到所有通过前向解析和后向解析识别得到的命名实体后，逐步式识别引擎 1604 将这些结果传递给拒识打分引擎 1607。对于前向解析和后向解析结果的每一个候选命名实体，拒识打分引擎 1607 触发全局特征抽取器 1608 的操作，并将候选的上下文传递给全局特征抽取器 1608。全局特征抽取器 1608 将抽取到的全局特征向量传回给拒识打分引擎 1607。在得到候选命名实体的特征向量时，拒识打分引擎 1607 触发对单类 SVM 识别器 1610 的操作，将该候选的特征向量传递给单类 SVM 识别器 1610。单类 SVM 识别器 1610 向拒识模型存储器 1609 请求得到拒识分类模型，并对输入的候选命名实体进行识别，将识别结果（可靠性得分）传回给拒识打分引擎 1607。单类支撑向量机识别装置，利用单类分类模型，对输入的全局特征向量进行测试得到其测试得分，减去不同的阈值得到拒识得分，根据拒识得分进行最优路径搜索，接受最优路径上的候选命名实体。

[0169] 在得到候选的识别结果（可靠性得分）后，拒识打分引擎 1607 根据前后和后向解析结果之间的位置关系，确定不同的阈值，从可靠性得分中减去该阈值，得到拒识得分，并触发对最优路径搜索器 1606 的调用。此后，拒识打分引擎 1607 将该候选及其位置和拒识得分传递给最优路径搜索器 1606。在得到一个候选及其位置和拒识得分时，如果得分大于 0，最优路径搜索器 1606 将该候选及其位置和拒识得分加入到候选网格中。根据计算出的拒识得分来接受或拒绝侯选命名实体。

[0170] 在一个输入句的所有候选均得到拒识得分的计算后，最优路径搜索器 1606 开始进行最优路径的搜索工作，搜索拒识得分之和最大的最优路径，根据并将最优路径上的命名实体作为系统的最后输出进行保存。

[0171] 本发明的命名实体离线训练系统和在线识别系统可以用计算机实现。如果用计算机实现，那么实现前向逐步式模型存储器 1501，后向逐步式模型存储器 1502，多类 SVM 训练器 1503，逐步式训练样本存储器 1504，逐步式训练引擎 1505，局部特征抽取器 1506，训练文本存储器 1507，拒识训练引擎 1508，全局特征抽取器 1509，拒识模型存储器 1510，单类 SVM 训练器，和拒识训练样本存储器 1512，以及前向逐步式模型存储器 1601，后向逐步式模型

存储器 1602，多类 SVM 识别器 1603，逐步式识别引擎 1604，局部特征抽取器 1605，最优路径搜索器 1606，拒识打分引擎 1607，全局特征抽取器 1608，拒识模型存储器 1609，和单类 SVM 识别器的程序保存在盘、半导体存储器、或其它记录介质上。计算机读取该程序，并且通过控制计算机的操作，在计算机上实现上述装置。

[0172] 根据本发明的识别自然语言中的命名实体的方法和系统，使用全局特征的命名实体识别方法。可以拒识逐步式命名实体识别方法产生的不可靠候选命名实体（具有不可靠起始边界或不可能结束边界）。另外，基于字的特征抽取避免了分词错误带来的影响。通过结合前向和后向两遍解析结果，使命名实体的识别性能得到提高。

[0173] 至此已经结合优选实施例对本发明进行了描述。应该理解，本领域技术人员在不脱离本发明的精神和范围的情况下，可以进行各种其它的改变、替换和添加。因此，本发明的范围不局限于上述特定实施例，而应由所附权利要求所限定。

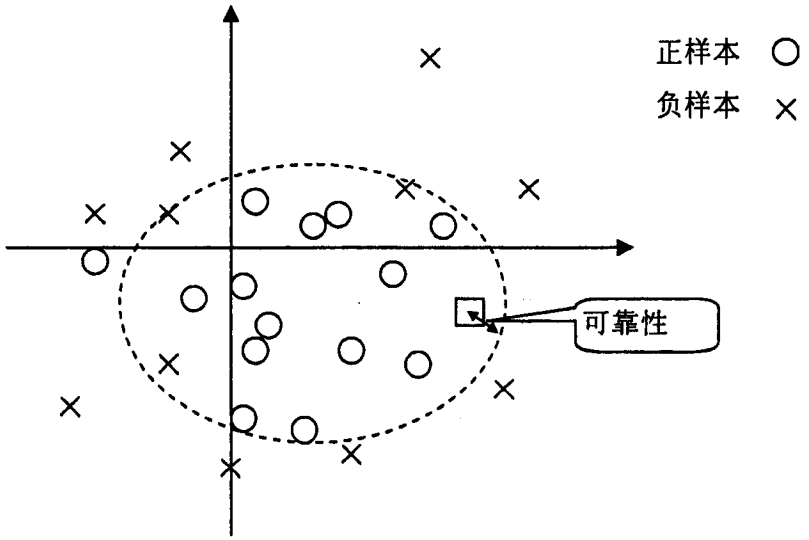

正样本
负样本
可靠性

图 1

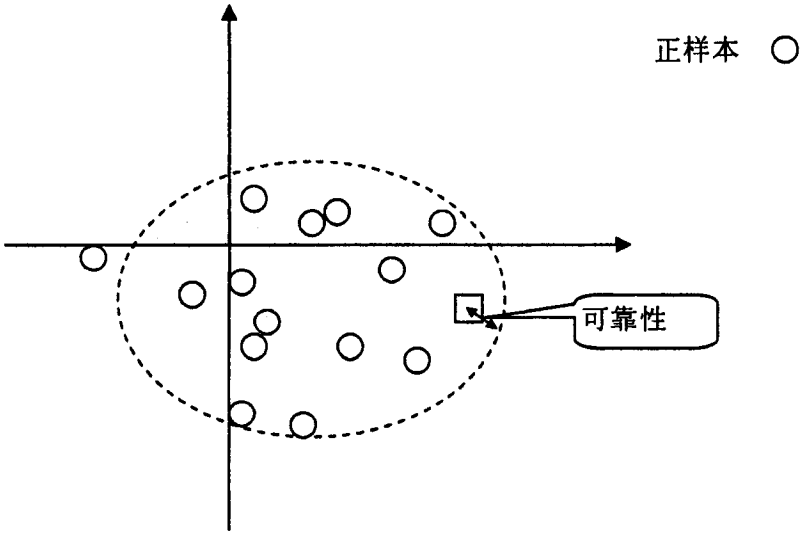

正样本
可靠性

图 2

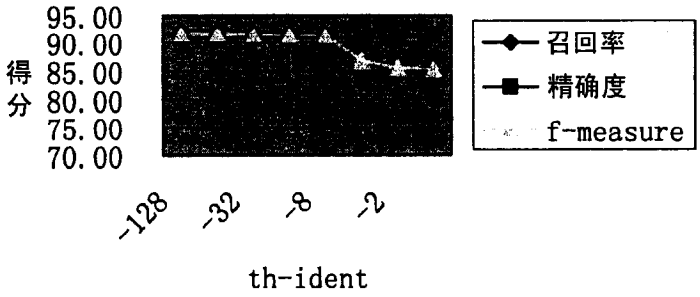

95.00
90.00
得 85.00
分 80.00
75.00
70.00
召回率
精确度
f-measure
-128
-32
-8
-2
th-ident

图 3

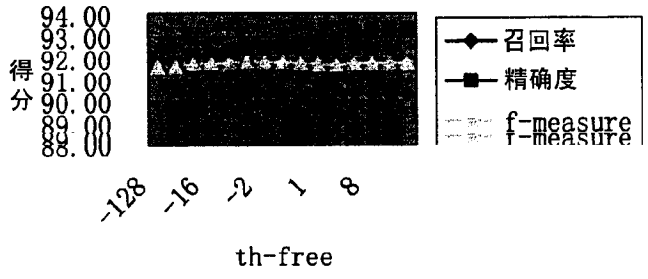

94.00
93.00
92.00
得分 91.00
90.00
89.00
88.00
召回率
精确度
f-measure
-128
-16
-2
1
8
th-free

图 4

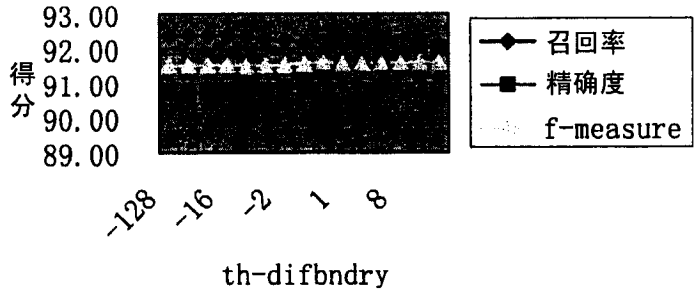

93.00
92.00
得分 91.00
90.00
89.00
召回率
精确度
f-measure
-128
-16
-2
1
8
th-difbndry

图 5

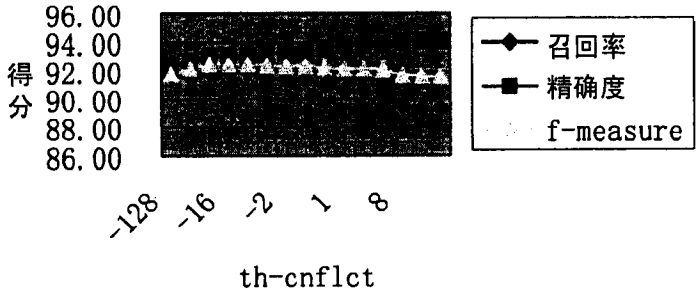

96.00
94.00
得分 92.00
90.00
88.00
86.00
召回率
精确度
f-measure
-128
-16
-2
1
8
th-cnflct

图 6

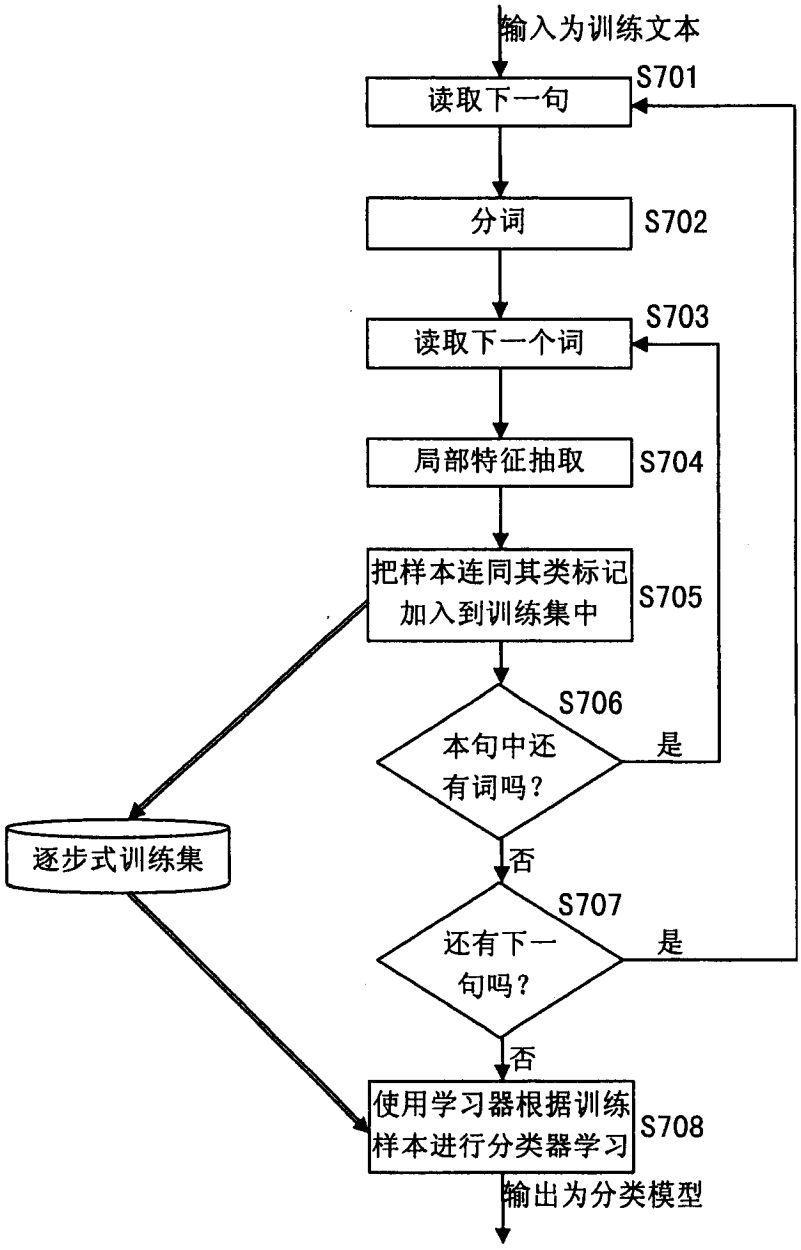

输入为训练文本
读取下一句
S701
分词
S702
读取下一个词
S703
局部特征抽取
S704
把样本连同其类标记
加入到训练集中
S705
S706
本句中还
有词吗？
是
否
逐步式训练集
S707
还有下一
句吗？
是
否
使用学习器根据训练
样本进行分类器学习
S708
输出为分类模型

图 7

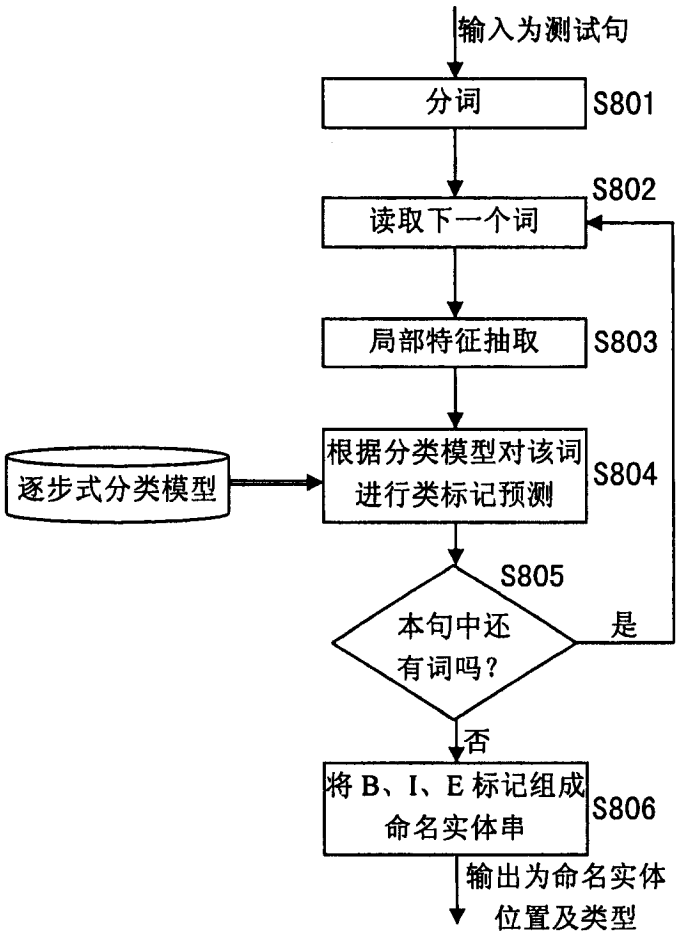

输入为测试句
分词
S801
S802
读取下一个词
局部特征抽取
S803
逐步式分类模型
根据分类模型对该词进行类标记预测
S804
S805
本句中还有词吗？
是
否
将 B、I、E 标记组成命名实体串
S806
输出为命名实体位置及类型

图 8

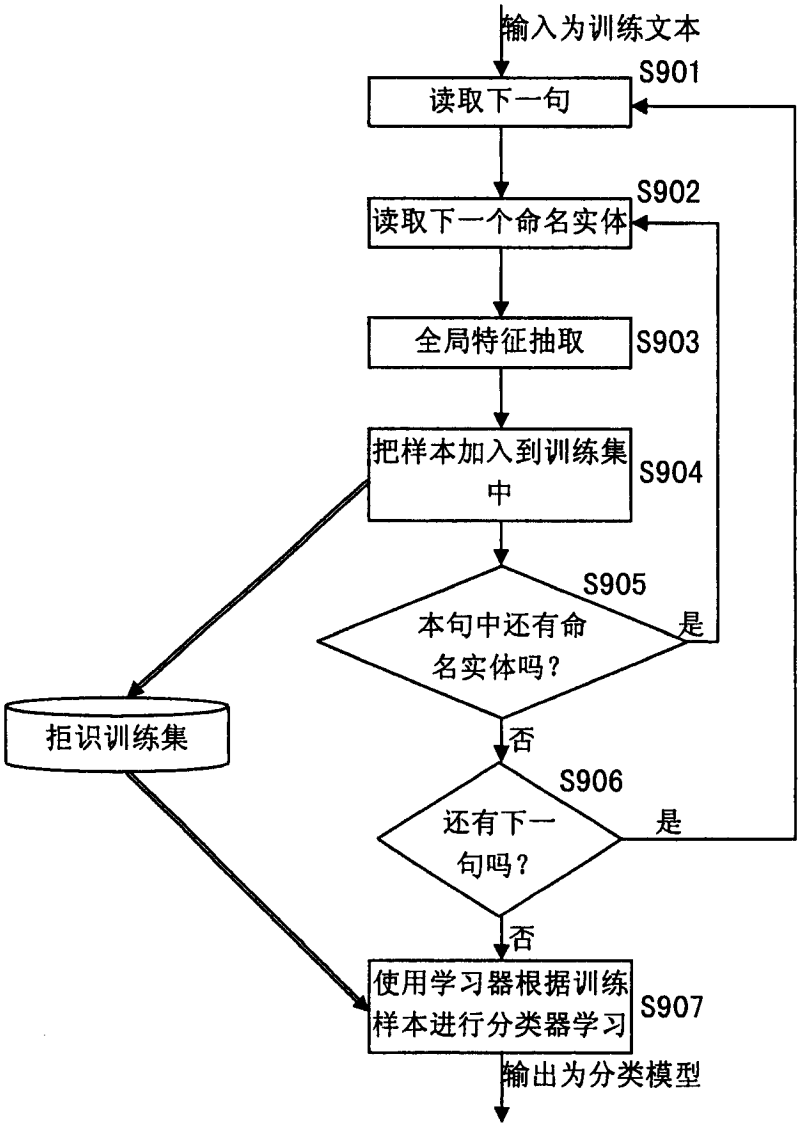

输入为训练文本
S901
读取下一句
S902
读取下一个命名实体
S903
全局特征抽取
S904
把样本加入到训练集中
S905
本句中还有命名实体吗？
是
否
拒识训练集
S906
还有下一句吗？
是
否
S907
使用学习器根据训练样本进行分类器学习
输出为分类模型

图 9

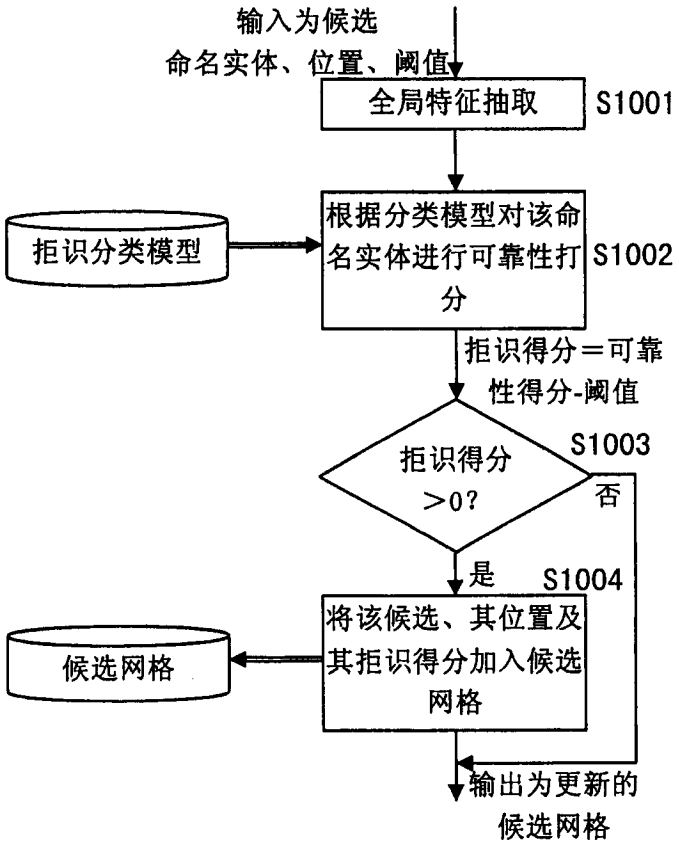

输入为候选
命名实体、位置、阈值
全局特征抽取
S1001
拒识分类模型
根据分类模型对该命名实体进行可靠性打分
S1002
拒识得分＝可靠性得分-阈值
拒识得分
>0?
S1003
否
是
S1004
候选网格
将该候选、其位置及其拒识得分加入候选网格
输出为更新的候选网格

图 10

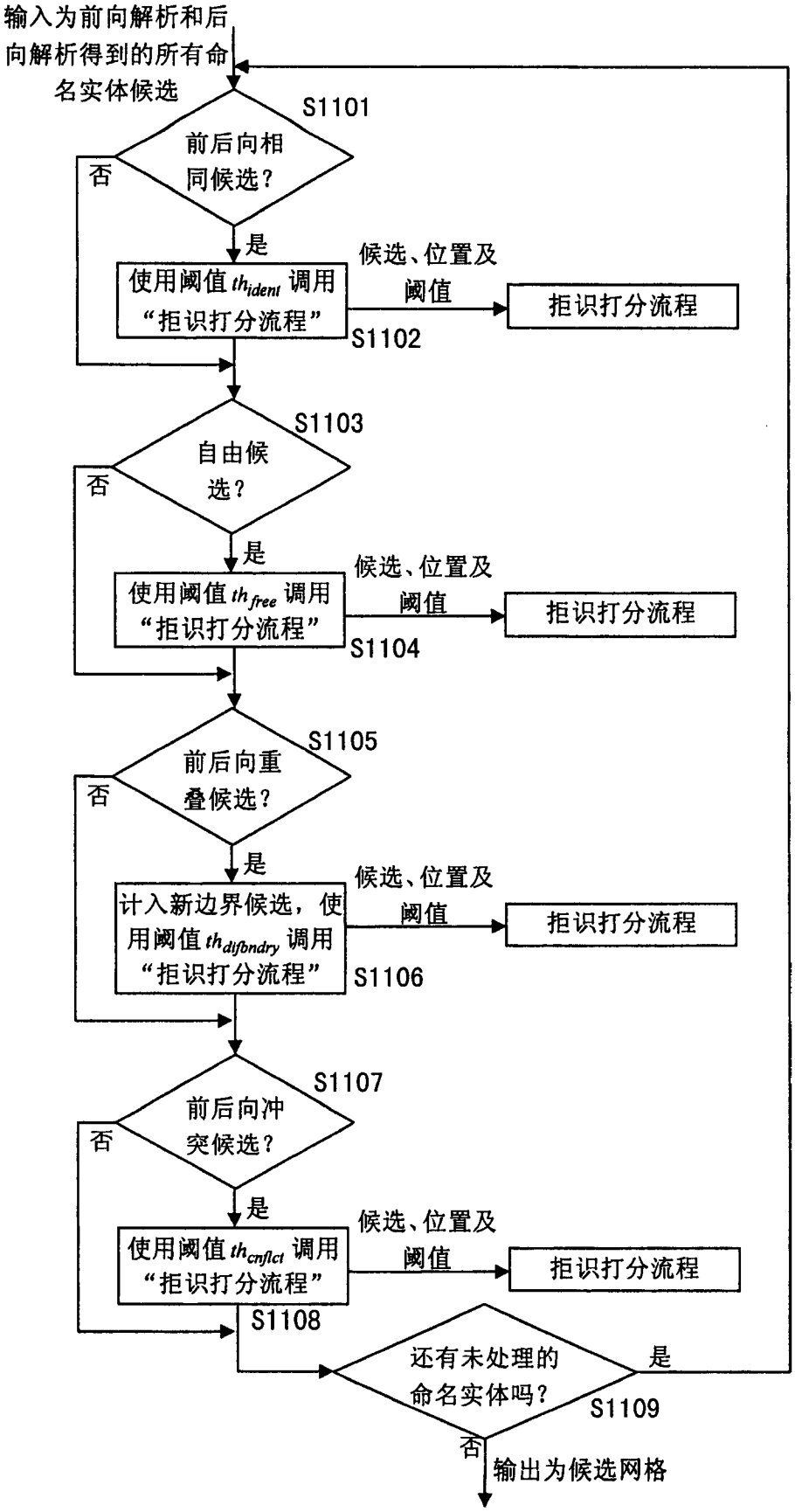

输入为前向解析和后向解析得到的所有命名实体候选
S1101
前后向相同候选？
否
是
使用阈值 th_{ident} 调用“拒识打分流程”
候选、位置及阈值
拒识打分流程
S1102
S1103
自由候选？
否
是
使用阈值 th_{free} 调用“拒识打分流程”
候选、位置及阈值
拒识打分流程
S1104
S1105
前后向重叠候选？
否
是
计入新边界候选，使用阈值 $th_{difbndry}$ 调用“拒识打分流程”
候选、位置及阈值
拒识打分流程
S1106
S1107
前后向冲突候选？
否
是
使用阈值 th_{cnflct} 调用“拒识打分流程”
候选、位置及阈值
拒识打分流程
S1108
还有未处理的命名实体吗？
是
S1109
否
输出为候选网格

图 11

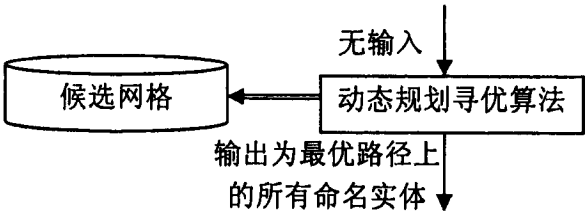

无输入
候选网格
动态规划寻优算法
输出为最优路径上
的所有命名实体

图 12

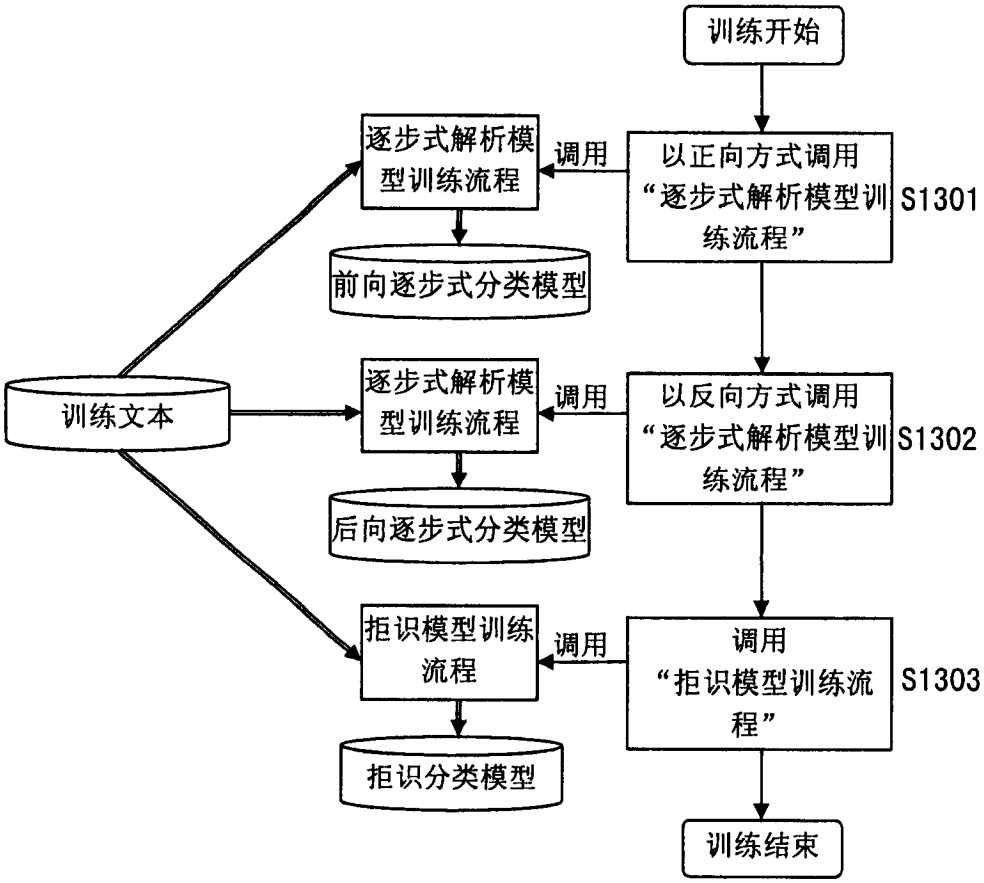

训练开始
逐步式解析模型训练流程
调用
以正向方式调用“逐步式解析模型训练流程”
S1301
前向逐步式分类模型
训练文本
逐步式解析模型训练流程
调用
以反向方式调用“逐步式解析模型训练流程”
S1302
后向逐步式分类模型
拒识模型训练流程
调用
调用“拒识模型训练流程”
S1303
拒识分类模型
训练结束

图 13

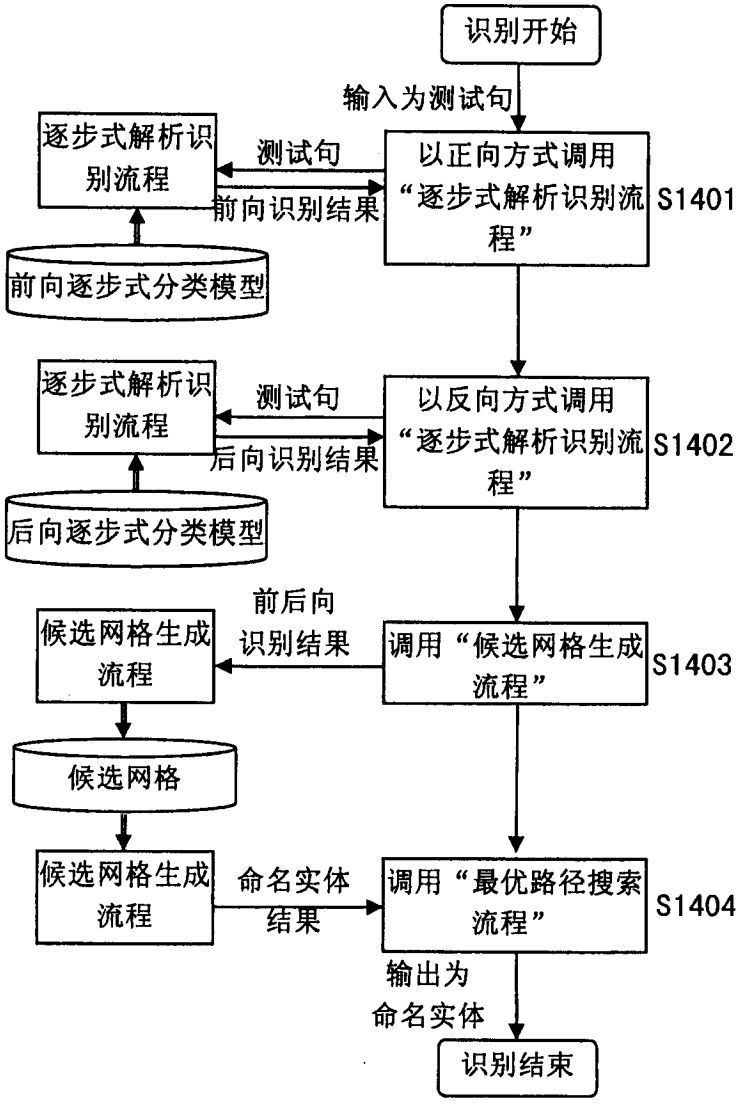

识别开始
输入为测试句
逐步式解析识别流程
测试句
前向识别结果
以正向方式调用“逐步式解析识别流程”
S1401
前向逐步式分类模型
逐步式解析识别流程
测试句
后向识别结果
以反向方式调用“逐步式解析识别流程”
S1402
后向逐步式分类模型
前后向识别结果
候选网格生成流程
调用“候选网格生成流程”
S1403
候选网格
候选网格生成流程
命名实体结果
调用“最优路径搜索流程”
S1404
输出为命名实体
识别结束

图 14

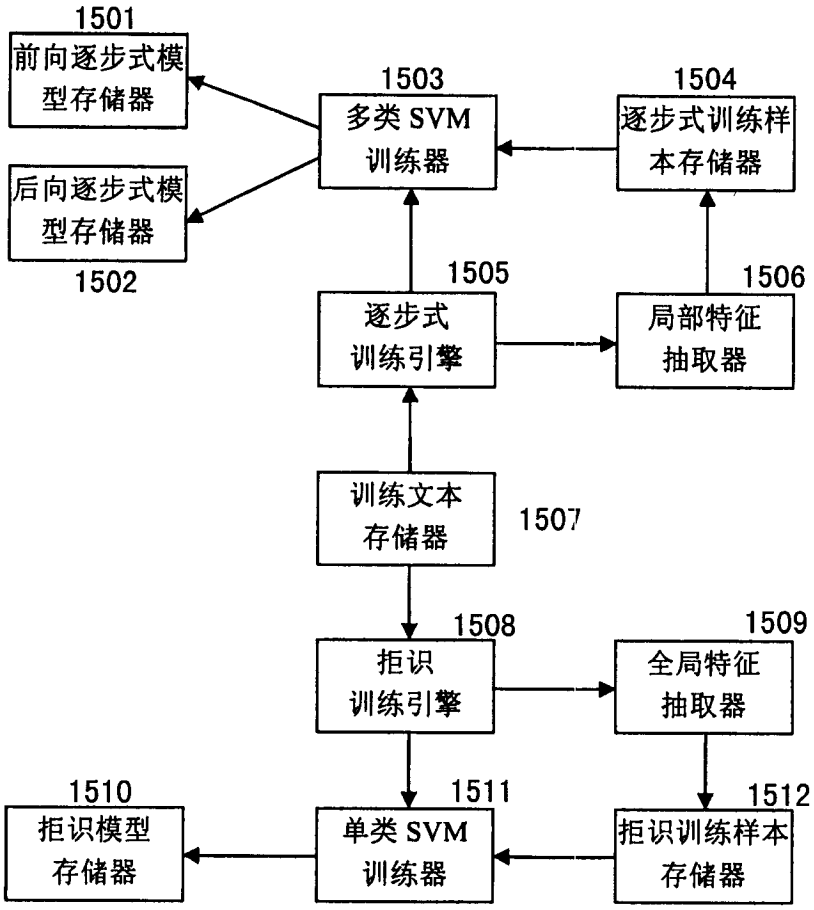

1501
前向逐步式模型存储器
后向逐步式模型存储器
1502
1503
多类 SVM 训练器
1504
逐步式训练样本存储器
1505
逐步式训练引擎
1506
局部特征抽取器
训练文本存储器
1507
1508
拒识训练引擎
1509
全局特征抽取器
1510
拒识模型存储器
1511
单类 SVM 训练器
1512
拒识训练样本存储器

图 15

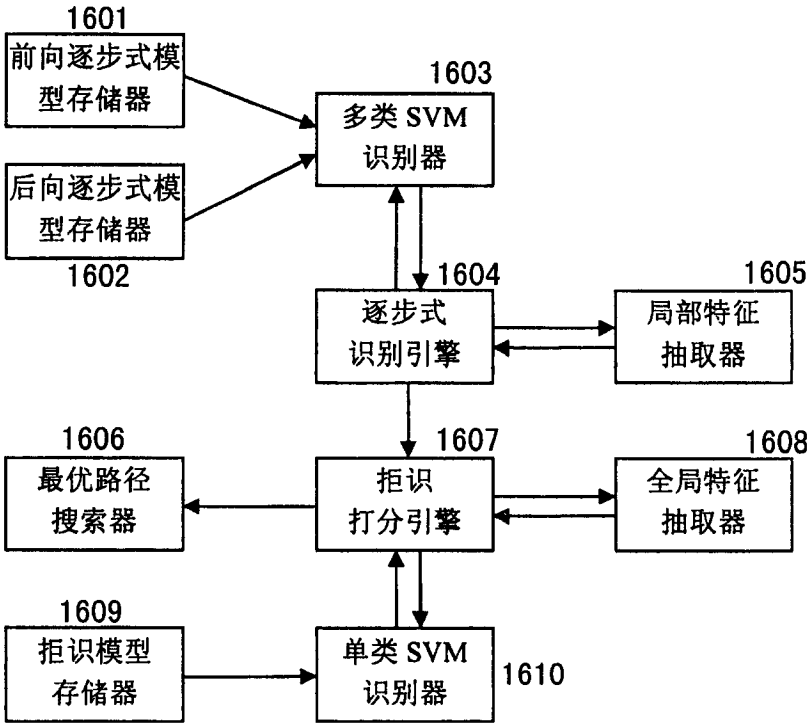

1601
前向逐步式模型存储器
1603
多类 SVM 识别器
后向逐步式模型存储器
1602
1604
逐步式识别引擎
1605
局部特征抽取器
1606
最优路径搜索器
1607
拒识打分引擎
1608
全局特征抽取器
1609
拒识模型存储器
单类 SVM 识别器
1610

图 16