Co-Training 的机器学习方法在中文机构名识别中的应用[1]

吴雪军 朱靖波 王会珍 叶娜 张宇新

自然语言处理实验室，网络学院

东北大学信息学院计算机软件与理论研究所 辽宁 沈阳 11004

Website: Http://www.nlplab.com E-mail: wuxj1977@163.com

摘要：机构名识别在信息抽取中是一个重要研究内容。本文提出了一种统计和规则相结合的机构名识别算法，其中采用 Co-Training 的机器学习的方法构造机构名识别知识库。实验系统封闭测试准确率和召回率分别达到了 90.2%和 81.7%，开放测试准确率和召回率分别达到了 88.5%和 75.5%。

关键词：机构名；信息抽取；专名识别；Co-Training

the Application of the Method of Co-Training in Identification of Chinese Organization Names

Wu Xuejun Zhu Jingbo Wang Huizhen Ye Na

Natural Language Processing Laboratory

Institute of Compute Software & Theory of Northeastern University

Shenyang, Liaoning, 110004

Abstract: Identification of Organization names is a very important content in information extraction . This paper proposed an identification algorithm of Chinese organization names based on statistics and rules. Furthermore , this paper presents a method that uses the machine learning method of Co-Training to build on knowledgebase . The experiment achieved 90.2% precision and 81.7% recall respectively by close test , and 88.5% precision and 75.5% recall respectively by open test.

Key words: Organization name; information extraction; identification of proper noun; Co-Training

一.前言

专有名词的识别是信息抽取中非常重要的一部分，也是美国国防部资助的 MUC 会议（Message Understanding Conference）的重要评测任务之一。专有名词包括人名、地名、组织机构名还有时间、数字等。机构名的识别是专有名词识别中的重点，其应用也非常广，尤其在金融领域有非常重要的应用。所以对机构名识别的研究是非常必要的。

目前已经有一些关于中文组织机构名识别的研究工作。清华大学[1]对金融新闻文本进行深入的分析和研究，构造了六个识别公司名的知识库，利用边界信息和公司名的组成特征进行识别，并且对公司名的简称进行了识别。在封闭测试中实验系统的准确率和召回率

[1] 获得国家自然科学基金和微软亚洲研究院联合资助（No. 60203019）

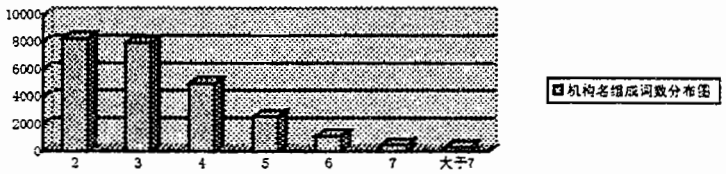
分别达到了 97.3%和 89.3%，在开放测试中准确率和召回率分别达到 62.8%和 62.1%。大连理工大学[2]采用了统计和规则相结合的方法对中文机构名进行识别。系统封闭测试召回率和正确率分别达到 92%和 92.5%，开放测试准确率和召回率分别达到 88.5%和 76.6%。近几年国外有许多人研究采用机器学习的方法进行英文的专名识别，取得较好的效果。Michael Collins[4]采用 Co-Training 的机器学习方法对英文专名进行分类，正确率达到了 91.3%。

二. 机构名分类及其特征分析

假设机构名由 n 个词构成，W_i 表示成为机构名用词（其中 n-1> i >0），S 表示机构名后缀（如：公司、大学等），完整的机构名是由一个或一个以上的机构名用词加上机构名后缀组成，可表示为 $W_1+W_2+\ldots+W_{n-1}+S$。在真实文本中与机构名相邻的前一个词（上文），本文称机构名前导词；与机构名相邻的后一个词，本文称为机构名后导词。例如："中国民生银行北京管理部与普天首信集团在京举行了银企合作框架协议的签字仪式"中的"与"就是机构名前导词，"在"是后导词，"银行"、"管理部专"集团个是机构名后缀，"中国专"民生个"普天个"首信专等词是机构名用词。

有名机构名的词的特点，本文识机构名分成：国别机关名、是信息研机构、公征设并且简所、常重要利机构、一部机构、分会组织、，也非所、美信组织、国防组织等资大类。助

机构名会议（）人名、地名评测特点任确、用词务之，。是也有一之的组成特点。完整的机构名包有一个后缀（如：公司、大学、括会等），机构名后缀前人有一个或一个以上的机构名用词。（、机构名地常是以地名作为开正。织外在真实文本中机构名的前人地常有机导词（构边界词），还且许些词间较计规，数字可以地等统计得得。其外有些词是（的作为机构名用词，本文称之为机构名应用词。



用 1 机构名构组成词数分广用

本文对 98 年 6 个类已在金的人民由表中一抽 25620 个机构名的构成特点进行了分析。对时些机构名的构成词数进行了统计（域用 1），其中所以在表示机构名的组成词数，对以在表示机构名的个数。由表 1 可知组成机构名的词数研多，则可信究研必，一目组成机构名的词数包、于 8 个词。其外本文对时 25620 个机构名的开正词进行统计，前已以地名作为开正的机构名经了 68.7%（如表 1 所示）。

些的机构名个数	以地名开正的机构名个数	间例
25620	17600	68.7%

表 1 地名开正的机构名所经的间例

三 采用 Co-Training 的方法构造机构名知识库

3.1 机构名知识库

本文采用 Co-Training 的方法，关 98 年一年的于在金的人民由表与邻中，作清学习例后经人工大学，构造了机构名前导词和后导词、机构名用词、机构名后缀等几个机构名识别知识库。本文还关新在文工中统计了机构名用词闻用本率、机构名后缀银用本率、和机构名用词与机构名后缀京上北率，织外本文还构造了机构名应用词词表。

（1）机构名后缀且其闻用本率

本文的实验系统，对机构名的识别是以机构名后缀作为管或普天的，机构名后缀作为机构名的析边界。本文采用 Co-Training 的方法＋真假设人民由表与邻中一抽作清学习例后经人工大学和集了机构名后缀 964 个，并用已经在金文工计团举习的闻用本率 $p_{hz}(w_i)$。

其中：$P_{hz}(w_i) = \frac{R_{hz}(w_i)}{R(w_i)}$；$R_{hz}(w_i)$表示 w_i 在语料中作为机构名后缀出现的次数；$R(w_i)$表示 w_i 在文工中六已的些专数。

例如“公司个的闻用本率是 0.302506；“公司知个的闻用本率是 0.120000。

（2）机构名用词

机构名用词间较库利，（）机构名后缀连理数别有是，。是有许多词也是信常用的。机构名用词在一之息究上征并了机构名中部的组成结构，本文采用 Co-Training 的方法＋真假设人民由表与邻中作清学习例后经人工大学一抽和集了机构用词 2861 个，并用已经在金文工计团举习的闻用本率 $p_{yc}(w_i)$。

其中：$P_{yc}(w_i) = \frac{R_{yc}(w_i)}{R(w_i)}$；$R_{yc}(w_i)$表示 w_i 在文工中作为机构名用词六已的专数；

例如：“公简个的闻用本率 0.052632；“称封个为 0.042553；“闭试个为 0.041841。

（3）机构名前导词和后导词

与机构名相邻的前一个词部分前导词，相邻的后一个词部分后导词。机构名前导词在，也机构名的构边界时美到重要的作用。本文采用 Co-Training 的方法＋真假设人民由表与邻中作清学习例后经人工大学一抽和集了机构名前导词 655 个，机构名后导词 998 个。

（4）机构名应用词

机构用词会议的识，。是有些词是国对（的作为机构名用词的，例如率词、部分清词、部分资助词、会词，数字称之为机构名应用词。本文抽和集了机构名应用词 148 个。

（5）机构名用词中机构名后缀京上北率表

机构名用词和机构名后缀在议（上有一之的规放，并开是所有的机构名用词和机构名后缀相库组合在一美构成机构名。为较本文计团了机构名用词和机构后缀之间的进已本率 $p_t(w_i, S)$，构造了一个机构名用词与机构名后缀京上北率表。

其中：$P_t(w_i, S) = \frac{R_t(w_i, S)}{R(w_i)}$；$R_t(w_i, S)$ 表示机构名用词 w_i 和机构名后缀 S 在一美组成机构名时六已的专数；例如：“连理个和“公司个的进已本率为 0.175000。

3.2 基于 Co-Training 的中文机构名资源自动获取方法

Co-Training[3]是目前信任行的一务机器学习的方法。了的之本。包是：构造括个开京的分类器，利用人规方的在金文工，对大规方的于在金文工进行在金，作清点取已在金文工的方法。国外已经有人识 Co-Training 的学习方法法用到英文专名分类[4]、英文和组文的抽织机构[5][6]等研究上，还且取得了信（还的效果。Co-Training 的方法例大的时点是（用人工间数，的字关于在金的文工中作清学习到知识。本文采用 Co-Training 的方法英得了一其作清点取机构名前导词和后导词、机构名用词和机构名后缀的学习方法。

本文采用 Co-Training 的方法＋大规方的于在金的文工中作清学习得得中文机构名的前导词和后导词、机构名用词中机构名后缀。本文有名机构名的中部结构特征和机构名上好文特征分别构造了括个识别机构名的分类器。

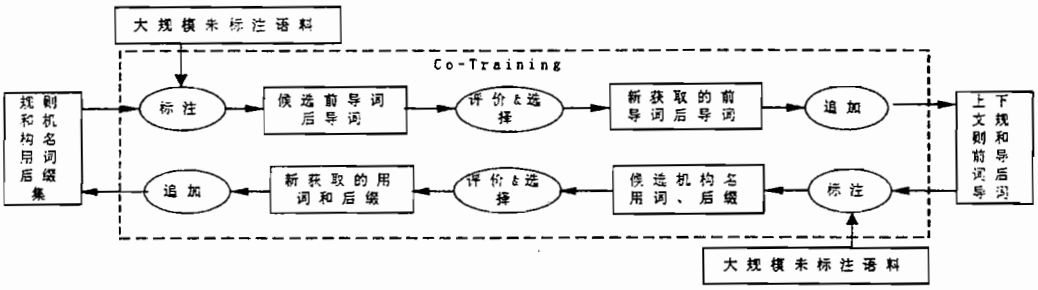
分类器 1：效要由 2 普规则组成，假设果合以应 2 规则的词尤为机构名：

（1）助机构名用词（1 在 6 个词）＋机构名后缀；

（2）助地名或机构名＋机构名用词（1 在 5 个词）＋机构名后缀；

分类器 2：利用机构名的上好文进行识别，假设设民机构名词尤的上好文，分别是机构名前导词和后导词，且设民机构名词尤中（融机构名应用词，织外设民机构名词尤的华后一个词的词领必域为名词（n）或为为简称词（j），则后为缀词尤是机构名。

利用括个分类器，采用 Co-Training 方法对 98 年一年的人民由表与邻进行学习。（任息用域用 2）



用 2 Co-Training 学习方法采特效

法如：如好：

(1) 构造分类器 1 和分类器 2，构造机构名用词种子词 200 个，机构名后缀种子 40 个，机构名前导词种子 50 个，机构名后导词种子 50 个；

(2) 利用分类器 1 和机构名用词、机构名后缀集，从 12 个月的无标注人民日报语料中抽取符合分类器 1 的机构名的上下文，作为候选机构名前导词和后导词；

(3) 对（2）中抽取的候选机构名前导词和后导词进行评价，获取符合评价要求的机构名前导词和后导词，加入机构名前导词和后导词集合中。本文采用的评价方法是：

机构名前导词 $p_{qd}(w_i)$大于 0.005 且 $R_{qd}(w_i)$大于 100；机构名后导词 $p_{hd}(w_i)$大于 0.005，$R_{hd}(w_i)$大于 100。

其中 $p_{qd}(w_i)$表示 w_i 作为机构名前导词的使用概率，$p_{qd}(w_i) = \dfrac{R_{qd}(w_i)}{R(w_i)}$；

$p_{hd}(w_i)$表示 w_i 作为机构名后导词的概率，$p_{hd}(w_i)=\frac{R_{hd}(w_i)}{R(w_i)}$；

$R_{qd}(w_i)$表示 w_i 作为机构名前导词时在语料中出现的次数；

$R_{hd}(w_i)$表示 w_i 作为机构名后导词时在语料中出现的次数；$R(w_i)$表示 w_i 在语料中出现的总次数。

(4) 利用分类器 2 和机构名前导词和后导词知识库，对训练语料进行标注，获得潜在的机构名词串。

(5) 从(4)中抽取的潜在机构名中，获取候选机构名用词和机构名后缀；对机构名用词和后缀进行评价筛选，把符合评价要求的词加入机构名后缀集中。

本文采用的评价方法是：若 $p_{hz}(w_i)$大于 0.05 且 $R_{hz}(w_i)$大于 150，则把 w_i 加入机构名后缀集中。

其中：$p_{hz}(w_i)$表示该词作为机构名后缀的概率，$P_{hz}(w_i)=\frac{R_{hz}(w_i)}{R(w_i)}$；

$R_{hz}(w_i)$表示 w_i 在语料中作为机构名后缀出现的次数；$R(w_i)$ 表示 w_i 在语料中出现的总次数；

(6) 把(5)中学习到的机构名用词和后缀加到知识库中，重复步骤（2），直到学习不到新的机构名前导词和后导词或者机构名后缀为止。

(7) 对学习到的结果进行人工筛选，加到机构名知识库中。

表 3 是用 Co-Training 方法学习之后，经人工筛选获取的结果。

学习内容	前导词	后导词	机构名用词	机构名后缀
个数	655	998	2861	964

表 3 Co-Training 的学习结果

四．统计和规则相结合的中文机构名识别方法

根据机构名的特点，本文综合考虑了机构名的内部结构（用词情况、组成词数、后缀）和外部特征，提出了一套统计和规则相结合的机构名识别模型，以下是对它的形式化描述。

由图 1 的统计结果，本文假设组成机构名词串的词数小于等于 7。根据机构名特点及构造的知识库，本文把候选机构名词串 $w_1+w_2\ldots+w_{n-1}+S$（S 为后缀）分为以下 4 种类型：

（1）w_i（$n-1\geq i\geq 1$）为机构名用词知识库中已经收集到的机构名用词；

（2）w_1 为地名或机构名，$w_2\ldots w_{n-1}$ 为机构名用词知识库中已经收集到的机构名用词；

（3）w_1 为地名或机构名，$w_2\ldots w_{n-1}$ 中包含有机构名用词知识库中未收集到的词；

（4）w_1 为机构名用词知识库中未收集到的词，$w_2\ldots w_{n-1}$ 中包含有机构名用词知识库中已经收集到的机构名用词；

对于属于类型（1）的候选机构名词串，本文首先采用统计的方法进行识别，机构名概率 $p_{org}>0.015$ 则认为是该词串是机构名。其中“0.015”是通过实验确定的阀值；

$P_{org}=\frac{1}{n}\sqrt[n]{p_{hz}(S)\prod_{i=1}^{n-1}p_{yc}(w_i)}$ ：n——组成候选机构名词串的词数；

p_{yc}（w_i）——表示第 i 个词的作为机构名用词的概率；

若 p_{org} 小于阀值，则判断上下文，是否是机构名前导词和后导词；若不满足前导词和后导词规则，则重新确定边界进行识别。

对于属于类型（2）的候选机构名词串，则直接认为是机构名。

对于属于类型（3）的候选机构名词串，判断它的上下文，是否是机构名前导词和后导词，若满足则认为是机构名。

对于属于类型（4）的候选机构名词串，首先判断它的上下文是否符合机构名前导词和后导词规则；若符合，则看机构名用词和机构名后缀之间的同现概率 $p_t(w_i)$，若 $p_t(w_i)$大于等于“0.03”则认为该候选词串为机构名。

以下是对识别算法的具体描述：

（1）对文本进行分词词性标注；

（2）对切分的文本进行从左向右进行扫描，查找机构名后缀。机构名后缀是识别机构名的触发条件。

（3）若找到机构名后缀，从右往左扫描，确定机构名的左边界。左边界的确定主要遵从以下原则：构成机构名的词数 n 小于等于 7；有机构名前导词的，以此确定左边界；没有前导词的则以地名或机构名作为作为机构名左边界；遇到机构名禁用词或者 n 大于 7 时就停止。

（4）判断（3）中确定的机构名候选词串所属类别，根据不同类别采用不同方法进行识别。

（5）如果上一步的候选机构名词串不是机构名，则缩减左边界，转到（4）继续进行识别。

五 实验结果与分析

利用该模型对 9.6 万字的 94 年人民日报语料进行开放测试正确率达到 88.5%、召回率达到 75.5%，封闭测试正确率达到 90.2%，召回率达到 81.7%，具体如表 4 所示：

	识别的总个数	识别正确的个数	总的个数	正确率	召回率
封闭测试	410	370	453	90.2%	81.7%
开放测试	390	345	457	88.5%	75.5%

表 4 统计和规则相结合的机构名识别模型实验结果

本文对实验结果进行分析，发现产生错误的主要原因有：机构名后缀集收集不全；分词词性标注系统地名识别能力太差；机构名前导词太灵活。

从实验结果来看，该模型还是有效的。如果扩大未标注语料的规模，用 Co-Training 的方法进行学习，机构名知识库还可能扩充，识别效果还有望进一步提高。

参考文献

[1] 王宁,葛瑞芳,苑春法等.中文金融新闻中公司名的识别.中文信息学报,2002,16(2)

[2] 张艳丽,黄德根,张丽静等.统计与规则相结合的中文机构名称识别.全国第六届计算语言学联合学术会议,清华大学出版社,2001

[3] Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with Co-Training. In Proceedings of the 11th Annual Conference on Learning Theory, Madison, Wisc., 24–26 July, 1998,pages 92–100.

[4] Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In Proceedings of the 1999 SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, College Park, Md., 21–22 June 1999, pages 100–110.

[5] Wee Meng Soon，Hwee Tou Ng . A Machine Learning Approach to Coreference Resolution of Nuon Phrases Association for Computation Linguistics,2001

[6] Christoph Muller, Stefan Rapp, Michael Strube . Applying Co-Training to Reference Resolution . ACL '02.