第16卷 第11期
2006年11月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 16 No. 11
Nov. 2006

基于属性标记的专有名词自动识别研究

王 蕾,杨季文
(苏州大学 计算机科学与技术学院,江苏 苏州 215006)

摘 要:提出了一种新的基于属性标记的专有名词统一识别方法。其基本思想是:根据专有名词的成词特点,利用标注语料库,设定词语属性作为标准属性重新进行标注,在此语料基础上进行专有名词成词结构、成词环境的实例提取,并采用基于转换的错误驱动方法对提取的实例进行适用规则提取。在提取的实例和规则的基础上进行属性标注,是一种基于转换的错误驱动规则自学习方法与基于实例的学习方法相结合的基于浅层句法分析的一种新的识别专有名词的方法。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。

关键词:中文专有名词识别;未登录词识别;属性标注;基于转换的错误驱动学习方法
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2006)11-0195-04

Recognition of Chinese Proper Noun Based on Attribute Tag

WANG Lei, YANG Ji-wen
(School of Computer Science and Technology, Suzhou University, Suzhou 215006, China)

Abstract: Introduces a new method to identify the Chinese proper noun. It is based on attribute tag. The basic thinking is: according the characteristics about the Chinese proper noun compages, using label corpus, enact the words attribute to be the standard attribute and relabeled it. Based on the corpus, distilling the Chinese proper noun instances about compages configuration and compages environment, using the transformation-based error-drive learning method to distill the fit regulation. Doing attribute label based on the instance and regulation which just distilled is the method combined the transformation-based error-drive learning and instance-based learning. Experiments proved this method ratio of nicety achieved 95.3% on testing stylebooks, the ratio of recall achieved 92.5%, so it is an effective method to identify Chinese proper noun.

Key words: Chinese proper noun recognition; unknown words recognition; attribute tag; transformation-based error-drive learning

0 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,这些词根据分类的不同可分为各类专有名词(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月份的语料(共计2305896字)统计,平均每100个字包含未登录词1.192个(不计数词、时间词)[1],仅专有名词就达到了总词数的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。

在现有的未登录词识别系统中,对专有名词的识别大多只是针对人名、地名或者机构名的一种进行的单独的识别方案。它们一般借助建立专门的地名库、人名库等资料库的方式或采用统计的方式进行人名或者地名的识别。采用统计与规则相结合的方法进行单独识别,这也是未登录词识别中常用的方法,正确率也比较高,但是当同时对专有名词中的人名、地名或者机构名识别时,由于各种规则和统计模型相互影响,所以识别结果往往不是很理想。而未登录词识别中用到的统计模型和语言规则一般是有语言学家大量试验和经验总结的结果,试验曲线中的阈值和经验总结出来的规则都具有相当的主观性。因此文中采用一种基于属性标记的专有名词方法,在应用规则时利用机器学习代替人工试验,从已经人工订正过的语料库中机器自动学习规则,挖掘出专有名词潜在的成词规律和适用环境的规则,避免了人工提取规则的局限性。将此方法应用于专有名词的识别工作中,试图解决专有名词统一识别问题。

1 问题的描述及方法知识介绍

根据日常生活中,人们对包括人名、地名、机构名在内的专有名词(尤其是对初次见到的专有名词)的获取,大多是根据此专有名词的成分结构来区分。例如:姓名结构中的姓氏,地名、机构名的后缀(省、市,电视台)等。而对于一些外国译名等,则是通过上下文的结构。例如:"克林顿对内斯塔尼亚胡说"。如果没有上下文的结构考虑,是很

收稿日期:2006-03-10
作者简介:王 蕾(1980-),女,河南开封人,硕士研究生,主要从事中文信息处理;杨季文,教授,主要从事中文信息处理。

难识别出“内斯塔尼亚胡”为人名的。基于以上考虑，将对专有名词的识别转化为对语料词语进行各类属性标记的过程。

1.1 属性定义

考虑到简单性和实用性，将主要识别的属性分为了 9 种类型，即：人名特征词、人名上文、人名下文、地名特征词、地名上文、地名下文、机构名特征词、机构名上文、机构名下文。

1.2 相关概念的定义

根据定义的属性的位置和结合情况的不同，又定义了以下一些概念：

1)专有名词特征词：包括人的姓氏、地名指示词、机构称呼词等可以反应未登录词类别特征的名词。其中既包括单字特征词(例如：省、市、街等)，也包括多字特征词(例如：酒店、电视台等)。

2)框架对模式：从语料中提取出来的上下文框架结构。例如：句子“李鹏在北京考察企业”中，“在……考察”即被认为是一个框架对模式。而其中的“李鹏在”也可以认为是“专有名词特征词属性……在”的框架对模式。

3)词序对模式：当专有名词只有上文或者只有下文的时候，从语料中提取出来的专有名词属性与上文或者下文形成的前后词序对结构。例如：句子“国务院侨办发表新年贺词”中，国务院侨办为机构名，那么“机构名末尾特征词属性 - 发表”即被认为是一个词序对模式。

4)样例对模式：专有名词首词与上文，或者专有名词尾词与下文成词的结构。例如：句子“厂长对于民红说”中，“对于……说”即被认为是一个样例对模式。

1.3 方法知识介绍

属性标注的过程采用的是基于转换的标注方法。但是考虑到识别专有名词的成词特性，结合了基于实例的方法来首先提取框架模式等必要实例。所以实质上是一种基于转换和实例相结合的方法。

基于规则转换的方法[2~5]是采用一个已经标注好的训练语料库和一个初始标注字典作为输入数据，然后用字典中最常用的标记来标注训练语料库中的每个词，接下来采用基于模板的学习算法构建一组转换规则集，然后对这组转换规则集进行优化后用来对新文本进行标注。Eric Brill 在 1995 年提出的这个基于转换的学习方法时，被用于词性标注，得到了令人满意的结果。而后又被 Ramshaw & Marcus 改进用于英语中的基本名词短语(base NP)识别上。

基于实例(example - based)[2]的方法是首先把人工订正过的语料库分成两部分，其中一部分为训练语料库，另一部分为剪枝语料库。在进行学习的过程中，首先从训练的语料中得到一组短语的组成模式规则，然后把得到的这些规则应用到剪枝的语料中，对这些规则进行优化和评价。最后得到一组正确率较高的短语组成模式，分析的时候再利用这样的模式去和文本中的词序列进行匹配。

2 专有名词框架结构的规则挖掘

文中主要对属性标注采用基于实例和规则转换相结合的思想进行规则提取：首先从训练语料库中抽取“专有名词(指包括人名、地名、机构)上文……下文”框架对模式，“姓氏特征词……人名下文”框架对模式，“地名上文……地名末尾特征词”框架对模式，“机构名上文……机构名末尾特征词”框架对模式，“人名上文 - 姓氏特征词”词序对模式，“地名末尾特征词 - 地名下文”词序对模式，“机构名末尾特征词 - 机构名下文词”词序对模式，以及“专有名词(指包括人名、地名、机构)特征词……连词……专有名词(指包括人名、地名、机构)特征词”框架对模式。然后把得到的这些模式应用到剪枝语料库中，把使用次数较少的模式进行剔除，对使用次数较多的模式采用基于规则转换的方法，从剪枝的语料中得到一组针对这些模式的一组转换规则集，从而得到一组正确率较高的专有名词识别模式和一组针对这些模式的一组转换规则集，并对这些规则集再进行优化，随后就可以使用优化规则模式集对采用最大概率法进行第一趟分词后的数据进行专有名词识别，文中以下将讨论专有名词识别模式的构造过程，如图 1 所示。

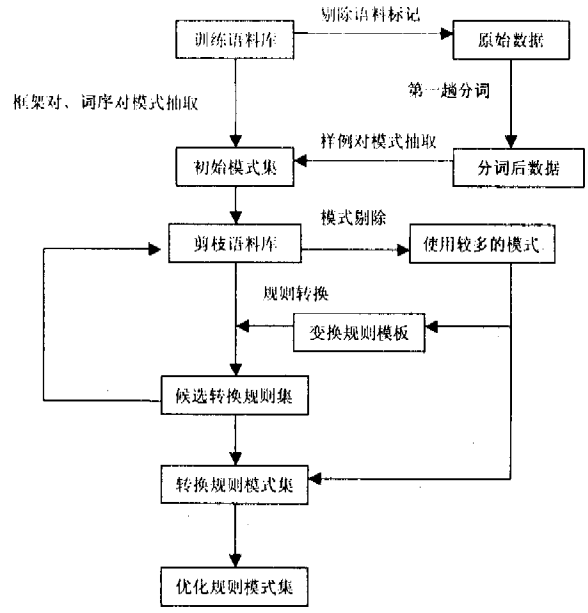

图 1 专有名词识别模式的构造过程

在图 1 中，有几点需要注意的地方是：

1)训练和剪枝语料库。

训练和剪枝语料库是在北京大学计算语言学研究所的标注语料库(1 月份)的基础上进行重新标注。标注的属性包括上文提到的 9 种主要成分属性和 4 种附近属性，具体地分为人名(姓 + 名)、人名上文、人名下文、地名、地名上文、地名下文、机构名、机构名上文、机构名下文、人名间连词、地名间连词、机构名间连词和其他。

2)剔除语料标记和第一趟分词。

由于文中的专有名词识别是建立在第一趟分词的基础上的，所以第一趟分词的好坏将会影响以后的专有名词

识别,为了尽量减少第一趟分词的结果对专有名词识别的影响,在进行模式集抽取之前,先剔除训练语料库中语料标记后对语料中的数据进行分词,然后把分词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,也就是一些样例模式,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,因此"对于……说"就是一个样列模式,在专有名词识别的过程中,碰到"对于……说"的模式时,就有可能要改成"对……说"的模式,而"于"则可能是专有名词特征词了,对于具体模式"对于……说"究竟是否要改成"对……说"的模式,还要根据针对"对于……说"模式的变换规则来定。例如,当前的句子是"对于他来说",而变换规则中有"对于……来说"的特例,这时"对于……说"模式就不需要进行改变。

3)模式集抽取。

它的主要任务就是要能从训练语料库抽取出上文所列举的那些模式,并且合并上上一步找出的样例模式,组成模式集。

4)模式剔除。

是把初始模式集中的模式应用于剪枝语料库中,把那些未被使用或使用次数小于一定阈值(f)的模式删除,这样有利于减少下步的候选转换规则集大小。

5)变换规则模板集。

变换规则模板集采用的策略是在模式的左右一个词的范围作规则变换,为了简化标注词属性种类,规则变换时,参照的对象是规则变换空间的每个词,并不是每个词的属性,因此搜集出来的规则就是针对具体模式中的特例情况。例如,"对……说"是一个具体的模式,它可能存在的一个变换规则是"对大家……说"。因此,如果在句子中碰到"对大家好好地说",就不会对"对……说"模式中的词进行专有名词识别。

3　专有名词识别

在有了专有名词识别规则模式集之后,接着将讨论基于实例和规则转换相结合的专有名词识别方法。

3.1　采用最大概率法进行第一趟分词

假设汉字串为 C,$P(C)$ 是汉字串的概率,W 是 C 汉字串的一种切分,$P(W)$ 是 W 这种切分的概率,那么 $P(W \mid C) = P(W)P(C \mid W)/P(C)$ 表示 C 汉字串中按照 W 这种方式进行切分的概率,由于 $P(C \mid W) = 1$,而 $P(C)$ 对于 C 串的所有切分又是常数,那么寻找汉字串 C 最大概率的切分方式,只需寻求使得 $P(W)$ 值最大的切分方式就可以了,而 $P(W)$ 的计算一般可以简化为用词与词之间的二元转移概率的乘积近似表示。

3.2　初始词标注

在用最大概率法进行第一趟分词结束后,紧接着的事件就是要用基于实例的模式集去给那些新分出来的词标注属性,在分析和标注词序属性时,将参照以下的步骤进行模式匹配:

1) 判断当前词序是否有符合样例模式集中的词序,如果有就再根据当前具体样列模式集的转换规则,查看是否需要改变模式,否则进入第 2)步;

2) 判断当前词序是否符合"专有名词特征词……连词……专有名词特征词"词序对模式,如果符合,标注这些词并继续后续词序属性标注,否则进入第 3)步;

3) 判断当前词序是否符合"专有名词上文……下文"词序对模式,如果符合,标注这些词并继续后续词序属性标注,否则进入第 4)步;

4) 判断当前词序是否符合"姓氏特征词……人名下文"词序对模式、"地名特征词……地名下文"词序对模式、"机构名特征词……机构名下文"词序对模式。如果符合,标注这些词并继续后续词序属性标注,否则进入第 5)步;

5) 判断当前词序是否符合"人名上文……姓氏特征词"词序对模式、"地名上文……地名特征词"词序对模式、"机构名上文……机构名特征词"词序对模式,如果符合,标注这些词并继续后续词序属性标注,否则,当前词标注为"其他"属性,并继续后续词序属性标注。

3.3　运用规则排除具体模式下的特例情况

这个阶段的主要任务就是从标注过程中已识别出的那些属性模式中提取"专有名词候选词",文中把在属性模式中的那些词称为"专有名词候选词"。然后再根据规则对这些模式中"专有名词候选词"进行筛选,例如:"克林顿对内斯塔尼亚胡说"中的"内斯塔尼亚胡"和"厂长对大家说"中的"大家"都是属于"专有名词候选词",但"内斯塔尼亚胡"是国外人名,而"大家"是一般的词语,这时就需要通过规则法排除这种模式下的特例情况。

3.4　专有名词提取

专有名词提取阶段的主要任务就是确定专有名词的边界,并把提取的专有名词进行归类。对专有名词边界进行确定可以按以下步骤进行:

1) 如果专有名词存在于"专有名词上文……下文"框架对模式中,那么专有名词上文的后续字就是专有名词的上界,专有名词下文的前序就是专有名词的下界。

2) 如果专有名词存在于"姓氏特征词……人名下文"框架对模式、"地名上文……地名特征词"框架对模式、"机构名上文……机构名特征词"框架对模式中。那么存在于其间的包括专有名词特征词在内的所有成分,即可认为是专有名词。

3) 如果专有名词存在于"人名上文－姓氏特征词"词序对模式、"地名特征词－地名下文"词序对模式、"机构名特征词－机构名下文"词序对模式中。对于"人名上文－姓氏特征词"词序对模式,姓氏特征词为专有名词上界,姓氏特征词后续碎片中的成分为专有名词内容。对于"地名特征词－地名下文"词序对模式、"机构名特征词－机构名下文"词序对模式,如果专有名词在分词的碎片中,那么专有名词的上界就是分词的碎片中的起始位置,下界就是专

有名词特征词,如果专有名词不在分词的碎片中,那么默认地名特征词的前两个词是专有名词的上界,地名特征词是专有名词下界。

4) 如果专有名词存在于"专有名词特征词……连词……专有名词特征词"框架对模式中,对于是"人名特征词……连词……人名特征词"框架对模式,"人名特征词……连词"间的专有名词的上界是人名特征词,下界是连词的前序词,"连词……人名词特征词"间的专有名词的上界是人名特征词,下界是人名后续词。对于是"地名特征词……连词……地名词特征词"框架对模式或"机构名特征词……连词……机构名词特征词"框架对模式,"地名特征词……连词"或"机构名特征词……连词"间的专有名词的上界是分词的碎片中的起始位置或者是地名特征词、机构名的前序词;"连词……地名词特征词"或"连词……机构名词特征词"框架间的专有名词的上界是连词后续词,下界是地名词特征词或机构名词特征词。

3.5 专有名词归类

如果专有名词的边界确定之后,专有名词的归类就相对容易多了,专有名词的归类是指识别出来的专有名词是属于人名、地名还是机构名类型。专有名词的归类只要看识别专有名词的模式,如果模式是"人名上文……下文","人名上文-姓氏特征词","姓氏特征词……人名下文",那所识别专有名词就属于人名类型;如果模式是"地名上文……下文","地名上文……地名特征词","地名特征词-地名下文",那所识别专有名词就属于地名类型;如果模式是"机构名上文……下文","机构名上文……机构名特征词","机构名特征词-机构名下文",那么所识别专有名词就属于机构名类型。

4 试验结果与分析

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

4.1 试验用语料和评测标准

试验使用了北京大学计算语言学研究所的标注语料库(1998 年 1 月)。在此语料的基础上根据自定义的属性重新进行标注后,作为试验用语料。

针对专有名词的识别,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 试验结果

根据模式提取中的阈值设定的不同,在封闭测试中的试验比较如表 1 所示。

表 1 试验结果

阈值(f)的取值	准确率(%)	召回率(%)	自学习到的规则数
1	95.3%	92.5%	6759
2	87.2%	83.7%	8432

4.3 试验结果分析及后续工作

封闭测试中,阈值的选取直接决定了规则的提取和准确率与召回率的结果。阈值的取值越小,造成采用的模板增多,在基于转换的错误驱动进行规则学习的时候学习到的规则就越少。但是,在封闭测试试验中,准确率和召回率反而增高。但是规则的减少,使模板应用的局限性加大了,这样不利于其在开放测试中的应用。所以,在更大的语料里面进行阈值确定和规则的大量提取是以下将要进行的工作。

5 结 论

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,提出了基于属性标记的专有名词的识别。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,重新标注语料,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,找出了一系列专有名词出现的上下文环境和规则。在此基础上进行了小规模语料的封闭测试识别,取得了相当好的效果。目前实验表明基于属性标记的专有名词识别方法是行之有效的。但此方法的有效运用,需建立在拥有大量的熟语料库的基础上,而且正确的模板阈值(f)的确定,也关系着系统的准确度。这些存在的问题也正是需要进一步完善的地方。

参考文献:

[1] 张华平,刘 群.基于角色标注的中国人名自动识别研究[J].计算机学报,2004,27(1):85-91.

[2] 孙宏林,俞士汶.浅层句法分析方法概述[J].当代语言学,2000(2):74-83.

[3] 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.

[4] Brill E. A Simple Rule-based part of speech tagger[C]//In: Proc 3rd Conference on Applied Natural Language Processing. Trento:ACL,1992.

[5] 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.

[6] 万建成,杨春花.书面汉语的全切分分词算法模型[J].小型微型计算机系统,2003,24(7):1247-1251.

(上接第 161 页)

Science. New York:Springer,1986,417-426.

[5] Koblitz N. Elliptic Curve Cryptosystems[J]. Mathematics of Compution,1987,48:203-309.

[6] 李道丰,揭金良.基于椭圆曲线的数字有序多签名方案[J].通讯和计算机,2005,2(2):36-38.