基于角色标注的中文机构名识别*

俞鸿魁 [1,2] 张华平 [1] 刘群 [1,3]

1 中国科学院计算技术研究所软件研究室 北京 100080

2 北京化工大学信息科学与技术学院 北京 100029

3 北京大学信息学院计算机系计算语言所 北京 100871

E-mail：yhk@public3.bta.net.cn

摘要：中文机构名自动识别是命名实体识别的重点和难点，目前各种解决方案的实际效果还难以满足人们的实际需求。本文提出了一种基于角色标注的中文机构名自动识别方法，其基本思想是：根据在机构名识别中的作用，采取 Viterbi 算法对切分结果进行角色标注，在角色序列的基础上，进行字符串识别，最终实现中文机构名的识别。识别过程中我们只需要某个词作为特点角色的概率以及角色之间的转移概率。该方法的实用性还在于：这些角色信息完全可以从真实语料库中自动抽取得到。通过对大规模真实语料库的封闭测试中，该方法取得了接近90%的召回率和准确率，即使在开放测试中，准确率也高达 88%。不同实验从各个角色表明：基于角色标注的机构名识别算法是行之有效的。

关键词：中文机构名识别；未登录词识别；角色标注；Viterbi 算法

Recognition of Chinese Organization Name Based on Role Tagging

YU Hong-Kui[1,2] ZHANG Hua-Ping[1] LIU Qun[1,3]

1 Institute of Computing Technology, The Chinese Academy of Sciences, Beijing, 100080 China

2 Information science & technology college, Beijing University of Chemical Technology, Beijing, 100029 China

3 Inst. of Computational Linguistics, Peking University, Beijing, 1000871 China

E-mail: yhk@public3.bta.net.cn

Abstract: automatic recognition of organization name is emphasis and difficulty for named entity identification. Because of their inherent deficiencies, previous solutions are not satisfactory. This paper presents an approach for organization name recognition based on role tagging. That is: tokens after segmentation are tagged using Viterbi algorithm with different roles according to their functions in the generation of organization name; the possible names are recognized after sting identification on the roles sequence. During the recognition process, only the possibilities of tokens being specific roles and the transition possibilities between roles are required. The significance is that such lexical knowledge can be totally extracted from corpus automatically. In both close and open test on large realistic corpus, its recalling rate and precision is nearly 90%, and precision is nearly 88% in open test. Various experiments show that: our role-based algorithm is effective for organization recognition.

---

*本文得到国家重点基础研究项目(G1998030507-4；G1998030510)和计算所领域前沿青年基金项目20026180-23资助

作者**俞鸿魁**，男，1978 年生，北京化工大学计算机系研究生，中科院计算所客座学生，主要研究方向为计算机语言学。**张华平**,男，1978 年生, 博士研究生，主要研究方向为计算语言学，中文信息处理与信息抽取。**刘群，**男，1966 年生，在职博士研究生，副研究员，主要研究方向为机器翻译，自然语言处理与中文信息处理。

Keywords: organization name recognition; unknown words recognition; role tagging; Viterbi algorithm.

1. 引言

命名实体识别是自然语言处理中的一项基本工作，命名实体的识别也是句法分析、机器翻译、信息抽取等任务的一个非常重要的预处理模块。一般来说，命名实体识别的任务就是对于一篇待处理文本，识别出其中出现的人名（Person）、地名（Location）、机构名（Organization）、日期（data）、时间(time)、百分数(percentage)、货币(monetary value)这七类命名实体。其中命名实体中人名、地名、机构名的识别是最难识别、也最重要的三类。

对于机构名识别来说，所要识别出来的机构名主要包括股票交易所、国际组织、商业组织、公私企业、电视台或广播台、政党、宗教组织、乐队或音乐组织、政府实体、运动队、军队等等。例如“中国国际航空公司”、“北京商业银行”、“北京电影学院青年电影制片厂”、“联想集团”、“国家经委”、“中直机关工委”、“中共中央统战部”等等。

人们已经对人名和地名的识别作了非常细致的研究[1-6]，提出了各种各样的处理方法。目前人名和地名识别已经能满足人们的需求，但是机构名无论是从理论上还是从实际上，都远远达不到人们的要求。

1.1. 机构名识别的难点

对于机构名识别来说，主要的瓶颈在于存在大量的未登录机构名。未登录词在人名、地名和机构名中都占有很大一部分的比例，未登录机构名的识别比未登录人名和地名的识别要难得多，归根到底还是由机构名的自身特点所造成的：

第一，中文机构名组成方式非常复杂。机构名识别中的机构种类繁多，各类机构都有其自已独特的命名方式。例如，公私企业命名大多以地名作为开头，中间加以企业字号，如“金山”、“亿阳”等等，结尾一般都是“公司”、“集团”类的普通名词。而机关团体类的机构名则相对比较正规，一般以上级部门开头，结尾为“所”、“部”、“院”、“委”等单字。序数词在一般的机构名中很少出现，但是在军队、医院类的机构名中，序数词确占有相当大的比例。而且机构名中还嵌套的情况，机构名中包含有另一个机构名，如“北京电影学院青年电影制片厂”。

第二，机构名中含有大量的其它命名实体。在这些命名实体中，地名所占的比例最大，其中未登录地名又占了相当一部分的比例。其它命名实体的识别大大制约了机构名的识别。

第三，中文机构名用词非常广泛。通过对 1998 年 1 月人民日报语料中的 10817 个机构名所含的 19986 个词进行统计，共计 27 种词，其中名词最多（9941 个），地名其次（5023 个）。所用词如此之广泛，是命名实体中绝无仅有的。最为严重的的是，在这些词中有很大一部分词是未登录词，例如大部分的企业字号。

第四，机构名的长度极其不固定。不像中国人名，一般为两到三个字，最多不超过四个字，地名最多也只是由三到四个词组成。机构名的长度少到两个字（“北大”、“首钢”），多到几十个字（“中国人民政治协商会议第八届全国委员会常务委员会”），在人民日报的真实文本中，由十个以上的词构成的复合机构名占了相当一部分的比例。机构名称长度的不确定性，导致机构名称的边界很难确定，加大了机构名识别的难度。

第五，大多数机构名都有其简称。简称一般都是取其全称中的几个关键字或关键词，例如“联

想”、“人大”。大量的机构名简称的出现，使得本来已经十分困难的问题变得更加困难。

综上所述，机构名的这些特点，使得机构名的识别变得困难重重。

1.2. 已有的工作

命名实体识别不外乎基于规则[7]的方法、基于统计的方法以及把规则和统计相结合[8]的方法。其实在实际应用中，纯的基于统计的方法并不多，统计中或多或少引入一些规则。

机构名大多都有非常有特点的词作结尾，尤其是在特定的领域内，例如在金融领域内的机构名，大多都是以“公司”、“集团”作为结尾。金融类机构名[7]的这种表面上的规律使得人们很容易就想到使用规则的方法来识别这类机构名。虽然在封闭测试中，能达到百分之九十多的准确率和召回率，但是在开放测试中，仅能达到百分之六十多一点，远远不能满足人们的实际需求。在特定领域内尚且如此，如果把基于规则的方法推广到全领域内，其效果是可以想像的到的，可见单纯地使用规则的方法来处理这种最为复杂的命名实体是不适宜的。使用基于规则的方法之所以行不通，关键是只注意到了机构名结尾的规律性，而忽视了机构名用词的无规律性。大量未登录词作为机构名用词，使得规则系统变得无能为力，这点在开放测试中，显得尤为突出。

文献[9]提出了一个专名的一体化识别方法，从语料和专名表中统计和分析了各种专名的内部构成，其中有关机构名的有：企业字号常用字（词）、企业经营内容、企业经营内容前修饰成分、企业机构类型等属性，然后对具有各种专名属性特征的单字和多字词进行穷尽式的标注，最后用一个逆向的规则系统，使用逆向扫描、尾字激活的策略，运用 27 条规则对机构名进行识别，在小规模的语料上测试，取得了不错的效果。不过识别规则过于复杂。

在机构名识别方面，前人们还一项非常有参考价值的工作，就是文献[11]提出的采用基于类的语言模型把中文分词和命名实体识别结合在一起，其中在机构名识别上也取得了不错的成果。

在总结前人工作的基础上，本文提出了一个新的机构名识别方法——基于角色标注的方法。首先，在人名和地名识别的基础上，对机构名内部构成角色进行有选择的分类，然后采用隐马模型[12][13]，对分词结果进行机构名构成角色的标注，最后，在角色序列上进行模式串识别，并最终识别出机构名。这套识别方法，已经实际应用到我们的汉语词法分析系统（ICTCLAS）中，取得了非常好的结果。本文以下将详细介绍有关基于角色标注的机构名识别的方法，然后给出详细的系统测试数据并分析有关试验的结果，最后阐明我们的结论。

2. 基于角色标注的中文机构名自动识别方法

2.1. 中文机构名的构成角色

就组成方式上来讲，机构名比其它专有名词复杂得多。基本上，完整的机构名可以为前段（名字部分），还有后段（关键字）两部分。关键字一般为普通的名词，用词也相对集中，是机构名中唯一较有规则可循的部分。附属的名字部分似乎毫无规律可循，可能是一些常见的词，也可能是被切分成碎片的单字。

但是通过对人民日报（我们训练和测试用的语料都是采用北大标注集的人民日报语料）1998 年 1 月中的 10817 个机构名的 19986 个前段进行统计，发现它们并非毫无规律可循。在从词性上

来分，地名、专有名词、简称、机构名占有相当一部分的比例，而且在普通名词中，又有许多在机构名中经常出现的高频词（其中，“国际”、“中央”等五个高频词占全部名词的四分之一）。

机构名不仅在内部用词的词性和用词上具有一定的规律，而且中文机构名的上下文用字相对来说也比较集中，同样具有一定的规律性，机构名的上下文大多是一些连词、动词或者表示职位的名词等。如“董事长”、“经理”等。

为了充分利用机构名构成上的这些特点，我们提出了基于角色标注的中文机构名自动识别方法。根据每个字词在机构名构成中的不同作用，我们把它们分成各个不同的角色。经过对角色集选取的反复试验，我们对机构名识别制定了以下角色表：

角色	意义	例子
A	上文	*参与*亚太经合组织的活动
B	下文	中央电视台*报道*
X	连接词	北京电视台*和*天津电视台
C	特征词的一般性前缀	北京*电影*学院
F	特征词的译名性前缀	美国*摩托罗拉*公司
G	特征词的地名性前缀	交通银行*北京*分行
H	特征词的机构名前缀	*中共中央*顾问委员会
I	特征词的特殊性前缀	*中央*电视台
J	特征词的简称性前缀	*巴*政府
D	机构名的特征词	国务院侨务*办公室*
Z	非机构名成份	

表格 1 中文机构名称构成角色

例如切分结果：

“在/１９９８年/来临/之际/，/通过/中央/人民/广播/电台/向/全国/各族/人民/致以/诚挚/的/问候/和/良好/的/祝愿/！”

我们对其进行角色标注，其相应结果就应为：

“在/Z １９９８年/Z 来临/Z 之际/Z ，/Z 通过/A 中央/I 人民/I 广播/C 电台/D 向/Z 全国/Z 各族/Z 人民/Z 致以/Z 诚挚/Z 的/Z 问候/Z 和/Z 良好/Z 的/Z 祝愿/Z ！/Z”。

2.2. 角色自动标注与中文机构名的识别

中文机构名构成角色的标注类似于一人简单的词性标注过程。

我们采用的是 Viterbi 算法［9］进行角色自动标注。即：从所有可能的标注序列中优选出概率最大的标注序列作为最终标注结果。其理论及推导如下：

我们假定 W 是分词后的 Token 序列(即未登录词识别前的分词结果)，T 是 W 某个可能的角色标注序列.其中 $T^{\#}$ 为最终标注结果，即概率最大的角色序列。则有：

$W=(w_1, w_2, \dots, w_m), T=(t_1, t_2, \dots, t_m), m>0,$

$$T^{\#}=\arg\max_{T} P(T|W) \quad \text{.....E1}$$

根据贝叶斯公式，有： $P(T|W)= P(T)P(W|T)/P(W)$ E2

对于一个特定的 Token 序列来说，P(W) 是一个常数，因此根据 E1 和 E2 我们可以得到

$$T^{\#}= \arg\max_{T} P(T)P(W|T) \quad \text{E3}$$

假定 w_i 为观察值,角色 t_i 为状态值。则 W 是观察值序列，而 T 为隐藏在 W 后的状态值序列。那么，我们可以引入隐马尔科夫模型[12]来计算 P(T)P(W|T)。因此：

$$P(T)\,P(W|T) \approx \prod_{i=1}^{m} p(w_i \mid t_i)\,p(t_i \mid t_{i-1}) \quad \text{E4}$$

$$\therefore T^{\#}= \arg\max_{T} \prod_{i=1}^{m} p(w_i \mid t_i)\,p(t_i \mid t_{i-1}) \quad \text{E5}$$

$$\Leftrightarrow T^{\#}= \arg\min_{T} \{- \sum_{i=1}^{m} [\ln p(w_i \mid t_i) + \ln p(t_i \mid t_{i-1})]\} \quad \text{E6}$$

因此，角色自动标注问题就转换为求解 E5 表达式最小化的问题。利用 Viterbi 算法[12][13]就可以求解 $T^{\#}$。 该方法的其中一个优点在于可以采取 E6 对识别出来的候选机构名根据其组成部分进行最终评分。

2.3. 角色信息的自动抽取

$p(w_i|t_i)$ 和 $p(t_i|t_{i-1})$是 E5 中两个关键的角色信息参数。其中 $p(w_i|t_i)$指的是角色为 t_i 的 Token 集合中 w_i 的概率；$p(t_i|t_{i-1})$表示的是角色 t_{i-1} 到角色 t_i 的转移概率。在大规模语料库训练的前提下，根据大数定理，我们可以得到：

$p(w_i|t_i) \approx C(w_i,t_i)/C(t_i)$ E7

其中 $C(w_i,t_i)$：w_i 作为角色 t_i 出现的次数；$C(t_i)$：角色 t_i 出现的次数。

$p(t_i|t_{i-1}) \approx C(t_{i-1},t_i)/C(t_{i-1})$ E8

其中 $C(t_{i-1},t_i)$：角色 t_{i-1} 下一个角色是 t_i 的次数；

$C(w_i,t_i)$, $C(t_i)$, $C(t_{i-1},t_i)$均可以通过对已经切分标注好的熟语料库进行学习训练、自动抽取得到。

首先要对已经词性标注好的语料库进行机构名的角色标注，

例如，原始语料为：

“在/p 1 9 9 8年/t 来临/v 之际/f ，/w 通过/p [中央/n 人民/n 广播/vn 电台/n]nt 向/p 全国/n 各族/r 人民/n 致以/v 诚挚/a 的/u 问候/vn 和/c 良好/a 的/u 祝愿/vn ！/w”

经过我们的转换程序，就变为了：

“在/Z 1 9 9 8年/Z 来临/Z 之际/Z ，/Z 通过/A 中央/I 人民/I 广播/C 电台/D 向/Z 全国/Z 各族/Z 人民/Z 致以/Z 诚挚/Z 的/Z 问候/Z 和/Z 良好/Z 的/Z 祝愿/Z ！/Z”。

再对角色序列进行训练，最终得到机构名的角色字典和各个角色之间的角色转移概率。在角色训练的过程中，将角色不是 Z 的词 wi 存入机构名识别词典，并统计 wi 作为 ti 的出现次 C(wi,ti)。同时累计所有不同角色的出现次数 C(ti)以及相邻角色的出现次数 C(ti-1,ti)。

2.4. 自动识别的最终实现

识别的过程就是在已经角色标注好的序列上进行的。识别的最大的特点就是无须复杂的规

则，而且高效准确。识别的策略就是找出满足“[CFGHIJ]D”的子串。

角色标注好的文本一般如下：

“在/Z １９９８年/Z 来临/Z 之际/Z ，/Z 我/Z 十分/Z 高兴/Z 地/Z 通过/A 中央/I 人民/I 广播/C 电台/D 、/X 中国/G 国际/I 广播/C 电台/D 和/X 中央/I 电视台/D ，/B 向/Z 全国/Z 各族/Z 人民/Z ，/Z 向/Z 香港/Z 特别/Z 行政区/Z 同胞/Z 、/Z 澳门/Z 和/Z 台湾/Z 同胞/Z 、/Z 海外/Z 侨胞/Z ，/Z 向/Z 世界/Z 各国/Z 的/Z 朋友/Z 们/Z ，/Z 致以/Z 诚挚/Z 的/Z 问候/Z 和/Z 良好/Z 的/Z 祝愿/Z ！/Z”。

应用上述的策略，识别出的潜在机构名为“中央人民广播电台”、“中国国际广播电台”以及“中央电视台”。还要根据机构名自身概率的大小对结果进行筛选，最后才得出最后的结果。

3. 试验结果与分析

3.1. 有关角色集的选取对机构名识别效果影响的试验

我们对机构名角色集的选取并不是主观臆断的，是经过我们不断筛选测试而得的。

测试一：我们仅把机构名内部用词分出特征词（D）和特征词前缀两个角色，而不把特征词前缀再细分类，统一当作是一般特征词前缀（C），角色集的其它成员为上文（A）、下文（B）以及其它成份（Z）。对《人民日报》一月的部分语料进行封闭测试，结果如下：

语料	TOTAL	FOUND	RIGHT	P(%)	R(%)	F(%)
人民日报 1 月	7836	9699	6407	66.1	81.8	73.1
人民日报 6 月	9065	11632	7249	62.3	80.0	70.0

表格 2 测试一的试验数据

注：(1)TOTAL：语料中所有的机构名数; FOUND：系统识别出的机构名数; RIGHT：系统识别正确的机构名数

(2)P：机构名识别的正确率=RIGHT/FOUND×100%; R：召回率=RIGHT/TOTAL×100%;F：综合指标=2×P×R/（P+R）×100%

在角色标注好的序列中，我们发现许多机构名内部成分被标注为非机构名成分。对于一些作为机构名内部成份出现次数相对较少的词来说，$p(w_i|t_i=C)$非常小，$p(w_i|t_i=Z)$相对来说比较大，而$p(t_i=Z|t_{i-1}=Z)$与 $p(t_i=C|t_{i-1}=C)$又相当无几。我们初步认为是由于角色分得过粗所造成的。

测试二：为了验证测试一的假设，我们将特征词前缀初步细化为地名性特征词前缀、译名性特征词前缀、机构名性特征词前缀、特殊特征词前缀以及一般特征词前缀五类。

测试结果如下：

语料	TOTAL	FOUND	RIGHT	P(%)	R(%)	F(%)
人民日报 1 月	7836	8167	6162	75.4	78.6	77.0
人民日报 6 月	9065	9736	6841	70.3	75.5	72.8

表格 3 测试二的试验数据

测试结果显示，正确率大幅度提高，整体性能也有了不小的提升。以“北京商业银行”为例，细化出一个角色，虽然 p(C|G)相对原来的 p(C|C)要小，但是 p(北京|G)的概率比原来 p(北京|C)要大得多，便得“北京商业银行”整体作为机构名的概率变大。实践初步证明，细分特缀词前缀对

提高机构名识别的效果有一定的影响。

测试三：进一步细化特征词前缀。在测试二的结果中，我们发现许多带有简称的机构名，例如“巴政府”、“美国务院”等非常简单的机构名都没有被识别出来。为此，我们特此在一般性前缀中，分化出一类简称性前缀。

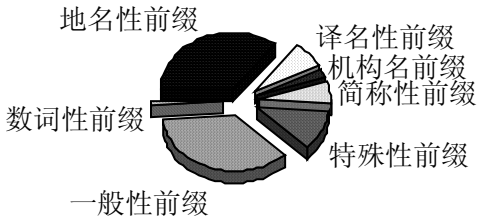
测试结果如下：

语料	TOTAL	FOUND	RIGHT	P(%)	R(%)	F(%)
人民日报 1 月	7836	8476	6317	74.5	80.6	77.5
人民日报 6 月	9065	10216	7136	69.9	78.7	74.0

表格 4 测试三的试验数据

细化特征词前缀的结果使得大量含有简称的机构名被识别出来，召回率和整体性能略有上升。进一步的细化前缀角色带来性能的进一步提升，但是是不是前缀化分得越细越好呢，为此，带着疑问我们进一步作了测试四。

测试四：我们在剩余的一般性前缀中，把所占比例最大的数词也单独化分为一类前缀，由图可见，相比如其它前缀，数词前缀所占的比例最小。



图表 1 测试四中各种前缀角色所占比例

测试结果如下：

语料	TOTAL	FOUND	RIGHT	P(%)	R(%)	F(%)
人民日报 1 月	7836	8491	6320	74.4	80.7	77.4

表格 5 测试四的试验数据

数据表明，虽然召回率有所提高，但是带来了整体性能上的下降。归其原因，主要是因为细分角色虽然使数词成为角色的概率提高，但是由于数词在所有前缀中所占比例过小，召回的少量机构名并不能弥补大量误报所带来的损失。实际结果明前缀并不是分得越细越好。

3.2. 机构名识别与人名识别和地名识别的相互影响

这次试验，我们把机构名识别集成到 ICTCLAS 中，在其它命名实体识别的基础之上进行机构名识别。

我们作封闭测试所用的训练语料是《人民日报》九八年一到六月的语料，开放测试时所用的训练语料是《人民日报》九八年二到六月的语料。封闭和开放测试时所用的测试语料都是九八年

一月的语料。

测试一：我们让系统只对人名和地名进行识别，结果如下：（注：结果中含词典中已收录的机构名）

	TOTAL	FOUND	RIGHT	P(%)	R(%)	F(%)
人名	15888	15940	15198	95.345044	95.657100	95.500817
地名	18462	23026	17736	77.025971	96.067598	85.499422
机构名	10817	4618	4072	88.176700	37.644449	52.763201

表格 6 测试一的试验数据

测试二：我们在人名和地名识别的基础之上进行基于角色标注的机构名识别，结果如下：

	TOTAL	FOUND	RIGHT	P(%)	R(%)	F(%)
人名	15888	15927	15197	95.416588	95.650806	95.533553
地名	18462	20848	17633	84.578856	95.509696	89.712541
机构名	10817	9049	7814	86.352083	72.238144	78.667069

表格 7 测试二的试验数据

结果发现，地名的总体性能大幅提升，机构名的性能相比实验一中的最好结果也有不小的提升。人名识别的性能也略有升高。

测试三：借鉴基于类的思想，我们对命名实体进行有选择的分类处理，例如，将地名识别后的结果进行分类，所有核心词典中未有的未登录地名归为未知地名类，已知地名各自为一类。经过这次改进后，

	TOTAL	FOUND	RIGHT	P(%)	R(%)	F(%)
人名	15888	15915	15198	95.494816	95.657100	95.575889
地名	18462	19009	17609	92.635068	95.379699	93.987350
机构名	10817	10520	9426	89.600760	87.140612	88.353564

表格 8 测试三的试验数据

与测试二相比，地名和机构名识别的指标又都有了大幅度的提高。

测试四：为了验证系统的实用性，我们对机构名进行开放测试。根据所掌握的材料，我们所作开放测试所选用的测试集之大，在相关的研究论文中是绝无仅有的，所以相比其它系统所得出的数据，我们的数据更有价值。

	TOTAL	FOUND	RIGHT	P(%)	R(%)	F(%)
人名	15888	15927	15199	95.429145	95.663394	95.546126
地名	18462	19930	17648	88.549925	95.590944	91.935820
机构名	10817	9279	8202	88.393146	75.825090	81.628185

表格 9 测试四的试验数据

最终结果显示，系统的整体性能依旧非常高，完全可以满足实际的需求。

4. 结论

本文系统地分析了中文机构名的特点与命名实体识别在机构名识别上的诸多难点，分析了各种典型解决方案，针对实际问题和已有方法的种种不足，同时吸收各种方法的精华，提出了一种基于角色标注的中文机构名识别方法。即采用 Viterbi 算法，利用中文机构名构成角色表及其相关统计信息，对句子中的不同成分进行角色标注，在角色序列的基础上进行字符串匹配，从而识别出中文机构名。中文机构名构成角色指的是各个分词片段在机构名识别过程中所扮演的不同角色。某个词作为特定角色的概率以及角色之间的转移概率，全部从训练语料库中自动抽取，从而降低了人工总结规则的高成本与内在缺陷。角色的标注过程就是选取角色序列概率最大的过程，避免了以前方法盲目触发的不足。通过对大规模完全真实语料库的封闭与开放测试，该方法取得了相当好的效果。各种实验表明基于角色标注的中文机构名识别算法是行之有效的。本文所用的方法虽然是一种纯统计学的方法，不过角色集合的确定却是人为的，需要引入人类的语言学知识和世界知识，不断地调试、修改角色集合，才能最终达到较好的效果。因此本文所用的方法较好地体现了经验主义和理性主义相结合的思想。

最后感谢为我们提供训练和测试语料的北大和富士通公司，感谢计算所 NLP 小组的所有成员。

参考文献

[1] 季妲，罗振声 基于反比概率模型和规则的中文姓名自动辨识系统 见：黄昌宁 张普 自然语言理解与机构翻译 北京：清华大学出版社 2001 p123-p128

[2] 何燕 基于单字词转移概率的未登录词识别见：黄昌宁 张普 自然语言理解与机构翻译 北京：清华大学出版社 2001 p141-p146

[3] 宋柔，朱宏，潘维佳，尹振海 基于语料库和规则库的人名识别法 陈力为主编《计算语言研究与应用》 北京语言学院出版社 北京 1993

[4] 吕雅娟 赵铁军等. 基于分解与动态规划策略的汉语未登录词识别 中文信息学报,VOL 15，No 1

[5] 孙茂松等， 中文姓名的自动辨识，中文信息学报，1994，No 2

[6] 沈达阳等，中国地名的自动辨识，计算语言学进展与应用，清华大学出版社，1995

[7] 王宁等, 中文金融新闻中公司名的识别，中文信息学报，VOL 16，NO 2

[8] 张艳丽 黄德根等. 统计和规则相结合的中文机构名称识别. 自然语言理解与机器翻译,清华大学出版社.2001.p233-p239

[9] 罗智勇 宋柔. 现代汉语自动分词中专名的一体化、快速识别方法.2001 国际中文电脑学术会议论文集.p323-p328

[10] 陈信希 李振昌 中文文本组织名之辨识 Communications of COLIPS, VOL 4, NO 2, DEC 1994, p131-142

[11] Sun J., Gao J. F., Zhang L., Zhou M Huang, C.N (2002). Chinese Named Entity Identification Using Class-based Language Model, Proc. of the 19th International Conference on Computational Linguistics, Taipei, pp 967-973

[12] L. R.Rabiner (1989) *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* Proceedings of IEEE 77(2): pp.257-286.

[13] L.R. Rabiner and B.H. Juang, (Jun. 1986) *An Introduction to Hidden Markov Models.* IEEE ASSP Mag., Pp.4-166.