基于概念属性特征的中文地名识别处理*

李 诺[1,2] 张 全[2]

[1]中国科学院研究生院 北京 100039；[2]中国科学院声学研究所 北京 100190

摘要： 在最大熵等统计机器学习模型当中，特征函数的选择可以说是对系统整体性能影响最大的部分。本文不仅使用了传统的词、词性等作为特征，同时基于HNC语言概念理论体系，以语义概念为特征进行训练。通过对语义概念符号的正确表示，把语义分析的内容加入到统计分析中去。把词语按照语义分类的部分属性加以利用。并且，本文中还尝试使用变长的训练窗口，对每一个特征刻画的更加仔细。最终在实际语料的测试中证明加入语义特征确实可以改进识别效果。

关键词： 最大熵模型、特征函数、HNC理论、语义概念属性、变长窗口

Chinese Place Name Recognition Based on Concept Features

Li Nuo[1,2] , Zhang Quan[2]

[1]Graduate University of Chinese Academy of Sciences Beijing 100039

[2]Institute of Acoustics, Chinese Academy of Sciences Beijing 100190

Abstract: The feature functions were reckoned as the most important part of the maximum entropy model which could affact the last result of system. In this paper, we no only select some common features as words and part-of-speech but also semantic features which based on HNC(Hierarchical Network of Concepts) theory. By semantic concept symbols, we could combine statistic analysis and semantic analysis together. Meanwhile, we also use moving window take the position of fixed window of maximum entropy model which could lead to more specific feature functions. As a result, it was demonstrated in real corpus test that semantic features surely contribute to better result.

Keyword: Maximum, feature function, HNC theory, semantic concept features, moving window

一、引言

中文地名识别是中文命名实体识别中的重要工作之一。一方面，地名等命名实体的识别工作作为中文分词的子任务，夹杂在分词碎片当中，正确地识别地名也就可以更好的分词；另一方面，随着问答系统、内容抽取等自然语言理解领域中的热点问题研究的发展，地名本身作为描述事件的一个重要方面越来越受到人们的重视，而对地名的识别工作逐渐成为独立的研究内容。

* 本文承国家 973 项目“自然语言理解的交互引擎研究”（2004CB318104）、国家科技支撑计划课题“搜索引擎中的语言翻译基础研究”（2007BAH05B02-05），中科院声学所知识创新工程项目“句群理解处理理论及其应用”（0654091431）、中国科学院声学研究所“所长择优基金”（GS13SJJ04）、中国科学院青年人才领域前沿项目（O754021432）的资助。

作者： 李诺 (1983 -)，男，硕士，研究方向：自然语言理解。 张全：研究员，博导。

E-mail: li_nuo@yahoo.cn

在相关研究当中，中文地名的识别大多还是作为命名实体识别的一部分加以研究的。而采用的识别方法也以识别命名实体通用的方法为主。具体而言，相关学者多是采用统计机器学习辅以规则的方法进行中文地名的识别。比较常见的统计机器学习方法如隐马模型的方法[1、2、5]、最大熵模型[1、2、4]、支持向量机[3]、条件随机场[1]、神经网络[5]等。地名识别 F 值结果也均在 86%-92%之间，效果比较理想。在这些模型当中，最大熵模型可以很灵活地选择特征。在建模时，实验者只需要集中精力选择特征，而不需要花费精力考虑如何使用这些特征。选择的特征函数可以来自同一类特征，也可以来自不同类特征。同时，特征间不冲突且无需满足独立性假设。因而可以先通过设置不同类型的特征提高系统的性能。再通过最大熵模型方便简单地把各种类型的特征融合到一起。这样，可以把能够想到的多种特征综合。也可以反复应用不同特征在同一词语处。由于上述原因，本文在设计融合语义概念属性特征时，首先想到采用最大熵模型。

而对于应用最大熵模型进行地名等命名实体识别的研究当中，一般比较容易想到的以词及词性的特征为主。除了词及词性的特征，诸如语态、动词框架、动词路径、中心词等句法特征都可以采用，但由于相关语料不完善，以及对上述概念的划分不如词、词性直接，所以应用性较差。文献[1]中为了识别中文地名，用到了词及词性的特征，同时考虑了对地名内部及外部出现的词进行分类。文献[2]同样应用了词及词性的特征，但在特征窗口长度的选择上，只选择出现 10 次以上的特征，这样可以使所取的特征更有一般性，忽略一些个别的训练目标，但同时也使得特征的区分度不够。文献[4]同样先建立较大的特征集合，然后再从特征集合当中通过计算熵增的方法逐步筛选最有区分度的特征。文献[6]的识别目标是中文机构名。文章中也试图用符号 I/S/O/C 对出现在机构名前后的词进行归类，希望在词、词性一级分析的同时，也能对机构名的上下文内容进行归类分析。综上所述，识别中文地名等命名实体时所取的分析特征还是以词、词性为主，虽然许多研究人员试图加入其他分类特征，但由于分类目标不明确，导致效果一般。

本文作者在应用词及词性为特征的同时，加入基于语义的词汇分类特征。在基于语义的明确分类方式的前提下，语义特征能够更好的把词语归类或标识。这种语义特征的表示是基于 HNC 语言概念层次网络理论体系，体系中通过词语符号表现出不同概念之间的关联性。也就是说，加入基于语义概念关联的分析结果作为最大熵特征。通过对真实语料的分析，与只用词及词性作为特征的最大熵模型相比，语义特征的加入确实可以提高系统的整体性能。另一方面，本文作者也分析了，对于变长窗口应用在最大熵模型中的效果并给出了在小规模语料测试中的结果。

二、基于 HNC 概念特征的分析

此处将主要介绍与提取 HNC 语义特征紧密相关的 HNC 基元符号体系。

HNC 是英文（Hierarchical Network of Concept, 概念层次网络）的缩写，HNC 理论，即概念层次网络理论，是一个关于自然语言理解处理的理论体系。

HNC 理论认为，世界上各式各样的语言可以表述为有着各式各样的语言空间的表述形式。而每一种语言空间的表述形式对应一种语言空间符号，即相应的文字。各种语言拥有同一个语言概念空间，即表述语义的空间。HNC 理论的一个重要假设就是假设世界上的各种语言只有一个语言概念空间，因而，只要通过语言概念空间研究清楚其概念、语义之间的联系，也就可以推广到所有语言空间。

概念基元是概念的最小单位，其特性为概念范畴的有限性。正因为概念基元有限，所以我们

可以通过概念间的联系表示出所有概念基元；而概念本身无限，我们可以通过各种概念基元的组合表示出各种概念来。至此，我们可以发现，通过用概念基元表示概念的形式，已经隐含着概念符号中将包含着概念基元间的关系，而其关系是通过概念组合符号具体标出。

HNC 理论对自然语言概念的符号化表述可以一般化为：

$$\sum\{\text{类别符号串}\}\{\text{层次符号串}\}\{\text{组合结构符号}\}\{\text{类别符号串}\}\{\text{层次符号串}\}$$

上式表示为概念表达式由类别符号、层次符号和组合结构符号三类符号构成，类别符号串和层次符号串构成一个概念基元的表达式，两个或多个概念基元通过组合结构符号的组合而构成新的概念。由单个概念基元构成的概念称为简单概念，有两个或多个概念基元组成的概念称为复合概念。

基于上述的 HNC 符号体系，可以把所以概念表示出来。例如：迅速 ul009c22，编辑 va34;pa34，承担 v901，发展 v10a8 等。HNC 符号中蕴含着概念联想的丰富知识，为计算机进行概念联想操作提供了简明有效的途径和工具。

本文中用到了如表 1 所示的 12 类概念类别：

序号	名称	概念类别符号	说明
1	抽象名词性概念	g/r/xr/gr	静态表示/ 效应表示/ 符合概念类别具有 g, r 属性的
2	具体名词性概念	p/w	人或者与其他概念复合表示人的/ 物或者与其他概念复合表示物的
3	特指概念	f	加在其他概念类别前，表特指概念
4	动态概念	v/vv	动态
5	属性和物性概念	u/ug/gu/x/jx/ px/gx/rx	属性/ 仅修饰 g, r 类概念的成分/ 既可以是静态表示又可以是属性表示/ 物性或与其他概念复合具有 u 属性的
6	副词属性概念	uu/uv	可修饰 v, u, ug 的成分/ 仅修饰 v 的成分
7	量词型概念	zz/zzv	名量词/ 动量词
8	时间类概念	j1	表示时间的基本概念，或与其他概念复合表示时间
9	空间类概念	j2	表示空间的基本概念，或与其他概念复合表示空间
10	表示切分组合的语言逻辑概念	l0-l5/l8	语言逻辑概念
11	表示其他功能的语言逻辑概念	l6-l7/l9-lb	语言逻辑概念
12	语习类概念	f1-fb	语习类概念

表 1 用作最大熵模型特征的概念类别

HNC 理论的基层中对概念的分类还有其他类别符号。不过，考虑到作为特征的概念属性在实际训练语料中既不能出现的太少，也不能出现的太多。如果出现次数过少，则训练后参数 λ_i 值就很小，说明这个特征不具有代表性，对系统最终结果影响很小；如果出现的次数过多，则区分度太差。

以概念类别符号作为特征函数，是把语义知识融入统计模型当中。一方面利用了最大熵模型灵活的优势，而另一方面也把对词的统计分析提升到语义分析。是先从词引申到概念，再应用概念代替词标签进行分析。

相对于仅使用词及词性作为最大熵特征，用 HNC 概念类别作为特征不仅为特征的选择提供了一种新思路，同时，相比词语作为特征还有如下优势：

（1）语义或概念是词的提升，能更好的揭示特征的本质。

（2）语义或概念也可以看作是词的聚合。同一种概念可以由多个不同词来表示，因而以概念类别符号为特征会使特征更加集中。如：多个词可能拥有一个概念类，或其概念类间有一定关联。

（3）概念类别的特征表示能够更好地对所有特征按照语义分类。更加整洁。当出现错误的时候也更容易发现是哪一类特征的错误。

三、最大熵统计模型与特征变长窗口

选取最大熵模型主要因为其可以很灵活地选择特征，和使用各种不同类型的特征，同时特征又容易更换，十分灵活。最大熵模型的原理部分可以参考文献[7]，这里毋庸赘述。迭代部分也是采用了较为常见的 IIS 算法。

影响最大熵模型识别效果最核心的部分即为特征函数选择部分。本文先选择待分析地名前后词及词性作为特征。之后，选择上文所述的 12 类 HNC 概念属性作为特征。特征函数具体形式举例为：

$$f(x,y)=\begin{cases}1 & \text{当} x\in \text{抽象名词属性概念} \,\&\, y=\text{地名} \\ 0 & \text{其他}\end{cases}$$

每次只分析一个地名前后的词或词性等特征的属性为单一属性，而单一属性的组合又可以组成复合属性。同时，文献[8]指出，为了简化分析和降低计算复杂度，可以只选择出现次数大于等于 5 次的特征。抛弃出现小于 5 次的特征代价相对较小。因而本文在比较加入语义特征前后的最大熵模型识别效果时均只选择出现次数大于等于 5 次的特征。

本文的另一个实验是尝试对同类特征使用变长的活动窗口。单一的特征往往会是刻画训练语料的粒度显得太粗。当使用复合特征的时候，会对训练语料的刻画更加细致，但是在提高识别正确率的同时，有时会降低系统的召回率。一个解决这个问题的方法就是使用变长的特征窗口。通过随时改变特征窗口的长度，可以更好的适应训练语料的变化，也能得到更准确的训练参数。以词的特征为例，变长窗口设计规则为：

1、初始化所有的特征窗口都为[-1, 1]，此时窗口长度 L=3

2、do

{

计算此时的训练特征结果 p1

计算 L+2，（范围变成[-2, 2]）后的训练结果 p2。

计算 L-2，变成单字特征后的训练结果 p0。

}while ((p1<p2 || p1<p0) && L>1 && L<7)

3、If L==1 此时为单字特征；

4、否则记住最终的窗口长度。

变长窗口的方法需要频繁的计算随窗口变化引起训练结果改变，以确定每一步的最佳窗口。计算复杂度更高。与单字特征加复合特征混合的方法相比，变长窗口的复杂度更高。因为虽然免去了单字特征与复合特征的重叠，但频繁计算系统结果的变化增加了系统的整体负担。

四、实验结果与分析

本文选择《人民日报》1998 年 1 月 1 日—20 日的语料作为训练语料和封闭测试集，取 1 月 21 日—31 日的语料为开放测试集。其中训练语料部分共有地名 17825 个，开放测试集部分共有地名 10065 个。对于只取词及词性的最大熵模型测试结果记为 ME，加入 HNC 语义属性特征后的最大熵模型测试结果记为 ME+HNCtag。识别结果如下：

		召回率	正确率	F 值
ME	封闭测试	91.04%	83.49%	87.10%
	开放测试	87.13%	79.88%	83.35%
ME+HNCtag	封闭测试	91.83%	85.29%	88.49%
	开放测试	88.29%	80.46%	84.19%

表 2　语义特征加入前后的比较识别结果

对于变长窗口的地名识别，计算复杂度较高。由于窗口每计算一次潜在地名都要循环变化一次，并且重新计算一次对于新窗口的特征函数取值下的概率结果。实验中只取了《人民日报》1998 年 1 月 1 日的语料作为训练集和封闭测试集。此部分语料只含有 557 个中文地名。特征函数的选择为词、词性的一般特征加 HNC 属性特征。未加变长窗口的记为 ME+HNCtag，变长窗口的最大熵模型记为 MWME+HNCtag。（MW=moving window）识别结果如下：

	召回率	正确率	F 值
ME+HNCtag	83.48%	77.50 %	80.38%
MWME+HNCtag	83.66%	82.77%	83.21%

表 3　最大熵模型定长窗口与变长窗口比较

对于上述实验结果进行分析有如下特点：

（1）在未加入语义特征时，识别结果在 84%左右，和文献[1][2]中同样运用最大熵识别方法所得结果相似。而文献[3][4]识别结果都在 91%左右，但均加入了大量规则。而本文在词及词性的特征基础上加入 HNC 属性特征确实可以提高系统的识别效果。

（2）变长特征窗口的选择对相同测试集的评价结果也有明显提高。但计算复杂度很高。

对于结果中的错误，也可以分为以下几类：

（1）普通用词。如“贫困地区”。这类普通词自身组成结构与真实地名相似，尤其当使用地名尾词作为特征时更容易出现这种情况。文献[2]也遇到了类似的问题。同时，加入语义特征分

析其上下文并不能对这种错误有所改善。

（2）地名中或前后含有介词、连词。如和平门。这种错误在加入 HNC 语义特征后有所改善，其原因是语义特征本质是对词语更好的分类，介词、连词归于虚词类，其在 HNC 属性标记时很多属于语言逻辑概念或语习类概念，区分较词汇明显。

（3）单字地名错误仍然是解决不理想的内容之一。这是由于在词语切分时单字地名更容易出现划分错误。

五、结语

使用最大熵模型进行中文地名识别，最重要的是特征函数的选择。而加入语义特征能够更好地发挥最大熵模型的优势。同时，基于 HNC 理论的概念基元符号体系能够很好的表示概念之间的关系。系统改进的目标及下一步研究工作为一方面更好地细化 HNC 符号，另一方面更好地选择 HNC 属性，选择那些更有分辨性的特征以提高系统整体性能。同时，研究目标也可以从地名推广到其他命名实体。

参考文献：

[1] 廖先桃. 《中文命名实体识别方法研究》[D]，哈尔滨工业大学，2006.

[2] 张晓艳. 《基于混合统计模型的汉语命名实体识别方法的研究与实现》[D]，国防科学技术大学，2004.

[3] 李丽双，黄德根，陈春荣，杨元生. 《SVM 与规则相结合的中文地名自动识别》[J]，中文信息学报，2005，5：51-57.

[4] 钱晶，张杰，张涛. 《基于最大熵的汉语人名地名识别方法研究》[J]，小型微型计算机系统，2006，27(9)：1761-1765.

[5] 欧嘉致，陈凯江，李宗葛. 《基于 NN_HMM 混合模型的汉语地名识别系统》[J]，计算机工程与应用，2002，23：220-222，228.

[6] 杨德来. 《SVM 和最大熵相结合的中文机构名自动识别》[D]，大连理工大学，2006.

[7] Adam L. Berger, Vincent J. Della Pietra, Stephen A. Della Pietra. 《A Maximum Entropy Approach to Natural Language Processing》[J], Computational Linguistics, Volume 22, Issue 1: 39-71.

[8] Adwait Ratnaparkhi. 《Learning to Parse Natural Language with Maximum Entropy Models》[J], Machine Learning, Volume 34, Issue 1-3, February 1999, Pages 151-175.

[9] 苗传江. 《HNC（概念层次网络）理论导论》[M]，清华大学出版社，2005.