文章编号：1003-0077(2007)06-0017-05

中文组织机构名称与简称的识别

沈嘉懿[1]，李 芳[1]，徐飞玉[2]，Hans Uszkoreit[2]

（1. 上海交通大学 计算机系 上海 200240；2. 德国人工智能研究中心 语言技术实验室）

摘 要：本文提出了一种基于规则识别中文组织机构全称和简称的方法。全称的识别首先借助机构后缀词库获得其右边界，然后通过规则匹配并借助贝叶斯概率模型加以决策获得其左边界。简称的识别是在全称的基础上应用其对应的简称规则实现的。在开放性测试中，该方法的总体查全率为 85.19%，查准率为 83.03%，F Measure 为 84.10%；简称的查全率为 67.18%，查准率为 74.14%。目前该方法已应用于中文关系的抽取系统。

关键词：计算机应用；中文信息处理；组织机构名称识别；组织机构简称识别；规则匹配；贝叶斯概率模型

中图分类号：TP391 文献标识码：A

Recognition of Chinese Organization Names and Abbreviations

SHEN Jia-yi[1], LI Fang[1], XU Fei-yu[2], Hans Uszkoreit[2]

(1. Department of Computer Science and Technologyp; Shanghai JiaoTong University, Shanghai 200240, China;
2. German Research Center for Artificial Intelligence)

Abstract: This paper proposes a method for recognizing Chinese organization names and their abbreviations based on rules. The right boundary of an organization name is identified with the help of the organization suffix lexicon. The left boundary is recognized by the optimum rules based on Bayesian probability model. After idendifying an organization name, we can get candidate abbreviations based on abbreviation rules accordingly. In open test, the recall is 85.19%, the precision is 83.03%, the F Measure is 84.10% for name recognition, and the recall is 67.18%, the precision is 74.14% for abbreviation recognition. This method has been applied in the Chinese relation identification system.

Key words: computer application; Chinese information processing; recognition of Chinese organization names; recognition of Chinese organization abbreviations; rule matching; bayesian probability model

1 引言

命名实体识别是信息抽取研究的前提。命名实体主要包括人名、地名、机构名、日期、时间、百分数、货币。其中人名、地名和机构名是最重要的三类。机构泛指机关，团体或其他企事业单位，包括院校、公私企业、政府部门、院校、宗教组织、科研部门、国际组织、体育团队、音乐团体、军队等。

1.1 机构名识别的难点

1. 中文机构名的用词十分广泛，并且很大部分是未登录词，例如大部分的企业字号。

2. 中文机构名的长度极其不稳定，短到两个字，多到几十个字，这就导致了机构名称的边界难以确定。

3. 机构名中含有大量其他的命名实体，这些实体也制约了机构名的识别。

收稿日期：2006-09-14 定稿日期：2007-05-22

基金项目：本项研究工作是在中德语言技术联合实验室进行，得到了上海市科委(045107035)和德方的赞助。

作者简介：沈嘉懿(1984—)，女，研究生，研究方向为自然语言处理；李芳(1963—)，女，博士，副教授，研究方向为自然语言处理，信息检索与抽取；徐飞玉(1968—)，Dipl. Ling. Senior Software Engineer，研究方向为信息抽取，问题回答。

4. 大多数机构名称都有其简称, 简称的构成没有稳定的规则, 甚至同一机构有不同的简称, 这也使得机构名的识别变得更加困难。

1.2 机构名识别研究现状

目前, 命名实体识别对人名[1] 和地名[2] 研究得较多, 而对于机构名实体的研究较少, 主要有:

文献[3] 从机构名称的语法特征和语义特征着手, 人工分析总结出机构名称的组织规律, 进而识别机构名称。经测试, 正确率达到 97.3%, 召回率达到 96.9%。

文献[4] 针对金融领域的机构, 在识别策略上综合考虑了机构名的结构特征和文本上下文信息, 利用机器统计和人工辅助相结合的方法进行识别。在开放测试中召回率达到 62.1%, 精确率达到 62.8%。

文献[5] 将机构作为命名实体的七种类型之一考虑, 通过模式匹配进行识别, 专名识别的召回率和准确率在含有 1 117 个 NE 的测试集上为 46%和 53%, 在含有 254 个 NE 的测试集上为 17%和 29%。

本文通过分析组织机构名称的构成特征, 建立了其专属的特征词库, 并运用机器学习的方法总结出构成规则, 从而对机构全称和简称进行识别。

2 中文组织机构全称识别

2.1 中文组织机构全称特征分析

通过对中文组织机构名称的构成分析可以发现: 机构名称通常是以 X^+Y 结构出现的定名型短语, 其中 X^+ 表示一个或多个定语修饰词, 它的词性一般为名词、形容词、动词、序数词; Y 表示机构后缀, 它主要集中在"公司","集团"等一些名词, 这些词一般情况下是特定的, 有限且为数不多的, 所以可以通过列举或者训练这样一个集合来帮助识别机构名称的右边界。

要确定机构名称的左边界, 就必须确定 X^+ 的长度 L, 由于中文机构名的长度不确定, 本文采取的策略是通过对大量的语料进行分词, 词性标注后, 统计机构名称中定语修饰词的可能词性序列, 形成规则集, 并对经过分词和特征词标注后的文档进行规则匹配, 从而确定中文机构名称的左边界。

2.2 中文组织机构特征词库及规则集

本文使用的语料搜集自网上, 由包含机构名称的句子构成, 共计 1 130 句, 包含 1 500 个真实机构名称。先对语料库进行分词、词性标注, 在此基础上, 建立了如下的特征词词典和定语修饰词规则库:

1. 机构后缀库: 对组织机构名称的识别首先从确定组织机构名称的右边界开始, 即通过找到"公司","银行","集团","企业"之类的机构后缀, 得到组织机构在文中可能出现的位置。所以, 建立机构后缀库, 作为识别的触发条件。

2. 地点词库: 地点特征词对标识机构左边界有很大的帮助, 例如"上海玩具厂"等。在词性标注的基础上, 引入 Gate① 的地点词库。

3. 独立机构名称库: 有大量的组织机构名称并不包含机构后缀, 比如"欧佩克","摩托罗拉","毕马威", 通常这些机构是一些英译过来的组织机构名称。

4. 定语修饰词规则集: 根据训练语料, 建立机构名称定语修饰词序列的规则集。

5. 机构类型库: 机构类型名包括"开发","责任"等附加在机构后缀前的词, 该词库是辅助系统在机构简称识别时界定机构名关键字。

2.3 中文组织机构全称识别

组织机构名称识别的整体结构如图 1 所示。

原始文档先进行分词②, 分词过程添加了分词专用词库: 机构后缀库、地名词库、独立机构名称词库。分词后的文本已经包含对组织机构名称识别有用的词性信息、地名、机构后缀、独立机构名称等。在识别系统的核心部分"组织机构名称识别模块"中, 先通过规则匹配得到候选规则, 接着, 通过贝叶斯概率模型对候选规则进行决策, 确定最优的规则, 从而最终确定组织机构名称的左边界。

2.3.1 规则匹配

定义文中出现的机构后缀集 S, 规则集 R, 候选规则集 CR, 以及指针 $*w$ 和 $*p$, 分别指向文档中的当前的词和规则中当前的词性, 匹配过程如下:

1) 对于 S 中下一个机构后缀 s, w 指向该机构后缀前一个词;

2) 对于 R 中下一条规则 r, p 指向规则 r 中的

① http://gate.ac.uk。

② 中科院的分词系统。

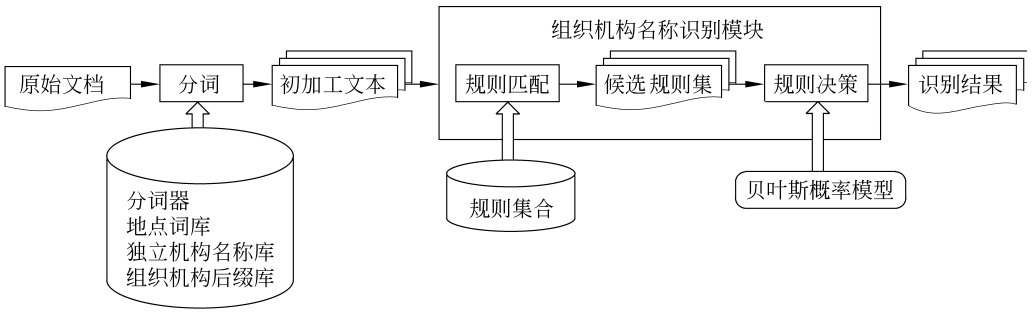

图 1　组织机构名称识别整体结构图

最后一个词性；

3）如果 w 指向的词的词性与 p 指向的词性相同，则 w——，p——；转 3）；

4）如果 p 指向 Null，则匹配成功，将 r 加入 CR，转 1）；

5）如果 p 指向非空，则匹配失败，转 1）。

2.3.2　规则决策

1）模型说明

经过上述规则匹配，同一机构后缀前可能有多条规则得到匹配，因此我们就需要借助贝叶斯概率模型决策出该机构最有可能由哪一条规则得到。

2）先验概率的确定

首先，对于每一条规则，假设它在训练集中出现次数为 n_i，可以计算该规则出现的频率：

$$P(r_i)=\frac{n_i}{\sum_{j=1}^{j=n} n_j} \tag{1}$$

其次，对每一条规则 r_i 计算其得到匹配时，组织机构名称被正确识别（即左右边界全部正确）的条件概率 $P(O|r_i)$。计算方法如下：

a）构造特征函数

$$f(j)=\begin{cases}1 & if(\text{规则 } r_j \text{ 的右半部分是 } r_i)\\ 0 & else\end{cases} \tag{2}$$

b）

$$P(O|r_i)=\frac{n_i}{\sum_{j=1}^{j=n} n_j\times f(j)} \tag{3}$$

3）最大后验概率

根据贝叶斯定理：

$$P(r_i|O)=\frac{P(O|r_i)P(r_i)}{\sum_{j=1}^{n}P(O|r_j)P(r_j)}\propto P(r_i)P(O|r_i) \tag{4}$$

以及最大后验概率假定：

$$r_{MAP}\equiv\arg\max_{r\in CR}P(r|O)$$

$$=\arg\max_{r\in CR}P(O|r)P(r) \tag{5}$$

找出使得 $P(r|O)$ 最大的规则 r_{MAP}，将其作为最终匹配的规则，获取其长度 L 后，从当前机构后缀向左回溯 L 个词，便可以得到机构名称的左边界。

2.3.3　机构名称合并

在组织机构名称中，存在这样一类情况，它是由上级机构＋下属机构或分支构成，比如“南昌市公安局西湖分局筷子巷派出所”，经过之前的处理，可以得到“南昌市公安局”，“西湖分局”，“筷子巷派出所”三个独立的机构名称。经分析，多个连续出现的机构在通常情况下存在着上下隶属关系，并且在语义上的重点也是落在最后的机构名称上，因此系统把这样连续出现的多个机构名称合并为一个。在本例中，我们最终标识出完整的一个机构名称“南昌市公安局西湖分局筷子巷派出所”，而不是三个。

3　中文组织机构简称识别

3.1　中文组织机构简称特征分析

对中文组织机构简称分析发现其构成与全称之间存在如下关系：

1. 取全称中每个词的首字如：华东师范大学——华师大；

2. 若全称中出现专有名词，取该专有名词，如：美国耐克公司——耐克；

3. 若全称以地点开始，取地点＋其他词的首字，如：上海交通大学——上海交大；

4. 取全称中除地点和机构后缀以外词的首字，如：中国南方航空公司——南方航空；

5. 取全称中除地点和机构后缀的所有词的首字，如：中国南方航空公司——南航；

6. 取除机构后缀其他词的首字＋机构后缀，

如：交通银行总部——交行总部。

如上关系都是建立在已有全称的基础上，因此在本文的研究中，简称的识别是在全称识别的基础上进行的。

3.2 中文组织机构简称识别

3.2.1 简称规则的构造

对于上述简称与全称之间的关系，本文采用了一种用数字序列来表示简称中各个字在全称中某个词中的位置，从而抽象该种关系的方法。例如：

全称："华东 师范 大学"——全称规则：ns＋n
（地名＋名词）；

1，2↓； 1↓； 1↓——简称规则：1，2；1；1

简称："华东 师 大"

全称："华东 师范 大学"

1↓； 1↓； 1↓——简称规则：1；1；1

简称："华 师 大"

于是对于全称规则 ns＋n，就得到了两条简称规则 1，2；1；1 和 1；1；1。

3.2.2 候选简称选取

设候选简称集合为 *CA*，对于每一个标识出来的机构全称：

它必定是由某条规则 *r* 得出的，对于规则 *r*，与之相应有简称规则集 *Ar*，构造方法如下：对于 *Ar* 中的每一条简称规则 *ar*，根据 *ar* 中每一个节点给出的位置信息，在全称中找出位于该位置上的字，再将所有的字相连作为候选简称，加入候选简称集合。

另外，在全称规则中，有这样一类规则，它是由地点开头，机构后缀结尾的，而规则的中间部分则是机构名关键字＋机构类型，比如"北京凯尔科技发展有限公司"，其中"科技"，"发展"，"有限"分别都是机构类型，这些机构类型是通过机构类型库识别出来的，而剩下的"凯尔"则认为是机构名关键字。通常此类机构全称的简称则包含机构名关键字(Keyword)，而地点(Loc)，机构类型(Type)，机构后缀(Suffix)则都是可选部分。于是将以下八条规则得到的候选简称加入候选简称集合：

1）loc＋keyword；

2）loc＋keyword＋type；

3）loc＋keyword＋type＋suffix；

4）loc＋keyword＋suffix；

5）keyword；

6）keyword＋type；

7）keyword＋type＋suffix；

8）keyword＋suffix。

3.2.3 简称筛选

经过候选简称的提取，得到候选的机构简称集合 *CA*，然后对该集合元素进行筛选，筛选方法如下：对于候选简称集合中的每一个候选简称，在文中搜索是否出现，若出现，就将其标识为机构简称，否则就认为这个候选简称是不存在的或者得到这个候选简称的那条简称规则是错误的。

4 实验结果与分析

4.1 实验结果

为了评估中文组织机构名称和简称的识别效果，我们从 Internet 上随机抽取了含有 654 个机构名称的 280 篇文章(含科技、体育、金融、房产、娱乐、旅游、教育题材)作为开放测试集。在不引入简称识别和引入简称识别的条件下，做了以下实验：

1. 不引入简称识别模块：

表 1 不引入简称识别模块测试结果

	Recall	Precision	F Measure
科技	85.25%	86.67%	85.95%
体育	78.18%	87.76%	82.69%
金融	87.32%	86.10%	86.71%
房产	86.00%	86.00%	86.00%
娱乐	81.82%	84.37%	83.08%
旅游	87.93%	86.44%	87.18%
教育	83.93%	85.45%	84.68%
	84.64%	86.2%	85.41%

2. 引入简称识别模块：

根据实验结果，引入简称模块后，Recall 略有提高，但是 Precision 却有了较大的降低。因为引入了简称识别，会将原本识别不出的简称识别出来，但是同样会产生错误的机构标识，从而使查准率降低。

4.2 结果分析

对实验结果分析，全称识别错误主要在于：

1）规则决策模型过于简单：在评价候选规则时没有对规则长度，规则的起始词性等因子引入其

表 2　引入简称识别模块测试结果

	机构名称识别整体			全　称　识　别			简　称　识　别		
	Recall	Precision	F Measure	Recall	Precision	F Measure	Recall	Precision	F Measure
科技	89.13%	82.00%	85.42%	94.44%	85%	89.47%	70%	70%	70.00%
体育	81.25%	81.25%	81.25%	88.46%	82.14%	85.18%	50%	75%	60.00%
金融	84.44%	82.61%	83.51%	93.94%	86.11%	89.85%	58.33%	70%	63.63%
房产	85.29%	82.86%	84.06%	88.89%	82.76%	85.72%	71.43%	83.33%	76.92%
娱乐	80.77%	84.00%	82.35%	82.35%	82.35%	82.35%	77.78%	87.5%	82.35%
旅游	85.11%	83.33%	84.21%	91.43%	84.2%	87.67%	66.67%	80%	72.73%
教育	87.50%	85.36%	86.42%	90.63%	87.88%	89.23%	75%	75%	75.00%
	85.19%	83.03%	84.10%	90.78%	84.62%	87.59%	67.18%	74.14%	70.49%

各自的权重，因此在界定左边界时产生错误。这类错误占到将近 60%。

2）机构名称不包含后缀：当机构名称不包含机构后缀，且独立机构名称库没有收录这个机构名称时，便产生此类错误。这类错误主要发生在体育类文章中，占 20%。

3）机构后缀误标识：由于不考虑上下文语义，一味地将搜索到的“机构后缀”当成真实的机构后缀，而没有考虑到有时候这个“机构后缀”只是另一个真正有意义词中的一部分，比如“电影业专业人士”中的“专业”，这类错误占 15%。

4）分词器本身存在一定的不合理，这类错误不多，仅为 5%。

简称识别错误原因主要在于：

1）简称规则集不够完善，即可能产生冗余也可能产生遗漏的情况。

2）机构全称未能被正确识别从而对简称识别结果造成影响。

3）机构类型名不像机构后缀那样特定有限，因此很难完备这样一个集合。

5　总结

本文系统地分析了中文组织机构全称与简称的特点以及识别上的诸多难点，提出并实现了一种基于规则匹配识别中文组织机构名称和简称的方法。通过对大量涉及不同领域，真实语料的测试，该方法达到了较高的查准率和查全率。基于该中文机构名称的识别，实现了中文关系的自动抽取系统[7]，目前，正开展对事件信息的抽取研究。

本文方法的改进，可以从以下方面入手：

1. 在规则决策过程中引入规则长度，规则的首词性等决策因子，通过训练得到各因子的权重，并最终用这些因子的加权和作为取舍的标准。

2. 中文机构名称的上下文用字比较集中，通常是一些连词、动词或者表示职位的名词等。如“董事长”、“经理”等。因此可以根据这些字词在机构名称构成中的不同作用，把它们分成各个不同的角色，然后训练得到角色集，最终在识别的过程中选取角色序列概率最大的情况。

参考文献：

[1] 刘秉伟，黄萱菁，郭以昆，吴立德. 基于统计方法的中文姓名识别[J]. 中文信息学报，2000，14(3)：16-24.

[2] 黄德根，岳广玲，杨元生. 基于统计的中文地名识别[J]. 中文信息学报，2003，17(2)：36-41.

[3] 张小衡，王玲玲. 中文机构名称的识别与分析[J]. 中文信息学报，1997，11(4)：21-32.

[4] 王宁，葛瑞芳，苑春法，黄锦辉，李文捷. 中文金融新闻中公司名的识别[J]. 中文信息学报，2002，16(2)：1-6.

[5] Erik Peterson. A Chinese Named Entity Extraction System[J]. http://epsilon3.georgetown.edu/petersee/Chinesee.html，1999.

[6] GATE 使用手册[EB]. http://gate.ac.uk

[7] Kebin Liu，Fang Li，et al. Embedding the semantic knowledge in convolution kernels[J]. In：the proceeding of 2nd International conference on Semantics，Knowledge and Grid (SKG 2006)，Nov. 2006.