一种基于 SVM/RS 的中文机构名称自动识别方法

宇 缨 王晓龙 刘秉权
(哈尔滨工业大学计算机学院 哈尔滨 150001)

摘 要 该文提出一种支持向量机(Support Vector Machines，SVM)和粗糙集(Rough Set, RS)相结合的中文机构名称短语识别方法。该方法借助词的基本语义搭配关系表示短语的构成规则，并通过粗糙集属性约简的方法自动学习到机构名称构成规则的无冗余集。识别时，首先寻找到与这些规则匹配的词串作为候选机构名，然后结合候选机构名以及其上下文词的语义特征，利用 SVM 分类器判断该候选是否是真正的机构名称。这种方法对 1617 万字人民日报语料开放测试的 *F* 值分别达到 82.06%。

关键词 模式识别，SVM，特征选择，语义，粗糙集，语义搭配
中图分类号：TP391.43 **文献标识码**：A **文章编号**：1009-5896(2006)05-0895-06

A Method of Automatic Recognition for Chinese Organization Name Based on SVM/RS

Yu Ying Wang Xiao-long Liu Bing-quan
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001, *China*)

Abstract A method to identify Chinese organization names by utilizing SVM (Support Vector Machines) and RS (Rough Set) is provided. Forming rule of organization name is defined based on semanteme collocation relation, and then the un-redundancy set of rough forming rules can be learned by employing attribute reduction in RS automatically. A chain of words matching forming rule is selected first as candidate, then a SVM classifier discern whether a candidate is real organization name according to candidate semanteme and its contextual semanteme while recognizing. Results of open testing achieve F-measure 82.06% in 16.17 million words news based on this project separately.

Key words Pattern recognition, SVM, Feature selection, Semanteme, Rough Set(RS), Semanteme collocation

1 引言

包括机构名称短语在内的名实体自动识别是自然语言处理中一项很重要的基础研究。目前，已经有许多这方面的识别方法，采用了如基于规则转换[1]、HMM(隐马尔可夫模型)[2,3]、最大熵方法[4]、统计语言模型[5]等许多学习方法。

本文专门讨论有关中文机构名称短语的自动识别问题，针对这类问题，王宁等[6]研究了中文金融新闻中公司名的识别，其根据公司名称的组成特征并结合上下文信息，建立了相关的六类统计知识库，通过两次扫描识别公司名，其对40篇新闻文本开放测试，准确率是62.8%，召回率是62.1%；Chen Keh-Jiann 和Chen Chao-jan [7]根据机构名称的构成特性、统计特征和语法知识识别机构名；张辉等[3]利用HMM进行机构名称的初选识别，并结合其收集建立的规则完成对侯选机构名称的识别，其对含155个机构名的2万字人民日报进行开放测试，准确率是89%，召回率是94.5%。

在这些方法中，大多要采用人工分析获得规则知识库，来识别或辅助识别机构名称，显然，这样不仅人工劳动量很大，而且也难以涵盖所有的机构名称类型。本文采用语义搭配关系表示机构名称构成规则，并通过粗糙集属性约简的方法自动获得无冗余的中文机构名称短语的构成规则。

在获得了大量机构短语(语义)构成规则的基础上，本文进而提出了一种基于 SVM 的机构名识别方法。识别的主要过程是：首先选择与机构短语构成规则匹配的词串作为候选机构名称，选择出候选机构名称之后，我们将机构名称的识别视为分类问题，另外，由于中文缺乏英文中较多的字形信息，相对而言，中文机构名称识别的难度更大一些，本文尝试借助语义信息来解决识别问题，因此，选择候选机构名称词串及其上下文词的语义作为特征空间，利用 SVM 分类器判断候选的机构名称是否真正的机构短语。

据我们所知，目前尚没有采用 SVM 进行中文机构名称短语识别方面的研究；借助语义搭配关系表示机构名称构成规则，并通过粗糙集属性约简的方法自动学习这样的短语组成规则也是新的尝试。

本文的后续内容安排是，第 2 节简要介绍本文利用的 SVM 和粗糙集算法；第 3 节介绍机构短语的语义构成规则及

2004-09-10 收到, 2005-03-31 改回
国家自然科学基金(60175020)和国家 863 计划(2002AA117010-09)资助课题

其自动学习方法；第 4 节介绍 SVM 算法如何应用于机构名称短语的识别；第 5 节是实验结果和简要分析；最后为结论。

2 SVM和粗糙集理论

2.1 一般的 SVM 算法

支撑向量机(SVM)[9]根据结构风险最小化(Structural Risk Minimization, SRM)原则，对训练样本进行优化学习，能够获得具有很好泛化能力的分类器。同时可以通过核函数变换的方法，将在低维空间无法线性分类的样本映射到高维空间进行线性分类。SVM已经在数字字符识别、基因检测、文本分类、图像识别等许多领域取得了较好的应用。

考虑一分类问题，对于给定样本点:

$$(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_i,y_i),\cdots,(\boldsymbol{x}_l,y_l),$$
$$y_i\in\{-1,+1\},\quad \boldsymbol{x}_i\in R^n$$

寻找使两类样本的分类空隙最大的最优超平面，SVM 可以表示为求解下面的二次式的最优化问题：

$$\underset{\boldsymbol{w},b,\Xi}{\text{Minimize}}\ \Phi(\boldsymbol{w},b,\Xi)=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}\xi_i^k,$$
$$\text{subject to } y_i(\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x}_i)+b)\ge 1-\xi_i,\quad \xi_i\ge 0,\ i=1,\cdots,l \qquad (1)$$

这里的参数 $C>0$ 被称为折中因子或惩罚因子，它决定了分类误差与类别间距之间的折中，分类器的泛化能力与间距分类面的间距同步变化，而所有非零的 ξ_i 决定了分类的误差。$K(x_i,x_j)=\phi(x_i)\bullet\phi(x_j)$ 称为核函数，用来实现特征从低维到高维的空间变换，常见的核函数类型有：多项式核函数、径向基函数 RBF、双曲正切核函数。

2.2 粗糙集理论

Pawlak[10]提出的粗集(Rough Set)技术作为一种新的处理模糊和不确定性知识的数学工具，在许多领域如自然语言处理、规则学习、数据挖掘获得应用，实践证明粗集技术比较擅长于处理大数据量的冗余信息消除和规则提取问题。

在粗集理论中，论域内的数据是采用信息表或者信息系统的形式来存储的，一个信息系统可以用以下四元组来定义：

$$I=(U,AV_a,f_a)_{a\in A} \qquad (2)$$

这里，U 是所有个体的非空有限集合；A 为属性的非空有限集合。对于任意一个属性 $a\in A$，对应有一个属性值的集合 V_a
以及一个信息函数 $f_a:U\to V_a$。信息函数 f_a 的作用是当给定一个个体时，可以通过该函数来确定给定个体中每个属性的值。给定个体 x 及属性 $p\in A$，$f_p(x)\in V_p$ 表示属性 p 在个体 x 中的取值，简记为 $p(x)$，对于属性子集 $P\subseteq A$，$P(x)=\{p(x):p\in P\}$。

在决策问题中，属性集合 A 一般分成两部分，即条件属性子集 C 与决策属性集 D，$C\cap D=\Phi$，$A=C\cup D$。我们把区分了条件子集和决策子集的信息表称为决策表。

粗集的一个重要作用就是适应于发现不准确数据或噪声数据内在的规律联系，即在保留原有信息系统的分类和概念表示能力不变的情况下，尽可能对信息系统中的冗余信息进行约简。对属性的约简，是粗糙集通过建立一个条件属性 $C=\{c_1,c_2,\cdots,c_n\}$ 与决策属性 $D=\{d_1,d_2,\cdots,d_m\}$ 之间对应的决策表，对决策表内条件属性 $C=\{c_1,c_2,\cdots,c_n\}$ 约简，获得一个简化不完全的决策表，仅包含决策所必须的条件属性值，但具有原决策表的全部知识。

本文对短语构成规则的提取，采用的就是这种粗糙集属性约简的方法，消除冗余信息，自动获得紧凑的规则集。

3 基于粗集理论的机构短语组成规则挖掘

3.1 机构短语的基本组成规则

根据中文的语法和习惯，机构短语一般属于“修饰性定语＋名词性中心词”一类的定中结构。其中，前缀修饰性定语可以是名词、形容词、数量词(包括形如“第＋(自然数)”形式的序数词)和动词，后缀的名词性中心词是表示机构称呼的普通名词(如：股份公司，集团，大学…),即机构短语一般构成形式是[11]：

{<地名><机构团体>}|<序数词>|{<人名>|<专有名词>}|{<产品,对象>|<功能/方式/等级>|<学科/行业>}+{机构称呼的普通名词}

可见，中文机构短语具有较为明确的规则，采用适当的方法能够自动挖掘出这些规则。

3.2 基于语义搭配关系的机构名称构成规则

虽然机构短语存在以上较为明确清晰的组成规则，但以往这些规则不仅需要大量的人为二次加工才能获得，而且难以量化被计算机自动处理。我们发现某些词搭配可以作为机构名称，而某些词搭配不符合语法习惯，不能作为机构名称，这种语法习惯表明：组成机构短语的词之间存在一种内在的逻辑关联关系，而这种关系与每个词的语义有关，只有符合一定语义关系的词组合才可能是机构名称。

因此本文尝试一种将短语组成规则映射到语义的相互关联规则，并通过粗糙集属性约简的方法，消除冗余信息，自动获得紧凑的规则集。这样，就可以大量减少人工的参与，并且这种语义表示的短语组成规则具有方便灵活的特点，方法阐述如下：

一个机构短语 $P=\{W_1,W_2,\cdots,W_n\}$ 表示P是由 $W_1W_2\cdots W_n$ 等n个单元词依次组成，而其中词 W_i 的语义定义为：$S_1\cap S_2\cap\ldots\cap S_t\to W_i$ (一般情况下，可设$t\leqslant 10$)，这里的 $S_1,S_2,\cdots,S_t$ 分别表示的是词 W_i 的基本语义或者词形特征(如数字、字母等)，W_i 的基本语义定义参照的是HowNet[12]。

如机构短语“哈尔滨工业大学”，我们可以建立如下的语义关联关系：

哈尔滨工业大学={哈尔滨,工业, 大学};

哈尔滨←地方 ∩ 市 ∩ 专有名词;

工业←事务 ∩ 工;

大学←场所∩教育;

根据以上的语义关联关系分析，我们可以获得如下形式的机构短语构成规则:

(地方∩市∩专有名词)∪(事务∩工)∪ (场所∩教育)→机构名称短语

它的含义是如果一个连续单词串$W_1W_2\cdots W_n$内的单词W_1的语义是“地方、市、专有名词”， W_2的语义是“事务、工”，W_3的语义是“场所、教育”，则该单词串$W_1W_2W_3$满足机构短语的构成规则，可以作为候选的机构短语名称。如“沈阳工业学院”、“哈尔滨林业大学”可以作为符合以上规则的机构名称，而“哈尔滨几所大学”、“参观哈尔滨工业大学”等等则显然不符合以上机构名称的构成规则。

要说明的一点是，这类规则一般是满足机构短语的必要条件，因而不能认为只要满足这类规则的连续单词串$W_1W_2\cdots W_n$肯定是机构名称，还需要进一步分析判断，才能最终确定。

3.3 机构名称的语义信息表

以上是仅通过一个机构名称样例，得到的机构短语构成规则。同样道理，多个机构名称$P_1,\cdots,P_k,\cdots,P_m$就可以构成如表1形式的机构名称语义信息表。

表1　机构名称的语义信息表

Tab.1　Table of organization name semanteme information

机构名	w_n^1	w_n^2	…	w_n^{10}	…	w_i^l	…	w_1^{10}
P_1	$S_{n,2}^1$	$S_{n,2}^1$	…	$S_{n,10}^1$	…	$S_{i,l}^1$	…	$S_{1,10}^1$
P_k	$S_{n,1}^k$	$S_{n,2}^k$	…	$S_{n,10}^k$	…	$S_{i,j}^k$	…	$S_{1,10}^k$
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋱	⋮
P_m	$S_{n,1}^m$	$S_{n,2}^m$	…	$S_{n,10}^m$	…	$S_{i,j}^m$	…	$S_{1,10}^m$

表中的w_i^l表示的是构成机构短语的第i个单词的第l个语义，P_k代表第k个机构短语名称实例， $S_{i,j}^k$填写的是具体的语义特征。这样的机构名称语义信息表，太过庞杂，使用不便，且可能存在大量的冗余信息，为了提高效率，根据RS的方法进行属性约简。

我们可以将机构名称语义信息表可以视为 RS 的决策表,其中条件属性$C=\{w_n^1,w_n^2,\cdots,w_1^{10}\}$,决策属性$D$相同(都是机构名称),因而可对该表进行属性约简，从而获得约减后的机构短语的组成规则，根据 RS 理论，这些规则具有$P_1,\cdots,P_k,\cdots,P_m$的全部知识。

4　用于机构短语识别的SVM算法和特征选择

4.1　SVM 算法的特征选择

假设在经过分词的文章中有这样的一个字串$W_pC_1C_2KC_nW_{n1}W_{n2}$，其中$C_1C_2KC_n$的语义组合满足我们得到的机构名称规则，因而被视为候选的机构名称，但仅满足这个机构名称规则，还不能确定$C_1C_2KC_n$就是真正的机构名称，还需要借助其它的特征(如上下文词)来综合判断，因此我们选择$C_1C_2KC_n$的上下文词W_p,W_{n1},W_{n2}的语义特征，作为SVM判断的上下文特征。

在计算时，特征集和采用如下的形式：

$$\text{Feature Set}=\{\text{Sem}(W_p),\text{Sem}(W_{n1}),\text{Sem}(W_{n2}),\text{Sem}(C_n),\text{Sem}(C_{n-1}),K,\text{Sem}(C_1)\}$$

其中$\text{Sem}(\phi)$代表词ϕ的语义，通过它的全部基本义原表示。

4.2　机构短语的识别算法

首先，根据学习到的机构短语构成规则，从文章内选择与这些规则匹配的单词组合的片断作为候选的机构短语名称，再根据预先学习到的 SVM 分类器判断该候选是否为真正的机构短语名称。需要说明的是，为了提高识别的效率，我们首先从寻找机构短语名称的右边界，即中心词开始，另外在本文中我们定义最长的机构短语名称由 10 个词单元组成。

具体的算法描述如下：

Input:要判断的文本

Output:地名

(1) 读入文本，并分词;

(2) 初始化，NowPosition=文本的结束位置，从文本的末尾向前依次扫描;

(3) 从当前位置 NowPosition 开始，寻找文本内可能的机构短语中心词 FocusWord，FocusWord 在文本中位置作为候选机构短语的右边界 RP;

(4) 计算 RP 前的第一个句首位置 CP;

(5) 确定 RP 前第 10 个词 fWord, 它在文本内的位置是 fP;

(6) If RP – CP≥10,则候选机构短语的左边界 LP=fP, else LP=CP;

(7) 获取 LP 与 RP 内词串的语义 semantic chain;

(8) If (semantic chain 满足机构短语组成规则),

该词串可作为候选机构短语，转到(9);

else

LP=LP−1，转到(7)继续判断，直至 LP=RP;

(9) 应用 SVM 分类器判断该候选机构短语是否真正的地名，并输出机构名称，并设定 NowPosition;

不是机构名称，则重新设定 LP，转到(7);

(10)如果 NowPosition 已经是文本的开始位置，则结束;

否则，转到(3), 继续。

如识别句子“西方将采取一切手段来确保联合国武器核查小组自由进入伊拉克武器基地检查大规模杀伤性武器。”经过分词后得到 “…确保/联合国/武器/核查/小组/自由/进入/…”，其中词串“联合国/武器/核查/小组/”与我们定义的机构名构成规则匹配，因而被确定为机构名候选，同时它的上下文词分别是“确保、自由、进入”,这样根据HowNet的语义定义，候选词串内的词“联合国、武器、核查、小组

”以及上下文词“确保、自由、进入”的全部基本语义集合，就构成了识别机构名称的特征空间，这样借助学习到的SVM分类器最终判断出“联合国武器核查小组”是一个机构名称。

5 实验结果

我们选择的实验语料是 1998 年 1 – 6 月的人民日报，其中训练语料是 2 个月大约 780 百万字，分别有未收录的 5851 个机构名称短语；测试语料是 4 个月大约 1617 万字，分别有未收录的 37409个机构名称短语。

实验的目的是检验算法对这些未收录的机构名称短语的识别能力，考察的各项指标定义是：

$$\text{正确率}\quad \text{precision}=\frac{\text{识别正确的机构名称数}}{\text{识别出的机构名称数}}$$

$$\text{召回率}\quad \text{recall}=\frac{\text{被正确识别出的机构名称数}}{\text{语料内的全部机构名称数}}$$

$$F\text{-measure}\quad F=\frac{(\beta+1)\cdot R\cdot P}{(\beta\cdot R+P)}$$

5.1 无冗余的机构名称短语构成规则的自动学习

我们从《人民日报》语料中随机选择了 22000 条机构短语，如根据“中央人民广播电台”这条机构名称，经过正确分词为“中央/人民/广播/电台”，对其中每个词的语义进行判断，我们就得到如下的逆序语义特征串到机构名称的一一对应实例：

(562∩1516)∪(239∩1536)∪(1024∩1648)∪(159∩480∩608∩1164∩1316)→中央人民广播电台(机构名称)

同样的方法，对其它机构名称如“中国航空工业总公司”、“法兰西国家队”、“沈阳市电话局”等也作同样处理，得到

(183∩1062∩1683)∪(450∩1140)∪(295∩484∩1660)→中国航空工业总公司

(1∩1275)∪(295∩ 484∩1660)→法兰西国家队

(562∩ 680∩ 864∩ 1163∩ 1516)∪(295∩ 1660)→沈阳市电话局

(注：式中数字代表的是词的基本语义索引)

这几条机构名称可以构成 3.3 节定义的如下形式“机构名称语义属性表”，那么借助粗糙集属性约简方法[10]可以发现其中的冗余信息，以求得最简约的规则集。

约简得到的 4 条规则是：

(1)(562∩1516)∪(239∩1536)∪(1024∩1648)∪(159∩480∩608∩1164∩1316)

(2)(183∩1062∩1683)∪(450∩1140)∪(295∩484∩1660)

(3)(1∩1275)∪(295∩484∩1660)

(4)(562∩ 680∩ 864∩ 1163∩ 1516)∪(295∩ 1660)

规则数目与实例相同，表明在粗糙集定义的范畴内它们之间不存在冗余信息。

同样方法，可以将全部的22000条实例，建立了如表1 的语义属性表，根据 3.3 节介绍的粗糙集属性约简方法，最后获得了 8142 条机构名称的粗规则。需要说明的是，约简结果表明每个语义属性对机构名称的构成都是有用的。

5.2 SVM 分类器的学习

从人民日报的训练语料中，选择 8497 条机构名称及其上下文，作为正例；另外，选择出虽然与规则匹配但不是机构名称的 21582 条词串及其上下文，作为反例；它们构成了SVM的训练集，采用的SVM学习算法是SVM$^{\text{-light}}$[13] 。

表2 机构名称语义属性表举例

Tab.2 Sample table of organization name semanteme information

规则序号	语言索引							
1	562	1516	0	0	…	239	1536	0
2	183	1062	1683	0	…	450	1140	0
3	1	1275	0	0	…	295	484	1660
4	562	680	864	1163	…	295	1660	0

为了寻找到最优的SVM分类函数，我们通过选择不同的核函数类型及其有关参数值，根据对训练集、测试集的识别结果，从而得到理想的SVM分类器。下面给出了部分实验结果。

下面首先给出的是若干采用内积线性核函数时，不同 C 值(区间：0.005—1.75)的实验结果。

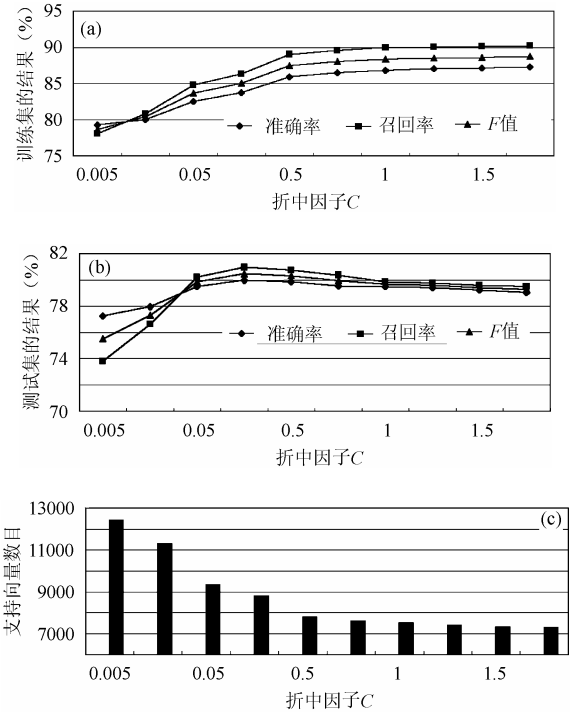

图 1 内积线性核函数时不同C值的结果

(a)训练集的实验结果(b)测试集的实验结果(c)支持向量数目的变化

Fig.1 Result under different C (dot kernel)

通过以上实验，我们发现这种内积核函数形式的SVM分类器识别机构名称短语所表现出的一些特点：

(1) 采用内积核函数时，可以控制的SVM参数是折中因子C，C的变化对训练集、测试集的识别结果影响是一致的，实验

结果的表现基本是同步的，这说明我们可以根据测试集的交叉验证结果选择最优的参数*C*；

(2) 伴随 *C* 的逐渐降低(由 1.75 到 0.0005)，支持向量的数目在增加，对应的性能指标在下降，说明支持向量数目的减少，有助于降低模型 VC 维和复杂度，从而提高泛化能力，这与结构风险最小化的原理是一致的；

(3) 伴随 *C* 值的加大，准确率和召回率的变化经历一个首先显著提高，达到最大值后，缓慢下降的过程，存在一个最佳的 *C* 值范围。

(4) 对不同类别采用不同 *C* 值的，对结果也会产生影响，因此可以根据问题需要，调整该值；

(5) 最优的 *C* 值取值范围是 0.1—0.75，这时训练集的准确率、召回率、*F*-measure 分别是 85.93%、89.04%、87.45；测试集的准确率、召回率、F-measure 分别是 79.85%、80.77%、80.30%。

我们也选择了 $K(x_i,x_j)=\exp\left(-g\left\|x_i-x_j\right\|^2\right)$ 这种 RBF 类型的核函数，重点观察改变参数 *g* 对识别结果的影响，下面给出的是，*g* 的取值在 2－0.0025 之间时的实验结果。

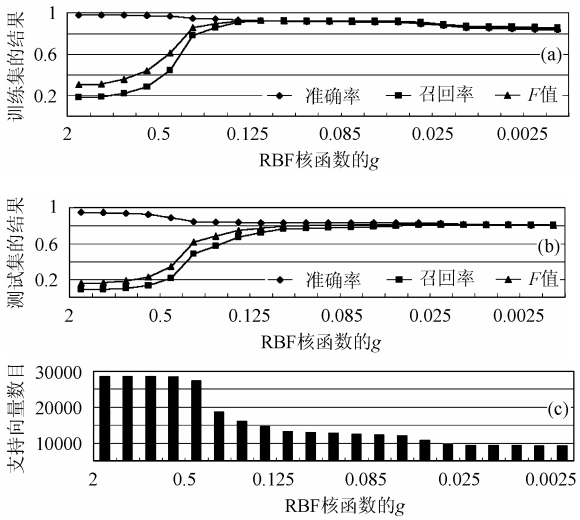

图2　RBF核函数在不同g值时的结果

(a)训练集的实验结果(b)测试集的实验结果(c)支持向量数目的变化

Fig. 2　Result under different g (RBF kernel)

我们观察到RBF核SVM分类器识别机构名称短语所表现出的一些特点：

(1) 采用RBF核函数时，可以控制的SVM参数主要是*g*，*g*的变化对训练集、测试集的识别结果影响是一致的，实验结果的表现基本是同步的，这说明我们可以根据封闭测试的结果选择最优的参数*g*；

(2) 伴随 *g* 的逐渐降低(由 2 到 0.0025)，支持向量的数目也同步降低，对应的性能指标 *F* 值在不断提高，说明支持向量数目的减少，有助于降低模型 VC 维和复杂度，从而提高泛化能力，这与结构风险最小化的原理是一致的；

(3) 伴随 RBF 核函数 *g* 的逐渐降低，导致准确率的逐步降低，但影响不明显，相对应召回率存在一个较为明显的拐点，当 *g*=0.25 时，召回率达到最大值，这之前提升明显，而后缓慢下降；

(4) 采用RBF核函数时，最佳的*g*值范围是0.15－0.025；

(5) 在采用 RBF 核函数时，随着 *C* 值的增加，使得分类的误差累计减小，准确率也随之提高，增加到一定程度之后，继续提高 *C* 值，不能明显改善识别效果，这表明，核函数保持不变的情况下，*C* 值对累计误差的控制能力是有限度的，了解了这一点，我们就可以把主要工作集中于对 *g* 值的选择上。

最后，选择的核函数是：

$$K\left(x_i,x_j\right)=\exp\left(-0.03\left\|x_i-x_j\right\|^2\right)$$

得到的这个 SVM 分类器支持向量是 9975 个，VC 维小于 8460，惩罚因子 *C*=1.5。

5.3 开放测试的结果

根据 5.2 节得到的分类器，对测试语料的实验结果是，共识别出机构名称 35107 条，其中正确的是 29756 条，错误的有 5351 条。得到的正确率是 84.75%，召回率是 79.54%;，*F*-measure($\beta=1$)是 82.06% .

5.4 识别结果分析

以下给出部分识别(下划线表示的是机构名称)的具体例句。

正确识别的例子如下：

(1) 记者去天津第一中级人民法院，旁听公开审理的天津对外经济律师事务所律师孔金荣诉天津市电话局侵权赔偿案,… 。

(2) 在 1995 年落成的法兰西国家图书馆的中间空地上，设计者还别出心裁地将诺曼底地区的一小块森林搬到了这里。

(3) 西方将采取一切手段来确保联合国武器核查小组自由进入伊拉克武器基地检查大规模杀伤性武器。

(4) 由清华同方股份有限公司承担的集装箱检测系统产品化工作通过了国家组织的专家审定。

(5) 本期国债由中央国债登记结算有限责任公司托管注册。

(6) 国际天文学联合会小行星命名委员会通报决定，将两颗永久编号为 6741、6742 的小行星以他们的名字命名。

(7) 今年 53 岁的穆勒曾经是前联邦德国国家队主力中锋。

识别错误的例子如下：

(8) 从北京百富勤投资咨询有限公司获悉：… 。

(9) 中国航空工业总公司，把考核工作同帮助领导班子加大企业内部改革力度，提高经营管理水平结合进来，… 。

(10) 荣获“中国明星企业”等荣誉称号。

未被识别出的机构名称如下：

(11) 1991 年，香港最大的彩电集团之一———讯科集团被香港录像大王瑞菱集团收购。

分析发现，识别的错误类型主要来自于 3 种情况：

机构名称的边界错误：如例句 9 中“中国航空工业总公司”是一个完整的机构名称，名称的左边界应该是地名“中国”，但识别时，得到“航空工业总公司”为一个机构名称，漏掉了前面的“中国”，导致识别结果的不完整。分析原因，具有重叠的机构名称候选短语，如“中国航空工业总公司”和“航空工业总公司”都是满足机构名称构造规则的候选短语，但最终识别的结果还受到 SVM 分类器的影响，这种影响与训练语料直接相关。

高频的机构名称用词被误认作机构名称：如例句 10 中“中国明星企业”被误认作机构名称。当上下文的语义特征不能够提供充分的分类信息时，高频的机构名称用词构成的候选词串有可能被误认作机构名称。

由于没有匹配的规则，使得该机构名称无法正确识别：我们在构造规则时，是假设词之间具有合乎语法习惯的合理语义逻辑搭配关系，而机构名称内出现的特殊专有名词会干扰语义的逻辑搭配，如果没有与它对应的规则相匹配，这些机构名称是无法识别的。如例句 11，其中机构名“讯科集团”、“瑞菱集团”因没有相应的规则可以匹配，未被识别出。这里对于“讯科集团”这类机构名称短语，如果还利用”讯、科、集团“3 个词的语义作为构成规则，不仅概括能力不够，而且也不尽合理，因此应该考虑其它类型的特征。

6 结束语

本文提出的机构名称短语识别方法，通过语义搭配关系表示机构名称构成规则，利用粗糙集属性约简的方法能够自动学习这些构成规则的无冗余集，避免了大量人工整理规则的劳动，具有精度和自动化程度较高的特点。同时基于 SVM 的识别策略利用上下文词以及候选机构名称短语的基本语义作为识别特征，将机构名称短语识别转化为一个二元分类问题，并能充分利用语义在解决这类问题时的重要帮助作用。在 1617 万字的人民日报语料上进行开放测试，实验结果显示 F 值达到 82.06%，尤其对较长的机构名称具备较高识别精度。

虽然，通过粗糙集属性约简的办法获得了较为紧凑的规则集合，但还可以从语义包含的角度出发进一步合并相关规则，提高规则的覆盖度，这项工作的完成需要依托于更准确、复杂的语义知识。

另外，本文的方法具有很好的移植性，如采用同样的粗糙集理论和语义映射方法，也可以自动学习到其它类型名词短语的组成规则，从而应用到对诸如时间、日期、会议、职位等特定形式的专有名词短语的识别。

参 考 文 献

[1] Tan Hongye, e*t al*.. Research on method of automatic recognition of Chinese place name based on transformation. *Journal of Software*, 2001, 12(11)：1608－1613.

[2] Bikel D, Schwarta R, Weischedel R. An algorithm that learns what’s in a name. *Machine Learning*, 1997, 34(1): 211－231.

[3] Yu S H, Bai S H, Wu P. Description of the Kent ridge digital labs system used for MUC-7. Proceedings of 7th Message Understanding Conference, Virginia, 1998.

[4] Borthwick A, Sterling J, Agichtein E, Grishman R. Description of the MENE named entity system as used in MUC-7. Proceedings of 7th Message Understanding Conference, Virginia, 1998.

[5] Sun Jian, Gao Jianfeng, Zhang Lei, *et al*.. Chinese named entity identification using class-based language model. In proceeding of the 19th International Conference on Computational Linguistics, Taipei, 2002.

[6] 王宁等. 中文金融新闻中公司名的识别. 中文信息学报, 2002, 16 (2): 1－6.

[7] Chen Keh-jiann, Chen Chao-jan. Knowledge extraction for identification of Chinese organization names. In proceeding of the 19th International Conference on Computational Linguistics, Taipei, 2002.

[8] 张辉等. 中国组织机构名自动识别系统的设计与实现. 电脑开发与应用, 2001, 15(1): 5－9.

[9] Cortes C, Vapnik V. Support-vector networks. *Machine Learning*, 1995, 20(3): 273－297.

[10] Pawlak Z. Rough Sets, Theoretical Aspects of Reasoning about Data. Boston: Kluwer Axademic Publishers, 1991.

[11] 刘开瑛. 中文文本自动分词和标注. 北京: 商务印书馆, 2000: 67－69.

[12] 董振东.知网(Hownet). URL: http://www.keenage.com/. 2001.

[13] Joachims T. 11 in: Making large-Scale SVM learning practical. Advances in kernel methods-Support Vector Learning, Schölkopf B and Burges C and Smola A (ed.), MIT Press, 1999.

宇 缨： 男，1968 年生，副教授，研究方向为人工智能、机器学习、模式识别、自然语言处理、支持向量机.E_mail: yuing@insun.hit.edu.cn .

王晓龙： 男，1955 年生，教授，研究方向为人工智能、自然语言处理.

刘秉权： 男，1970 年出生，副教授，研究方向为自然语言处理.