ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.5,No.8,March 2009, pp.1928-1930

E-mail: xsjl@cccc.net.cn
http://www.dnzs.net.cn
Tel:+86-551-5690963 5690964

基于最大熵模型的中文名词短语识别

张惠春

(晋城职业技术学院,山西 晋城 048000)

摘要:中文名词短语识别在自然语言处理已经得到了广泛应用。该文首先对名词短语识别问题进行描述,然后利用最大熵模型建立名词短语识别系统,通过实验选取最大熵模型的特征,最后利用选取的特征进行名词短语识别,实验结果表明系统达到了较高的准确率和召回率。

关键词:最大熵模型;名词短语;特征选择

中图分类号:TP391 **文献标识码**:A **文章编号**:1009-3044(2009)08-1928-03

Chinese NP Recognition Based on a Maximum Entropy Model

ZHANG Hui-chun

(JinCheng institute of Technology, JinCheng 048000,China)

Abstract: Chinese NP Recognition had wide . This paper firstly described the chinese NP recognition, constructed a system for chinese NP recognition using the maximum entropy model. It selected a feature by testing the system. Finally It recognized the Chinese NP uesed the feature which was selected. The experimental result proved that the system reach high precision and recall.

Key words: Maximum entropy; Chinese NP; Feature selection

1 名词短语识别问题描述

给定经过分词和词性标注的文本 $T=w_1/t_1\ w_2/t_2\ \cdots\ w_n/t_n$,目标是识别 T 中所有的名词短语。我们将识别名词短语的任务定义为给名词短语添加边界标记。[1]

例:南帆/nrm<O> 说/vs<O> ,/w<O> 没有/vs<O> 任何/rg<B> 作家/ng<I> 可以/vs<O> 垄断/vs<O> 一/mx<B> 种/qk<I> 文体/ng<I> 。/w

B 表示名词短语的开始,I 表示在名词短语中,O 表示非名词短语,那么上例中我们可以清楚地看到"任何作家"和"一种文体"是名词短语。

2 最大熵模型

设随机过程 P 所有的输出值构成有限集 Y,对于每个 $y \in Y$,其生成均受上下文 x 影响,x 属于有限集 X。对于名词短语识别问题,x 表示待识别词 w_0 的上下文环境,y 表示输出,在此只有三个值:B,I 和 O。模型的目标是:构造一个可以精确描述随机过程行为的模型,该模型对于给定上下文 $x \in X$,可以计算输出为 $y \in Y$ 的条件概率 p(y|x)。[2]

2.1 问题描述

在机器学习的过程中,我们需要对随机过程进行观察,从专家标注的语料库中抽取大量的训练样本$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$,即语料库中每个词的上下文及其确定的输出 y。可以用训练样本的经验分布来表示它,即

$$\tilde{p}(x,y) \equiv \frac{freq(x,y)}{\sum_{x,y} freq(x,y)} \tag{1}$$

其中,freq(x,y)表示特征(x,y)在样本中出现的次数。

2.2 相关定义

定义 1.特征

设 $x \in X$ 且 $x= w_1w_2\cdots w_n$,设 c 是 x 的子串(长度$\geqslant$1),若 c 对 $y \in Y$ 具有表征作用,则称(c,y)为模型的一个特征。特征分为原子特征和复合特征:若串 c 的 1,则称(c,y)为原子特征;否则,称(c,y)为复合特征。

定义 2.特征函数

特征函数 fi(x,y)是一个二值表征函数,是我们对所获取的特征的一种表示方法,表示(x′,y′)是否与特征(c,y)有关。定义(x′,y′)关于特征(c,y)的特征函数为

$$f_{(c,y)}(x',y') = \begin{cases} 1 & if\ \ c是x'的子串,\ 且y'=y \\ 0 & else \end{cases} \tag{2}$$

由以上定义可以看出,样本中待识别词 w_0 和出现在 w_0 周围的词都可以作为模型的特征,因此与模型有关的备选特征的集合会很大,但模型选出的特征只是真正对模型有用的特征,是候选特征集合的一个子集,它能较完整地表达训练语料中的数据。在此引入约束。

定义 3.约束

收稿日期:2009-02-25

作者简介:张惠春(1981-),女,助教,硕士,主要研究方向:自然语言处理。

假设$\tilde{p}(f)$为特征 f 对于经验概率分布$\tilde{p}(x,y)$的数学期望,表示为

$$\tilde{p}(f)=\sum_{x,y}\tilde{p}(x,y)f(x,y) \tag{3}$$

p(f)为特征 f 对于由模型确定的概率 p(x,y)的数学期望,表示为

$$p(f)=\sum_{x,y}p(x,y)f(x,y) \tag{4}$$

其中,

$$p(x,y)=p(x)p(y|x) \tag{5}$$

令 $p(x)=\tilde{p}(x)$,则限定所求模型的概率为在样本中观察到的事件概率,而不是所有可能出现的事件的概率。若 f 对模型有用,则令 $p(x)=\tilde{p}(x)$为约束。

概括地说,最大熵模型的基本思想是:给定训练样本,选择一个与训练样本一致的模型,最大熵模型应选择与这些观察相一致的概率分布,而对于除此之外的情况,模型赋予均匀的概率分布。

2.3 最大熵模型的特征

由于是对每一个词进行名词短语标注,每一个词的名词短语标注过程都被看作是一个事件,因此由当前词及它的上下文环境来确定一个事件的特征集合。根据影响当前词名词短语标注的各种因素,可以定义特征空间。[3]

定义 4:

特征空间由以下两部分内容组成:

1)词性,当前词及其前后各两个词的词性;

2)词,当前词及其前后各两个词。

根据定义 1, 我们可以知道最大熵模型中的特征可以分为原子特征和复合特征两种。从以上特征空间的定义 4,可以得到十个原子特征,例如当前词的词性,当前词的前一个词的词性等等,表 1 列出了特征空间中所有的原子特征表达式及各个特征所代表的意义。

表 1 原子特征

序号	原子特征	特征意义
1	CurPos	当前词性标注
2	CurWord	当前词
3	Pos-2	前后词的词性标注
4	Pos-1	
5	Pos+1	
6	Pos+2	
7	Word-2	前后特定的词
8	Word-1	
9	Word+1	
10	Word+2	

由于在上下文中,仅仅用原子特征不足以表示上下文中的一些呈现规律性的语言现象, 对文本中词的名词短语标注贡献较小,必定造成低识别率,所以应该在最大熵模型中使用复合特征。复合特征是由原子特征组成, 而复合特征集又由复合特征组成。复合特征表示为二值特征函数的形式与原子特征相似, 只是在取值时需要满足的条件变多。本文中定义的复合特征集中共含有八条复合特征,这些复合特征能够对名词短语的标注提供较多的信息,可以提高名词短语标注的正确率。符合特征集如表 2 复合特征表所示。

表 2 复合特征表

序号	1	2	3	4	5	6	7	8
复合特征	CurPos, Pos+1	Pos-1, CurPos	Pos-1, Pos+1	Pos+2, Pos+1, CurPos	Pos+1, CurPos, Pos-1	Pos-2, Pos-1, Pos+1, Pos+2	Pos+2, Pos+1, CurPos, Pos-1, Pos-2	CurWord, Pos+1

3 特征选取及实验

上面定义了原子特征集和复杂特征集,然而并非所有特征都适合引入到最大熵模型中去。对于要处理的问题,特征所含的信息量越大,该特征就越适合引入到模型中。通过原子特征和复合特征得到的特征构成候补特征集合,然后从中选取对模型最为有用的特征。本文通过采用实验的方法选取对模型最为有用的特征。

3.1 系统结构

如图 1 是所示,系统主要有模型训练和名词短语识别两大主要模块构成。在模型训练模块中,系统在训练语料上一次进行特征提取、特征选择和迭代训练工作,从而得到模型的参数。在名词短语识别模块中,系统用训练过的最大熵模型在自动分词后的待标注文种上进行名词短语的标注。在模型训练模块中,最大熵模型采用文献[3]上用 JAVA 语言开发的 Maxent 软件包。此软件包运算速度快,占用机器内存少,使得我们的大量数据在较短的时间内完成训练及测试。如图 1 名词短语识别系统流程图。

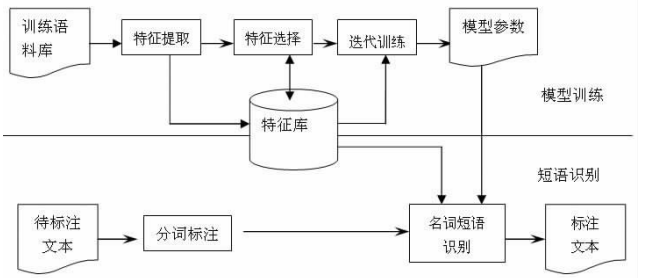

图 1 名词短语识别系统流程图

3.2 实验结果及分析

在表 1 中我们给出了实验中用到的特征,但是并不是所有的特征对名词短语的标注贡献都是相同的。有一些特征对标注贡献较大,可是还有一些特征可能对标注起到了副作用。所以接下来我们必须通过实验对这些特征进行检验,选出合适的特征组合,作为最大熵模型的输入。对表 2 中的八个复合特征分别进行实验,进行了基于最大熵模型的名词短语标注实验。表 3 是这八个复合特征的实验结果。

表 3 复合特征实验结果

复合特征序号	1	2	3	4	5	6	7	8
正确率	79.30%	87.32%	86.00%	85.60%	85.43%	86.03%	85.89%	66.33%

从表 2 中可以看到,系统 1 采用当前词的词性和当前词的后一个词的词性作为复合特征,系统 2 采用当前词的词性和当前词的前一个词的词性作为复合特征,而这两个系统的结果相差很大,正确率相差了 8.02%,而这两个系统恰恰又是这八个系统中正确率次低和最高的两个系统。这充分说明当前词的前一个词性对分类的贡献非常大,而后一个词的词性对分类的贡献相对较小。接下来我们再来看特征 6 和特征 7。特征 6 是采用当前词的前两个词的词性和后两个词的词性作为系统实现的复合特征,特征 7 是采用当前词的前两个词的词性、后两个词的词性以及当前词本身的词性作为复合特征,从表 3 中可以看到,特征 6 的准确率要比特征 7 的准确率高一点点,所以可以得出结论,当前词的词性对分类的贡献不大,而且会起到负面的作用。再来看特征 8,特征 8 的准确率非常低,和其它系统的准确率不在同一个水平上。特征 8 采用的复合特征由当前词和当前词后的第二个词的词性组成的,由于使用词作为特征,容易引起数据稀疏,再加上当前词后的第二个词的词性这一特征,距离当前词比较远,对分类的影响较小,所以,它的分类结果会很差。

综合表 3 的实验结果和以上对实验结果的分析,最后我们选定原子特征为 Pos-1,复合特征选取在上述实验中效果较好的,在表 2 中的复合特征 2,3,6。所以最终系统的特征模板为:

Pos-1,CruPos,Pos-1,Pos-1,Pos+1,Pos-2,Pos-1,Pos+1,Pos+2

用此特征模板进行实验,实验结果就是基于最大熵模型名词短语识别的最终实验结果,如表 4 试验结果表所示。

表 4 实验结果表

特征	准确率	召回率
Pos-1, CruPos, Pos-1, Pos-1, Pos+1, Pos-2, Pos-1, Pos+1, Pos+2	88.01%	90.32%

4 结束语

该文首先介绍了最大熵模型的基本原理,然后针对名词短语识别的任务为最大熵模型选取特征,构建了复合特征。接下来设计了系统的结构,针对 8 个复合特征分别进行实验。通过对基于复合特征系统的实验结果的分析,选定了 4 个特征作为最终系统的输入,结果显示了较高的正确率和召回率,完成了使用最大熵模型实现名词短语的识别的任务。

为了进一步提高名词短语的识别的准确率,可以在加入词、词性信息的基础上加词语的语义信息,在统计方法上增加规则弥补模型的不足,并考虑如何减少自动分词错误对识别结果的影响。

参考文献:

[1] 赵军.基于转换的汉语基本名词短语识别模型[J].中文信息学报,1999,13(2):1-7.
[2] 周雅倩,郭以昆,黄萱菁,等. 基于最大熵方法的中英文基本名词短语识别[J].计算机研究与发展,2003,40(3):440-446.
[3] 周明.基于语料库的中文最长名词短语的自动抽取[J].计算语言进展与应用,北京:清华大学出版社,1995:50-55.
[4] 周强.汉语最长名词短语的自动识别[J].软件学报,2000,11(2):195-201.

(上接第 1927 页)

而在对用户兴趣的分段存储可以由用户制定个性化信息服务来实现,首先应能够满足用户的个体信息需求,即根据用户提出的明确要求提供信息服务,或通过对用户个性、使用习惯的分析而主动地向用户提供信息服务

利用 CSCL 系统对用户的历史记录数据进行分析,可以根据用户的访问兴趣、访问频度、访问时间等改进服务;对 CSCL 系统使用记录中的序列模式进行分析,并利用分析结果协调用户与数据库之间的信息交互,开展个性化服务工作;通过对具有相似浏览行为的用户进行分组,分析其共同特征,从而准确地把握信息用户的个性和需求,及时调整服务的角度和内容,有效向用户提供更适合、更有针对性的个性化服务。

建立定制化个性化服务 在此项模块内设制功能包括:

1) 定制用户想要的资源。

2) 书签功能,该功能类似于浏览器提供的 bookmark,允许用户挑选若干个资源放入书签。

3) 最新信息通告。在 CSCL 系统的首页中显示每天的最新资料或者消息。

4) 更多的信息资源主页链接。例如在显示的首页放入用户可能使用到的资源的网站地址。

5) 定制用户页面的个性化(不仅能让用户按自己的喜好来更换页面颜色外,最好有能依照自己所想学习的专业内容的要求或个人兴趣)。

6 结束语

本文讲述的是 CSCL 系统中的时间对比表中的记录可以获取到用户的稳定兴趣因素。针对目前的 CSCL 系统的发展需要,结合时态信息处理技术和数据库技术,提出一种新的结合方法。本文提出的方法不局限于 CSCL 系统,可以拓展到其他应用领域。

参考文献:

[1] 黄楠,刘爱琴.时态数据库技术[J].微机发展,2002(1).
[2] 郭家义.个性化信息环境研究[J].中国图书馆学报,2004,30(3)
[3] 汤庸.时态信息处理技术研究综述[J].中山大学学报:自然科学版,2003(4)
[4] 唐常杰.时态数据库的成果、缺陷与未来—时态数据库二十年回顾之二[J].计算机科学,1999,26(3).