CN43-1258/TP
ISSN 1007-130X
计算机工程与科学
COMPUTER ENGINEERING & SCIENCE
2007 年第 29 卷第 7 期
Vol 29, No 7, 2007

文章编号: 1007-130X(2007)07-014F-04

一种基于框架结构的专有名词自动识别方法*

A Framework-Based Approach for the Automatic Identification of Proper Nouns

王 蕾[1,2], 李培峰[1], 朱巧明[1], 杨季文[1]

WANG Lei[1,2], LI Pei-Feng[1], ZHU Qiao-Ming[1], YANG Ji-Wen[1]

(1 苏州大学计算机科学与技术学院,江苏 苏州 215006;2 苏州卫生职业技术学院信息中心,江苏 苏州 215009)

(1 School of Computer Science and Technology, Suzhou University, Suzhou 215006;
2 Information Center, Suzhou Health College of Technology, Suzhou 215009, China)

摘 要: 本文提出了一种基于框架结构的专有名词统一识别方法。该方法首先根据专有名词的成词特点及出现的上下文环境,重新定义语料属性; 然后,提出了属性标注点(AP)的概念,对训练语料进行初次标注,并采用错误驱动的学习方法来获取规则; 最后,结合规则和实例对文本进行专名识别。实验表明,该方法在测试样本集上准确率最高可以达到92.3%,召回率最高可以达到80.4%,是一种有效的专有名词识别方法。

Abstract: In this paper, a method based upon the framework structure is proposed to identify proper nouns. First, the properties of corpus are defined according to the characteristics of proper nouns and their contexts. Then the concept of attribute point is put forward, and rules are collected through an error-driven learning algorithm after labeling train corpus for the first time. Finally, rules and instances are assembled together to identify proper nouns in the texts. The results of experiments show that the precision and the recall rate are up to 92.3% and 80.4% respectively, which illuminate that our method is effective in identifying proper nouns.

关键词: 专有名词识别; 框架结构; 属性标注; 错误驱动; 规则和实例

Key words: proper noun recognition; framework structure; attribute tagging; error-driven learning; rule and instance

中图分类号: TP391.1 文献标识码: A

1 引言

在大规模中文文本词法分析的自然语言处理中,未登录词的识别一直是自动分词过程中的一个难点,这些词一般包括各类专有名词(人名、地名、机构名等),某些术语、缩略语和新词等,而各类专有名词在未登录词中占有较大比重,也是未登录词识别的主要难点。

据人民日报1998年1~5月份的语料(共计切分出6 060 606词次)统计,仅专有名词就出现了369 980词次,达到了总词数的6%左右。因此,对各类专有名词进行统一识别对提高汉语自动分词和词法分析的准确性都有很重要的意义。

在现有的专有名词识别方法中,多数采用的是基于统计或者统计结合规则(以统计为主)的方法。文献[1]使用从大规模地名词典和真实文本语料库得到的统计信息以及针对地名特点总结出来的规律,通过计算地名的构词可信度和接续可信度从而识别中文地名。文献[2]提出了句子切分结果可信度等概念建立了统计模型,识别中文姓名。文献[3]使用从中文姓名进行抽样综合统计的结果,建立了姓氏频率表与名字用字频率表,又根据从大量统计数据中总结的规律信息建立了称谓表和简单上下文的匹配模式,是基于统计和规则的一种辨识方法。对于中文机构名的识别,目前来看,研究得还是较少。

本文采用了一种经改造的基于转换的标注方法,其实质是一种基于实例和转换相结合的方法。另外,本文在基于转换的初始属性标注中,提出了属性标注点(AP)的概念,从而在转换的初始属性标注中采用了实例的方法进行标注。因此,本文把这种基于实例和转换结合的识别方法称为基于框架结构识别方法,同时将这种方法应用于人名、

* 收稿日期: 2005-12-29; 修订日期: 2006-05-10
作者简介: 王蕾(1980-),女,河南开封人,硕士,研究方向为中文信息处理; 李培峰,副教授,研究方向为中文信息处理; 朱巧明,教授,研究方向为中文信息处理技术、嵌入式系统及应用; 杨季文,教授,研究方向为中文信息处理。
通讯地址: 215009 江苏省苏州市苏州卫生职业技术学院信息中心; Tel: (0512) 62690162, 13771916712; E-mail: leiwang@zhhz.org
Address: Information Center, Suzhou Health College of Technology, Suzhou, Jiangsu 215009, P.R.China

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

地名、机构名统一识别工作上。

实验证明, 该方法在测试样本集上准确率达到95 3%, 召回率达到92 5%, 是一种有效的专有名词识别方法。

2 问题的描述及方法知识介绍

在日常见到的语料中, 尤其是新闻语料中, 人名、地名和机构名的出现总是会伴随着它的上文或下文同现的特征, 特别是对于首次提到的不常见人名、地名或者机构名, 例如,“克林顿对内斯塔尼亚胡说”, 如果不是从人名的上下文环境出发, 很难识别出“内斯塔尼亚胡”是国外人名。因此, 本文在讨论专有名词(指包括人名、地名、机构等)统一识别时, 充分考虑专有名词的上文或下文和专有名词特征词之间的关系等这些框架结构, 力图运用浅层的句法结构分析来识别出专有名词。

本文从专有名词角度出发, 重新定义语料中的标注成分, 具体定义如下:

定义**1** 专有名词特征词(简称C)是指包括人的姓氏, 地名指示词和机构称呼词等可以反应未登录词类别特征的名词。其中单字特征词(如省、市、街等)和多字特征词(如酒店、电视台等)。

定义**2** 填充词(简称O)是指与专有名词特征词结合组词专业名词的词汇。

定义**3** 专有名词的上文(简称U)、下文(简称D)是从语料中提取的伴随专有名词左右的一对词语。例如,“李鹏在北京考察企业。”, 其中“在……考察”即被认为是专有名词“北京”的一对上下词语。

定义**4** 连接词(简称L)是指连接两个专有名词的词语。例如,“梁山伯与祝英台”, 其中“与”即被认为是连接两个人名的连词。

定义**5** 定义1~4定义的各个成分通称为专有名词相关成分。

定义**6** 其它成分(简称E)指专有名词相关成分以外的词。

定义**7** 专有名词(简称PN)指人名、地名、机构以及它们各自简称在内的统称。

定义**8** 基本专有名词串(简称BasePNS)指由专有名词的上文(U)、下文(D)、连接词(L)以及专有名词特征词(C)和一些填充词(O)组成的词串。

定义**9** 由专有名词上下文组成的词对、上文与专有名词特征词词对、专有名词特征词与下文词对、专有名词特征词与连词词对都认为是框架结构。例如, 专有名词上下文词语“在……考察”, 就认为是一个框架结构。

表1给出了PN的基本组成形式, 它根据专有名词的组成形式将专有名词分成前缀型专有名词和后缀型专业名词两种。

表2和表3给出来BasePNS的基本组成形式。根据BasePNS中专有名词的类型可以把BasePNS分成前缀型基本专有名词串(简称Q_BasePNS)和后缀型基本专有名词串(简称H_BasePNS), 而此两种名词串可近一步细分, 具体细分类型见表2和表3。

表1 **Q_PN** 和 **H_PN** 例句

类型	组成形式	例句
前缀型专有名词(Q_PN)	<O><O><C>	中央人民广播电台, 老王
后缀型专有名词(H_PN)	<C><O><O>	刘德华, 何教授

表2 **Q_BasePNS** 例句

类型	组成形式	例句
由上下文构成的	<U><Q_PN><D>	在/北京大学/上学
仅由上文构成的	<U><Q_PN>	出访/美国
仅由下文构成的	<Q_PN><D>	国务院侨办/发表
连接两个Q_PN形式的	<Q_PN><L><Q_PN>	中国/驻/美国/大使

表3 **H_BasePNS** 例句

类型	组成形式	例句
由上下文构成的	<U><H_PN><D>	记者/樊如钧/摄
仅由上文构成的	<U><H_PN>	主席/江泽民
仅由下文构成的	<H_PN><D>	李鹏/说
连接两个H_PN形式的	<H_PN><L><H_PN>	梁山伯/与/祝英台

3 专有名词框架结构的规则挖掘

从中文信息处理的角度出发, 要对经初步切分的文本进行专有名词识别, 首先就是要构造一个BasePNS识别模型, 然后对切分好的词串 $S = C1/C2/\cdots Cn$, 进行BasePNS句型识别, 最后再从识别出来的BasePNS提取专有名词。本文采用一种改造的基于转换的BasePNS识别模型, 该方法是将BasePNS所在上下文环境中的分布特征用一组上下文有关规则表示, 获取识别BasePNS的上下文有关规则采用的是错误驱动的学习方法[4~6]。

3.1 构造基于转换的 **BasePNS** 识别模型

(1) 根据训练语料库构造BasePNS初始标注器。

本文把人民日报1998年1~5月份人工订正过的语料库根据新定义的标注成分重新标注, 并从中抽取BasePNS句型。然后找出各BasePNS结构中的专有名词相关成分并统计实例次数。对所有提取出来的实例(未做实例优化和归纳)情况统计如表4。

表4 语料实例统计

BasePNS类型	专有名词相关成分	提取的实例个数
<U><PN><D>	<U><D>词对	57 545
<U><PN>	<U>	5 401
<PN><D>	<D>	12 775
<PN><L><PN>	<L>	846
<PN>	<C>	4 742

(2) 对训练语料库进行BasePNS初始标注。

定义**10** 文本中可以进行属性标注的位置称为属性标注点(简称AP): Eric Brill的初始属性标注[1,2]是分别为每个词进行单独地属性标注, 每个词都可以称作一个合法的 AP, 是个一维的点; 而本文的初始属性标注要对多个点同时进行标注, 而且并不是任意多个点都能组成一个合法的 AP, 而是那些点上的词组合起来能满足某种特征才能是一个合法的 AP, 它的 AP 是多维的点, $AP = (P_1, P_2, P_3) | (P_1, P_2)$, 其中 $P_n (0 < n < 4)$ 表示词在一个句子中的位置, 句子首词的 P_n 值可记为0, 第二个词的 P_n 值为1, 依次类推。

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

例如, U+ Q_PN+ D 的(P_1, P_2, P_3)是: P_1 与 P_3 位置上的词是 Q_BasePNS 中 U、D 词对中的用词, P_2 位置上的词是 Q_PN 里 C 的用词, 并且满足 $P_3 - P_2 = 1$, $P_2 > P_1$; Q_PN+ D 的(P_1, P_2)是: P_1 位置上的词是 Q_PN 里 C 的用词, P_2 位置上的词是 Q_BasePNS 中 D 的用词, 并且满足 $P_2 - P_1 = 1$。

定义 **11** 任意两个 $AP = (P_1, P_2, P_3) | (P_1, P_2)$, $AP' = (P_1', P_2', P_3') | (P_1', P_2')$。

若 $P_1 > P_1'$ 或($P_1 = P_1', P_2 > P_2'$) 或($P_1 = P_1', P_2 = P_2', P_3 > P_3'$), 则 $AP > AP'$。

若 $P_1 = P_1', P_2 = P_2', P_3 = P_3'$, 则 $AP = AP'$。

若 $P_1 < P_1'$ 或($P_1 = P_1', P_2 < P_2'$)或($P_1 = P_1', P_2 = P_2', P_3 < P_3'$), 则 $AP < AP'$。

定义 **12** 当前 AP 上 BasePNS 使用频度是指当前 AP 上 BasePNS 句型中各个专有名词相关成分在语料中出现的次数总和。

把训练语料库中语料标记剔除后, 提取训练语料库文本中所有合法 AP, 并按从小到大的顺序进行标注。在每个 AP 上, 标注为在当前 AP 上 BasePNS 中使用频度最高的形式。例如, S = "重庆市委书记张德邻说", 语料库中切分好的词串 S = "重庆/ 市委/ 书记/ 张/ 德\邻/ 说"。经提取后再初始标注, 所有合法 AP 从小到大的顺序是: $AP = (1, 2)$, 重庆/O 市委/C 书记/D; $AP = (2, 6)$, 书记/U 张/C 德/O 邻/O 说/D; $AP = (3, 6)$, 张/C 德/O 邻/O 说/D。

(3) 采用错误驱动的学习方法获取 BasePNS 的上下文有关规则。

规则的获取流程如图 1 所示。

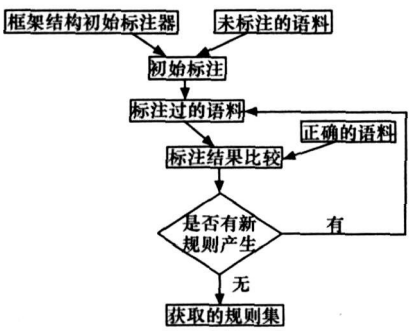

图 1 规则获取流程图

本文采用的规则体系参照了文献[7]中定义的(与 Eric Brill 的规则体系有所不同)。本文定义一个规则体系中包含多个规则块, 每个规则块内又包含一条或多条规则, 每个规则块都是针对 BasePNS 中一个具体用词的规则的集合, 并以这个用词标识这个规则块。

本文中规定可以标识规则块的用词有: 以 Q_BasePNS 的 U、D 用词对, U 用词, L 用词, 以及 Q_PN 里 C 用词; H_BasePNS 的 U、D 用词对, U 用词, L 用词, 以及 H_PN 里 C 用词。

规则体系形式化描述如下:

规则集= 规则块; {规则块}

规则块= 规则条; {规则条}

规则条= 触发条件- > 转换动作

转换动作= 修改当前标注结果

触发条件= 测试项; {测试项}

获取上述规则体系的关键在于转换规则模板集和评价函数的定义。而每一条转换规则模板包括转换动作和触发条件两个要素。其中触发条件限定了上下文环境的若干特征, 转换动作根据这些特征更新标注结果。本文定义了如下两种转换动作:

定义 **13** 转换动作 1: 当前标注更改为无效标注; 转换动作 2: 当前规则更改为有效标注。

触发条件是从触发环境中产生的, 而本文的触发环境根据 BasePNS 组成形式有所不同, 因此触发环境定义如表 5 所示。

表 **5** 触发环境定义

BasePNS 类型	PS1	PS2	PS3	PS4
< U> < PN> < D>	U 的前一个词	U 的后一个词	D的前一个词	D的后一个词
< U> < Q_PN>	U 的前一个词	U 的后一个词	C的前一个词	C的后一个词
< U> < H_PN>	U 的前一个词	C的后一个词		
< Q_PN> < D>	C的前一个词	D的后一个词		
< H_PN> < D>	C的前一个词	C的后一个词	D的前一个词	D的后一个词
< Q_PN> < L> < Q_PN>	前一个 Q_PN 中 C的前一个词	L 的后一个词	后一个 Q_PN 中 C的前一个词	C的后一个词
< H_PN> < L> < H_PN>	前一个 H_PN 中 C的前一个词	C的后一个词	L 的前一个词	后一个 Q_PN 中 C的后一个词

本文的触发条件是直接从触发环境中抽取的, 没有其它的特别限制, 所以最多可以产生 15+ 15+ 15+ 3+ 3+ 15 + 15= 81 种触发条件。每一条转换规则模板包括转换动作和触发条件两个要素, 因此文中将有 81× 2= 162 条转换规则模板。

下面是一个 H_BasePNS 的 U、D 用词对为"对, 的"规则块中部分例子:

①PS2= 华; PS3= 政策 - > 当前标注更改为无效标注。

②PS2= 全 - > 当前规则更改为无效标注。

③PS2= 全; PS3= 哲洙- > 当前规则更改为有效标注。

当前句子为"对/华/政策/的/正确/方向", 其初始标注结果是"对/U 华/C 政策/O 的/D 正确/E 方向", 在初始标注中"对, 的"被认为 H_BasePNS 中的一个合法 U、D 的, "华"被认为是特征词, 然后依次使用"对, 的"规则块中的规则对初始标注进行修改: 由第一条规则, 把当前标注更改为无效标注。

定义 **14** 评价函数是在错误驱动的学习中用来判断规则选取好坏的准则, 本文定义的评价函数是 $F(r) = C(r) - E(r)$, 其中 F 为转换规则的评价函数, r 为规则, $C(r)$ 为应用规则 r 后错误标记改为正确标记的数目, $E(r)$ 为应用规则 r 后正确标记改为错误标记的数目。

3.2 进行 BasePNS 句型识别

文中的专有名词识别是为了提高自动分词的正确率, 是在第一趟分词的基础上进行识别。因此, 为了尽量减少第一趟分词的结果对以后专有名词识别的影响, 本文在构

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

造 BasePNS 初始标注器之前, 先剔除训练语料库中语料标记, 对语料中的文本进行分词, 把切分的结果和原语料库的结果进行对比, 找出分词过程中专有名词的边界词和专有名词上文或下文成词的句子, 也就是一些歧异实例。例如, "厂长对于民红说", 第一趟分词后的结果是: 厂长/对于/民/红/说, 而正确的结果是: 厂长/对/于/民/红/说。因此, 分词结果中的"对于"就是由于专有名词的边界词和专有名词上文结合成词的缘故, 因此在进行 BasePNS 句型识别过程中, 当出现"对于……说"这个实例时, 则转换成"对/于/……说"的切分方式进行识别。为了提供转换的正确率, 系统在识别这种转换时, 也采用了基于转换的方式。例如, 当前的句子如果是"对于他来说", 这时"对于……说"实例就不需要进行改变。

在对第一趟分词的结果进行校正后, 采用初始标注器, 对切分文本中所有合法的 AP 按从小到大的顺序进行标注。在每个 AP 上, 初始标注为在当前 AP 上 BasePNS 中使用频度最高的形式。然后依次利用所得的转换规则对初始标注结果进行修改, 直至转换规则集中的所有规则都已用过。

3.3 从 BasePNS 中提取专有名词

定义 **15** 任意两个 $AP=(P_1, P_2, P_3)|(P_1, P_2)$, $AP'=(P_1', P_2', P_3')|(P_1', P_2')$。

下面两种情况认为 AP 与 AP' 是相交的:

(1) 若 $P_1=P_1'$ 并且 $P_2|P_3<P_2'|P_3'$;

(2) 若 $P_1<P_1'$ 并且 $P_2|P_3\leqslant P_2'|P_3'$。

专有名词提取阶段的主要任务就是对识别出来的 BasePNS 句型进行确认, 找出存在相交的 AP, 若 AP 与 AP' 相交并且 $AP<AP'$, 那么删除 AP'。然后再根据 BasePNS 组成形式提取 Q_BasePNS 与 H_BasePNS 中的 Q_PN 和 H_PN, 即可认为是专有名词。

4 实验结果与分析

下面给出我们在实验过程中采用的语料和指标, 然后给出实验的一个初步结果及相应的分析。

4.1 实验用语料和评测标准

实验使用了由北京大学计算语言学研究所和富士通研究开发中心有限公司共同制作的 1998 年 1~6 月份人民日报标注语料库。其中 1~5 月份语料作为训练语料, 6 月份语料作为开放测试用语料。

针对专有名词的识别, 我们采用了两个评测指标, 即准确率(P)、召回率(R)。其定义如下:

准确率= 系统识别的正确词数/系统识别的总词数×100%

召回率= 系统识别的正确词数/总的正确词数×100%

4.2 实验说明及分析

根据需要, 我们主要进行了两组, 一组实验是证明提取实例的可行性, 一组实验是方法识别的效果分析。

(1) 实例提取。为了说明实例框架提取的可行性, 我们进行了逐次增加语料提取实例框架的方式(将 1998 年 1~5 月份语料, 大约共 600 万条分成 6 等份)。通过观察实例框架的增长幅度, 来验证提取典型性实例的可行性。具体结果参见表 6。

表 **6** 实例提取结果

提取用语料	提取实例总数	优化后实例数量(进行实例归纳后)	比前次增加的数量(进行实例归纳后)
100 万词条	21 416	4 265	
200 万词条	38 649	8 554	4 289
300 万词条	53 743	12 744	4 190
400 万词条	67 588	16 299	3 555
500 万词条	73 494	17 694	1 395
600 万词条	80 930	18 219	525

从上面的实例提取结果(表 6)可以看出, 最初实例提取时, 虽然实例增长的种类数量依次在变小, 但是总体增长幅度还是很大。当语料数量增加到 500 万词的时候, 实例种类数量的增长有了一个明显的降低, 增长幅度有了一个显著的变化。这说明随着语料数据的增加, 提取出的实例的典型性也就越强了。可以预见, 如果继续增加语料数量, 实例的增长会更小, 甚而只是少量的增加, 而认为实例种类不再变化。

在实例的提取中, 对实例结构的归纳抽取也是使得实例具有典型性的一个方法。例如, 机构名实例"进入……读"和"进入……读书", 有明显的框架相似性, 那么将这两个实例归纳为一个"进入……读*(即下文词首次为读)"框架, 就更加具有代表性。

所以说, 大量的语料提取实例+实例的归纳合并是一个有效获得典型性实例的方法。

(2) 实验结果。实验分封闭测试和开放测试两部分, 封闭测试语料和开放测试语料的规模都约为 25 000 字, 其中封闭测试语料来自训练集(一月份语料中的部分文本); 开放测试语料来源于训练集外(采用 6 月份语料中的部分文本)。测试结果如表 7 所示。

表 **7** 专有名词识别的测试结果 %

测试类型	正确率	召回率
封闭测试	98.7	95.6
开放测试	92.3	80.4

根据目前所进行的封闭测试的结果, 发现情况如下:

①框架结构提取方面: 某些结构的出现并没有严格的语法结构, 但是却频繁出现在语料中, 并且对识别专名起到很好的作用。而这类框架的提取是必须依赖大量语料, 并不是只用单纯的语言学可以得到的。

② 特征词的提取必要性: 在最初单纯使用框架结构没有考虑特征词识别专有名词的时候, 正确率是非常低的, 原因就是识别结果中单纯考虑框架却没有考虑框架中间成分而识别出了很多非专名内容。加入特征词之后, 也就是加入了框架识别结果的正确性验证, 正确率就有了明显的提高, 所以特征词的提取是非常重要的。但是, 特征词采用也有其局限性, 例如对识别中国专名有良好的作用, 对外国译名的识别所起的作用就较小(因为外国译名的特征词提取有难度, 并且没有一个规范), 而单纯的框架结构在识别外国译名上却有好的效果。

③规则提取方面: 目前的规则提取只是考虑了框架结构中的左右相邻的各一个词, 如果增加其考虑范围, 相信框

(下转第 154 页)

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

虚信道(VL),其中一条为管理虚信道,15 条为数据虚信道。所有虚信道共享同一物理链路,各个虚信道都有各自的收发缓冲区和各自独立的基于信用的流控机制,通过仲裁器来决定哪条虚信道使用物理链路。虚信道提高了网络利用率,提供了不同的服务级别,增强了链路的可用性,并有效防止死锁。

(6) 串并转换部件(SERDESS):完成数据的串并转换。发送部件完成发送数据包的并行数据到串行物理链路传输的串行数据转换;接收部件则相反,将收到的串行链路数据转换成并行数据。

(7) 收发物理链路(PHY):实现 IBA 串行物理链路协议,完成串行数据的收发。对于发送方,完成控制信息的插入和 8b/10b 编码;对于接收方,要丢弃控制信息、进行 8b/10b 解码及物理层传输错误的检测和处理。物理层采用串行链路、差分信号传输,单线传输速率为 2.5Gb/s,可通过 4 或 12 线并行来扩展通道带宽,带宽高达 2.5、10、30Gb/s(1×、4×、12× 线)。

HCA 作为 HOST 的 IBA 接口必须支持所有的通道功能,以满足不同应用的需求。从 IBA 系统角度,HCA 可分为两个层次:HCA 硬件层和 HCA 驱动层。HCA 必须支持 IBA 的 Verb,同时要支持 UD、RC、UC 服务,并支持虚实地址转换和 Memory 保护。

3.2 TCA 设计

TCA 是 I/O 设备和存储设备的 IBA 接口,由发送和接收两部分组成,其基本功能是通过队列管理器、报文管理、虚信道管理以及链路管理完成 IBA 的五层协议。收发部件一般采用嵌入式微处理器实现工作队列(WQ)的管理。TCA 操作完全由硬件实现,无需处理器干预。TCA 必须支持 UD 服务,其它服务都是可选的。TCA 作为 I/O 设备的 IBA 接口部件,必须根据 I/O 设备的类型和功能决定所要实现的 IBA 所规定的通道功能。TCA 的具体实现与 HCA 类似,这里就不详细说明。

4 结束语

采用 InfiniBand 之后,I/O 不再是服务器的组成部分,而可以看成机箱的一部分。这时,远程存储设备、I/O 设备和服务器之间的互连是通过 InfiniBand 交换机和路由器完成的,由此 InfiniBand 根本解决了传统 PCI 总线存在的距离问题。采用 InfiniBand 技术,使用铜线传输时外部设备可以放到距服务器 17 米远处;若使用 4×、12× 光缆,最远距离可达 300 米;若使用 1× 光缆,则可长达 10 千米。未来的诸多信息存储方案中,InfiniBand SAN 具有无与伦比的性能优势,其关键实现技术值得国内同行进一步深入研究。

参考文献:

[1] InfiniBand Trade Association. InfiniBand Architecture Specification Volum1[S]. 2000.

[2] InfiniBand Trade Association. InfiniBand Architecture Specification Volum2[S]. 2000.

[3] 李琼,汪审权,方粮,等. SAN 存储技术研究[J]. 计算机工程,2004,29(19):165-169.

[4] 郭御风,李琼,屈婉霞,等. 基于 InfiniBand 的磁盘阵列系统研究[J]. 计算机研究与发展,2002,39(增刊):168-172.

(上接第 144 页)

架的使用会更加准确,对识别准确率和召回率的提高有更大帮助。

(3) 方法评价。本方案由于充分考虑了上下文语言环境,所以对复合结构专有名词有较好的识别效果。开放测试中,包括地名、机构名在内的复合结构专有名词召回率达到了 75.9%。

4.3 后续工作

后续主要完成以下一些工作:近一步完善实例的提取和归纳,建立典型性的实例库;增大规则的考虑范围,找出合适的规则考虑阈值;考虑多种规则判断方法进行实验;考虑规则提取的数量对开放测试的影响。

5 结束语

本文首先分析现阶段专有名词识别存在的问题和局限性,从人自身在阅读时候区别专有名词和普通用词的特点,提出了基于框架结构的专有名词识别方法。此方法从专有名词自身特点(姓氏用词等)和上下文环境特点出发,重新定义语料属性,然后采用基于转换错误驱动和基于实例相结合的学习方法对文本进行标注,从而识别专有名词。避免了传统专有名词识别中人工统计的局限性。在小规模语料的封闭测试中,该方法取得了相当好的效果。实验表明,基于框架结构的专有名词识别方法是行之有效的。但是,此方法的有效运用,需建立在拥有大量熟语料库的基础上,只有足够的学习语料,才可能得到较完备的框架结构模式和规则集合。另外,正确合理的 BasePNS 句型成分确定,也关系到整个识别过程系统的准确度;规则的提取范围和合理性判断,也是识别效果一个不可缺少的关键因素。

参考文献:

[1] 黄德根,岳广玲,杨元生. 基于统计的中国地名识别[J]. 中文信息学报,2003,17(2):36-41.

[2] 孙茂松,黄昌宁,高海燕,等. 中文姓名的自动辨别[J]. 中文信息学报,1994,9(2):16-27.

[3] 黄德根,杨元生,王省,等. 基于统计方法的中文姓名识别[J]. 中文信息学报,2001,15(2):31-37.

[4] Brill E. Transformation-Based Error-Drive Learning and Natural Language Processing: A Case Study in Part of Speech Tagging[J]. Computational Linguistic,1995,21(4):543-565.

[5] Brill E. A Simple Rule-Based Part of Speech Tagger[A]. Proc of the 3rd Conf on Applied Natural Language Processing[C]. 1992.

[6] 孙宏林,俞士汶. 浅层句法分析方法概述[J]. 当代语言学,2000,2(2):74-78.

[7] 陈文亮,朱靖波,吕学强,等. 词性标注规则的获取和优化[J]. 术语标准话与信息技术,2004,(2):23-26.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net