北京外国语大学中国语言文学学院 编

人文丛刊

第八辑

- 当前中文组织名识别困境与解决方案
- 探讨对外汉语词汇教学的有效方法
- 张恨水作品的北京地域文化书写及意义
- 跨文化视野下的圆积者现象研究
- 中古汉译佛经复句的文体差异

出版社

北京外国语大学中国语言文学学院 编

人文丛刊

第八辑

學苑出版社

图书在版编目（CIP）数据

人文丛刊．第8辑 / 北京外国语大学中国语言文学学院编．—北京 ：学苑出版社，2014.6
ISBN 978-7-5077-4553-5

Ⅰ．①人… Ⅱ．①北… Ⅲ．①人文科学－文集 Ⅳ．①C53

中国版本图书馆CIP数据核字(2014)第143875号

责任编辑：战葆红
封面设计：徐道会
出版发行：学苑出版社
社　　址：北京市丰台区南方庄2号院1号楼
邮政编码：100079
网　　址：www.book001.com
电子信箱：xueyuan@public.bta.net.cn
销售电话：010-67675512　67678944　67601101（邮购）
经　　销：新华书店
印 刷 厂：北京京华虎彩印刷有限公司
开本尺寸：889×1194　1/16
印　　张：24.25
字　　数：500千字
版　　次：2014年6月第1版
印　　次：2014年6月第1次印刷
定　　价：100.00元

编 委 会

目 录

语言本体

语言教学

文学文化

博士专栏

语言本体

当前中文组织名识别困境与解决方案

陈 慧

[内容摘要] 中文组织名识别可以作为命名实体识别一揽子解决方案中的一个任务，也可以作为个别解决方案中的独立任务。无论哪种方式，识别效果都取决于对实体本身的特征和实体所在上下文的特征的知识和应用。当前的中文组织名识别实验结果离实际应用还有很大距离。本文分析了造成中文组织名识别困境的多重原因。结合命名实体识别基本模式，基于对中文组织名简称识别结果的分析，初步提出分词标注系统的几个改进方向。

[关键词] 中文组织名 规则 统计 上下文信息 资源库

一、中文组织名识别的技术路线

目前命名实体识别的解决方案主要分为两种：一揽子解决方案和个别解决方案。所谓一揽子解决方案就是将所有的命名实体按照基于类的语言模型进行识别，适用于所有类型的命名实体。组织名是其中的一类识别对象；所谓个别解决方案，则是针对某一个类型的命名实体，各自建立识别模型和识别方法，实行一对一的解决方法。对于中文组织名，则是专门的中文组织名识别。无论哪种方式，命名实体识别取决于实体本身的特征和实体所在上下文的特征，不同类型的命名实体，其特征是不一样的。因此命名实体识别的做法，通常是每一类命名实体都有自己的统计模型以及特征函数。其识别方法主要有基于规则和基于统计学习两类。

在近二十年的研究中，统计语言模型在偏向应用的自然语言处理系统中起了关键的作用，统计方法似乎在整个自然语言处理领域中占据了主导地位。因而后者是近年来命名实体识别研究的主流，具体常采用隐马尔科夫模型、最大熵模型、

决策树模型、条件随机场、支持向量机等技术。但在统计学习中,规则和语言本体的价值也重新得到认识和重视。

1.1 一揽子解决方案中的组织名识别

在一揽子解决方案方面,英语命名实体识别研究主要集中在机器学习的方法上,主要包括:错误驱动的学习方法[Aberdeen,1995]、决策树的方法[Sekine,1998]、HMM[Bikel,1997]、最大熵模型[Borthwick,1998]、支持向量机[Asahara,2003]、条件随机场(Conditional Random Fields(CRF)[McCallum&Li,2003],命名实体识别任务几乎成了各种形式的有指导学习的实验沙盘。目前最好的命名实体识别系统是 MUC—7 上 Mikheev 等人开发的系统,准确率 95%,召回率 92%。

虽然汉语的命名实体识别面临的困难远远大于英语等其他语言,其识别研究仍然处于未成熟的阶段,但我国的语言信息处理专家仍在为识别精度的提高而不断努力。目前,中文命名实体识别方法多采用"统计+规则"方法,即统计模型和识别规则相互结合。代表性的中文命名实体识别方法有:

(1)规则和决策树相结合,MET—2 测试数据上 F—1 值 91%[CHUA,1998]。

(2)上下文模型和形态模型结合,应用词性、语义标记和命名实体列表,F—1 值 86.38%[Shihong Yu, Shuanhu Bai and Paul Wu, 1998]。

(3)基于类的统计模型与各种知识(包括姓氏表、同义词等)相结合,IEER—99 测试集上 F—1 值 84.61%[WU Y et al, 2003]。

(4)HMM 词性标注与具优先级别的匹配规则结合,在参加 2004 年 863 命名实体识别评测中,系统的准确率、召回率、F—1 值分别达到了 81.93%、78.20%、80.02%[向晓雯、史晓东等,2005]。

(5)NTU 系统用统计模型识别人名,用规则识别地名和组织名,MUC—7 评测 F—1 值 79.61%)[H. H. Chen,1998]。

(6)统计和词性相结合,召回率 89.9%,准确率 71.5%,F—1 值 79.65% [庄明等,2004]。

(7)基于分类,上下文模型和实体模型结合,MET—2 测试 F—值 81.79;IEER 测试 F—1 值达了 78.75%[Jian Sun et al,2002]。

姜维、王晓龙等在《人民日报》半年语料上对几种通行的命名实体识别办法进行测试,其结果如下:

表 1 几种通行命名实体识别方法测试结果

识别模型	召回率%	准确率%	F—值
最大匹配法	73.54	68.99	71.19
基本 HMM 模型	79.96	79.20	79.58

续表

识别模型	召回率%	准确率%	F一值
基本最大熵模型	83.23	84.77	83.99
HMM 模型	85.20	83.68	84.43
最大熵模型	84.62	87.95	86.25
混合模型	87.81	89.32	88.56

其中基本 HMM 模型、基本最大熵模型是未使用前缀、中缀、后缀角色的最基本分类器。而 HMM 模型、最大熵模型使用了这三种标记信息。混合模型以 HMM 模型与最大熵模型作为序列标记器。可见，混合模型的效果相对较好。

1.2 个别解决方案中的组织名识别

在个别解决方案方面，过去的一些研究主要集中于某一领域内的组织名的识别，如金融公司名、高等院校名。前者是组织名中的热点，后者结构规则较为清晰。采用的方法主要还是基于规则。基于纯的统计的方法并不多，统计中或多或少会引入一些规则。而近来的一些研究则引入了统计学习的许多模型，采用规则加统计的方法进行识别。代表性的识别方法和成果有：

1. 规则模式匹配

主要靠计算语言学家根据语言学知识经验，手工构建大量的上下文敏感推导规则构成。

(1)文献[张小衡，王玲玲等，1997]对中文组织名称尤其是中文高校名称的组成和特征进行了深入的分析，并采用基于规则的方法对中文中的高校名称进行识别，取得了很好的效果，在 600 多万字的测试集上准确率和召回率分别为 97.13% 和 96.19%。

(2)文献[王宁等，2002]综合考虑了公司名的结构特征和文本上下文信息，建立了六个用于公司名识别的知识库，公司名后缀库，公司类型名库，公司名禁止词性库，公司名完全禁止库和公司名不完全禁止库，并提出了一个基于两次扫描过程的识别策略，实验结果是正确率和召回率分别为 97.3%和 89.3%。

(3)文献[Keh-Jian Chen& Chao-jan Chen，2000]在 3 万多个新闻文本上，采用规则法研制的 NTU 系统，封闭测试的准确率、召回率、F－1 值分别为 85%、78%、81.3%；开放测试的准确率、召回率、F－1 值分别为 61.79%、54.50%、57.92%。

(4)文献[Jian Sun，2002]等基于分类，上下文模型和实体模型结合，在 IEER 测试集上准确率 76.79%，召回率 59.75%。

(5)文献[罗智勇等，2001]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,运用 27 条规则对组织名进行识别,在小规模的语料上测试,取得了不错的效果。

(6)文献[Erik Peterson,1999]将作为命名实体的七种类型之一考虑,通过模式匹配进行组织名识别,在含有 1117 个命名实体的测试集上召回率和准确率为 46%和 53%,在含有 254 个命名实体的测试集上为 17%和 29%。

组织名大多都有非常有特点的通名作右边界,组织名的这种规律使得人们很容易就想到使用规则的方法来识别这类组织名。虽然在封闭测试中,能达到百分之九十多的准确率到召回率,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。这是因为目前组织名的规则特征和领域相关,企业名、管理机关名、学校名等,其用词用字和结构规则差别很大,基于一个较小规模的随机抽取的语料获得的规则知识,往往失于片面,在一个领域内都无法全面覆盖,更无法推向其他领域。另外以往基于规则的识别系统,其规则制定一般带有较强的主观色彩,一旦调查不充分,或语言学背景不强,则规则系统的性能都会大打折扣。所以必须在一个领域覆盖全面的语料基础上针对整个组织名和不同类别的组织名制定相应的规则。

规则的制定工作量很大,而且需要我们具备深厚的语言学背景,但是一旦我们具备条件和能力,构建出了这样一个规则系统,对于组织名识别来说无疑是一个基础性的贡献。而如果我们将理性主义和经验主义结合,借助统计和计算机技术,获得这样一个规则知识,则比完全依靠简单统计和机器学习获得的规则系统的鲁棒性更强,更科学可靠。

与规则方法相比,统计方法不是由人工结构一些规则来判别命名实体,而是依赖于大规模的语料库,通过对标注语料的训练,模型从语言现象中的学习,自动获取语言学知识。与提取规则相比较,带标注语料的构建代价相对要小很多,对构建者的计算语言学的知识要求也比较低。基于统计的方法的关键在于建立一个合适的统计模型,然后利用大规模的真实语料对模型中的基本参数进行训练,另外语料库的标注质量和规模对模型的最终训练结果也有很大影响。基于统计的方法一般来说效率要比基于规则的方法高。

规则和统计方法本身均存在优缺点,但两者并不矛盾。语言规则来自于语言学家对大量语言现象的研究归纳和长期的实践检验,比完全基于语料统计出的规律具有更强的可靠性。而统计方法正可以弥补语言规则在处理例外现象方面的不足。在统计方法中,构建的模型以及模型的训练过程中往往包含了许多隐含的语言学规则。在许多规则方法中,在规则中加入概率统计信息,会比纯规则的方法更加灵活。基于这种思想,许多现有的方法都结合了规则与统计的方法,而这些规则与统计相结合的混合方法,而这些方法往往比单一的方法更有效,正确率更高。目

前，在中文组织名识别领域中融入规则知识，以统计技术为主的代表性工作主要有如下几项：

2. 基于角色标注

(1)文献[Hua-Ping Zhang& Qun Liu，2003]以 110 万词语料库为基础，基于角色标注，封闭测试的 F—1 值达到 81.63%。

(2)文献[俞鸿魁等，2003]提出了一种基于角色标注的中文组织名自动识别方法，其基本思想是：根据在组织名识别中的作用，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，最终实现中文组织名的识别。通过对大规模真实语料库的封闭测试中，他们的方法取得了接近 90%的召回率和准确率，在开放测试中，准确率在 88%左右。

3. 最大熵模型

(1)文献[冯冲等，2006]通过主动学习策略的最大熵模型训练算法，F—值达到 82%，2006)。

(2)文献[Deyi Xiong，2006]采用多层最大熵模型，准确率 82.1%，召回率 53.8%，F—1 值 65%)。

最大熵模型的优势是：在很多不同的任务领域中都能取得较高的准确率；其可以跨距离地选取特征；可以准确为变量间的细微依赖关系建模。但基于最大熵模型进行组织机构名的识别过于依赖标注数据。组织机构名的识别是一个和应用领域密切相关的任务，例如在国际新闻语料上训练出的模型用到金融领域中就很难保证识别效果。因此对标注语料的依赖已成为制约其走向真实应用的主要因素之一。

4. 条件随机场

文献[周俊生等，2006]提出了一种新的基于层叠条件随机场模型的中文组织名识别算法。对各粗分词串先在低层进行人名与地名的识别，将识别结果传递到高层模型，为高层组织名条件随机场模型对复杂组织名的识别提供决策支持，最后采用约束的前向后向算法对识别的结果进行可信度计算。在大规模真实语料的开放测试中，召回率达到 90.05%，准确率达到 88.12%。

5. 隐马尔科夫模型

文献[郑家恒、张辉，2002]用基于统计的方法，利用隐马尔科夫模型(HMM)在粗切分的基础上进行中文机构名的识别，在近 2 万字《人民日报》下载语料集上测试，结果准确率为 89%，召回率为 94.5%。从以上一些研究结果可以看出，似乎中文组织名的识别正确率已经取得了不错的结果，但如果客观地讲，有几点值得引起重视：

首先，上述一些文献对中文组织名的识别只是针对特定的领域来进行的，其规则的制定的结果的测试也是在一个小的相关的领域的语料集上进行的，如果在个

通用的领域进行识别，正确率必然会大打折扣。其次，上述的一些实验所用语料规模较小，其识别出来的组织名只有一百多个，甚至只有几十个，结果的偶然性很大，正确率中有不少随机性。因此，从这两点看来，中文组织名的识别的正确率还是被过高估计的，特别是在开放测试中，正确率和召回率往往只有 60%左右。例如，在 2004 年度国家 863 中文信息处理与智能人机接口技术评测的命名实体识别评测结果显示：中文组织名识别的召回率仅为 57.41%，准确率仅为 64.64%。这也是实验结果距离实际应用差距较大的原因。

二、中文组织名识别技术的困境分析

对于不同的语言，语言本身的特点决定导致了组织名识别方法的差异[宋柔，2001]。比如，英语中单词间有间隔，且组织名采用首字母大写，这样识别难度很小；德语中单词有间隔，但无论专名还是普通名词都一律大写，所以无法直接识别组织名；汉、泰语单词间无间隔，所以组织名的识别还受到了分词结果的制约。尤其对于汉语来说，困扰汉语自动分词的一个主要难题就是未登录词的识别，而中文组织名又是未登录词的一部分。如果文本中存在未被识别的未登录词（包括组织名），会造成难以弥补的分词错误，直接影响到汉语分词及整个句法分析的正确率。汉语较之英语有一系列难点，如：没有首字母大写这一特征、词间无空格、不同领域组织名有不同的结构、很少有专供组织名识别的语料库等等。在研究中文组织名识别时，我们可以借鉴其他一些与汉语有类似难点的语言。如，对于汉语词间无空格的特征，可以借鉴具有相同特征的泰语；还可以借鉴所有名词都大写从而区分不出组织名的德语，来识别中文组织名。

从目前国内中文命名实体识别的研究结果上看，人名和地名的识别效果要比组织名好很多，人名和地名的采用识别方法也和机构名识别有很大的不同，这和人名、地名与组织名在构词规律上的不同有很大的关系。

中文人名识别的研究是三类专有名词中开始的最早也是最集中的，所取得效果也最好，这和中文人名的构词规律有关。从历史上看，中文人名的姓氏用字是比较复杂的，如台湾出版的《中国姓氏集》收集姓氏 5544 个，其中单姓 3410 个，复姓 1990 个，三字姓 144 个。但这些姓氏到现代大部分已经不再被使用，现代中国人的姓氏趋于简单，用字相对集中，这为自动识别中文人名提供了方便。并且中文人名中，姓氏和名字用字相对集中，其概率分布符合 Zipf 定律，极少数高频姓氏和高频名字用字覆盖了大多数的人名。例如，文献[13]中从真实语料中进行统计发现：前 15 个高频姓氏的累积覆盖率达 50.8%，前 65 个高频姓氏则达到 80.4%，前 114 个就已经达到了 90%；而对于名字用字，前 71 个高频姓氏的累积覆盖率达

50.8%，前 410 个则达到 90.0%，前 1141 个的覆盖率达到了 99%。

由于中文人名构词的规律性较强，姓氏和名字用字相对集中，因此中文姓名的识别多采用概率统计加规则的方法，利用人名姓氏作为启发信息，采用这些方法进行人名识别的正确率和召回率大多数达到 90%以上。

中文地名的识别同中文人名的识别相比要更复杂。但地名相对比较固定，它总的来说是有限的，有的分词系统甚至采用地名库穷举的方法来进行地名的识别，但这种方法对于面向真实文本的系统来说，还是不大可行的。中文中的地名的构成有以下的特点：一是一部分高频出现的地名，如“北京”、“上海”等，已经包含在词典中，这部分地名在总的地名中占很大一部分；二是很大一部分地名中包含有地名特征词这样的启发信息，而这些词相对比较集中。另外，地名识别还有很多像地名库、地名词典这样的资源可以利用。因此，很多地名识别的研究也是采用概率统计加规则的方法，正确率和召回率也在 90%左右。最近的一些研究的趋势是将统计学习的方法引进到了地名识别当中包 ME，HMM，NN 和 SVM，以及一些混合模型的学习方法。组织名识别是命名实体识别任务中最困难的一部分。在 CoNLL2002 和 CoNLL2003 两次多语种命名实体识别测评中，组织名的识别效果和人名地名等相比是得分最低的。在我国 863 智能接口与技术专题的支持下，汉语的命名实体识别评测进行了三次，前两次（分别于 1995 年、2003 年举行）都是与汉语分词标注结合在一起的，2004 年单独对命名实体进行了评测。其中组织名的评测结果依然最差。

为什么中文组织名识别效果难度这么大呢？首先，中文组织名具有量大、低频、层出不穷的特点。因此组织名是未登录词的主要部分。我们不可能无限制地扩大词典规模来识别中文组织名。识别面临的中文组织名对象绝大多数为未登录词。组织名识别是命名实体识别任务中的重点。以 MUC—7 评测语料为例，组织名占实体总数的 46%，人名和地名两类实体分别仅占 22%和 32%。英文命名实体识别任务中组织结构名的比例为 79.8%；中文命名实体中组织名的比较更高达 80.9%。其次，未登录的中文组织名的识别比未登录人名地名的识别要难得多，归根到底还是由组织名的自身特点造成的：

（1）缺乏形式标记。汉语是分析型语言。书面汉语的单词基本上没有形态变化，而且一个方块字接一个方块字的文字书写方式决定了中文中的词语没有形式间隔，而汉语词汇缺乏形态标记也决定了中文组织名的识别天然的困难。而拼音文字如英语中的专有名词首字母大写和词语间隔书写都使它们的组织名识别问题变得相对容易。如：Microsoft Corporation and Lenovo Group——微软公司和联想集团。

（2）与分词任务互相影响。将文字序列切分为有意义的词语序列后，才能对词语进行词性标记。而对文字序列意义的理解也决定了文字序列的切分形式。所以

词性标记和分词实际上是互相影响的。如:重言,而非行。形式上"非行"可以表示为组织名"非洲开发银行(hang2)"的简称。而这里实际上是"非+行(xing2)"这一文言结构。意义理解和分词相互影响。

(3)中文组织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,地名最多也只是三到四个词组成。中文组织名可以是一个词(如"联想"(联想集团)),也可以是一个短语;其长度范围可以是两个字("中共"),也可以是几十个汉字(如华中科技大学同济医学院附属协和医院肿瘤科第一研究室)。在我们所考察的语料中,由十个以上的词构成的复合组织名占了相当一部分的比例。组织名长度的不确定,导致组织名称的边界很难确定,加大了组织名识别的难度。较长的中文组织名往往会被切成几个碎片,而较短的词又往往被识别为一个普通的词或被包含在一个文字串中捆绑识别错误。

(4)简称用法对识别的困扰。中文组织名都可以有全称和简称两套指称方式。而在简称方面,有时不止一种简称方式。这些简称形式灵活,或是全称中的一部分,或是全称中几个语素的组合,或是词语和语素的混合,或是字母词语。如"中央电视台",可以简称为央视、中央台、CCTV 等。"联想集团有限公司",可以简称为联想集团、联想公司、联想等。简称中通常不包含机构名称呼词等对识别有重要作用的启发信息,如"上(海)交(通)大(学)"、"全国人(民代表)大(会)"、"中(央)纪(律检查)委(员会)"等等。大量的组织名简称的出现,使得本来已经十分困难的问题变得更加困难。实际上语言经济原理的作用下,组织名简称已经是中文组织名识别任务中的主要识别对象。

(5)中文组织名用词非常广泛,和普通用字用词无异,左边界无明显特征。中科院计算所研究人员对 1998 年 1 月人民日报语料中的 10817 个组织名所含的 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。最为严重的是,在这些词中有很大一部分词是未登录词,如大部分的企业字号。中文组织名的用字用词和普通词语用字用词并无二异。如"军"、"系"、"室"等都是重要的中文组织名右边界,可以作为重要的中文组织名识别的依据之一。但实际上系统会将"王子军"、"夏利车系"、"a 栋 1309 室"等识别为中文组织名,而这样的识别错误正是由组织名字词使用和普通词语形式无二异造成的。

(6)中文组织名结构灵活,一般涉及地名、上级组织名、人名的结构嵌套。如"中国银行湖北分行洪山支行"、"北京语言大学应用语言学研究所"、"宋庆龄儿童基金会"等。这和组织的来历直接相关。显然银行是特定的某一银行的某一分行的某一支行,研究所是某一地域里某一大学某一专业学科的研究所,"基金会"是为纪念宋庆龄同志创立的。这种结构特征实际上也是组织的建立和层级关系的反

映。所以中文组织名的识别要建立在地名、人名等命名实体研究的基础上，这就决定了中文组织名的识别精度势必小于其他命名实体。

(7)“组织名的用词和结构规律和领域相关，不同领域间差别较大，所以组织名识别系统的效果严重依赖领域[李江波，2006]”，也限制了系统在不同应用领域之间的移植。各类组织都有其独特的命名方式。例如，公私企业命名大多以地名开头，中间加上企业字号，如“金山”、“全友”等等，结尾一般都是“公司”、“集团”类的普通名词。而机关团体类的组织名则相对比较正规，一般以上级部门开头，结尾为“所”、“部”、“院”、“委”等单字。序数词在一般组织名中，很少出现，但是在军队、医院、学校类组织名中，序数词却占有不小的比例。而且组织名中还有嵌套的情况，组织名中包含有另一个组织名，如“北京电影学院青年电影制片厂”。总的来说，中文组织名识别要比中文人名、地名的识别困难得多，识别的效果也和人名、地名相比有较大差距。

三、对策：加强相关语言研究与资源库建设

一个好的组织名识别模块，除了要应用成熟的统计技术，还离不开必要的语言资源的支持，包括词表和语料库。一方面，它们为开采分词系统所需要的各类知识提供了丰富的“矿藏”。如：一个常用组织名表的简单匹配可以解决待分词语料中的大部分组织名识别问题，一个组织名禁用词表可以避免一部分组织名识别错误，组织名识别所需的全局统计量需要一个好的训练语料库等等。另一方面，熟语料库又可作为测试材料对组织名识别模块的性能进行定量评估。因此，“语言资源的构造同样是机构名识别研究不可或缺的一环。[宋柔，2001]”我们先看命名实体识别的基本模式：中文命名实体识别的过程是[杨尔弘，2005]：

(1)在分词阶段，与分词任务同时进行，标注出词表中收集的命名实体；

(2)在此基础上，调用命名实体识别模型，对文中的所有命名实体进行识别。识别任务可描述如下：

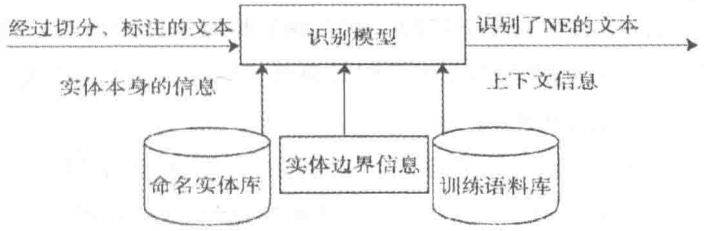

图 1　命名实体识别一般模型

在模型设计时，遵从如下原则：

(1)利用上下文信息和实体本身的信息；

(2)利用词语和词类的信息;

(3)合理使用先验知识。

上下文信息指实体所在的上下文对实体的约束作用,实体本身的信息是实体的构成信息。通过对已有资源的统计、计算,获得关于一个命名实体的特征集合,以此构造命名实体识别模型。从图 1 可见,命名实体识别模型需要三个知识库的支撑:命名实体库、实体边界信息、训练语料库。这三个知识库都是资源。命名实体库中收录使用度高的命名实体。实体边界信息库收录对识别有用的关于实体本身和实体左右边界特征的信息,如用词、用字、规则等。训练语料库则是前两个库的语料基础,是命名实体识别模型训练的基础。面向识别的资源建设不外乎这三个方面。

遗憾的是,目前中文信息处理界和语言学界在这方面的工作结合得还不够紧密。就拿中文组织名识别来说,语言学界对中文组织名的研究和关注远不能解决中文组织名识别问题。而以往识别研究者大多是技术研究人员,主要立足于技术层面,即便运用到了规则,也往往是通过统计或机器学习的方法获得,不加甄别分析地直接应用。很少有语言学专业研究者有机会或有意识投入到面向识别的组织名本体研究中去。不仅两个学科的研究实力和成果无法得到整合,组织名识别研究人员彼此的资源也得不到整合。另外,尽管以往的研究者也建设了不少资源,但总体上规模小、数据稀疏、内容主题面窄、数据陈旧。

当前中文信息处理的一个基础任务就是要建立一个大规模的,数据丰富,内容主题全面,数据历时更新的,体现了语言学研究成果的资源。如果我们在一个大规模真实语料库上对中文组织名进行全面深入的统计分析,得到一个大规模的,真实反映当前现实的中文组织名数据知识库,直接服务于中文分词和中文组织名识别,恰恰满足了当前中文信息处理的需求。这就是我们为本研究定的初始目标。总而言之,组织名识别任务面临的最大困难是面向识别的中文组织名有效资源和本体研究的缺乏。正如孙茂松教授所提到的,要继续提高组织名识别精度,我们还要使已有的组织名识别机制更加精细化;研究各种组织名与其他实体名称之间的冲突处理机制。我们以简称前后 8 个词的上下文作为校对窗口对组织名简称进行逐一校对,所发现的简称识别错误引发了我们对分词系统进一步改进的思考,因而初步提出一点分词标注系统改进意见:

首先,要解决标记集中标记符号功能的单一性、确定性问题。分词系统用 ORG 可以标记组织名和组织名简称,j 标记所有简称。使 AORG 设置冗余,标记标准不一致。

其次,以往的命名实体库往往来自已有的公司名录、人名录。但是因为它们在真实文本中的分布差异性非常大,所以面对实时更新的真实文本时捉襟见肘。如

果从一个动态更新的真实语料库上得到一个动态更新的带有统计信息的命名实体库，及时为分词标注系统提供资源，则会大大地提高分词标注的效率。因此我们可以进一步从 DCC 中构建组织名的系列知识库，为分词系统和命名实体识别服务。

再次，“观其伴，知其义”（Firth），组织名简称的标记错误也来自对上下文特征的分析不足上。比如，当“世贸”和大厦、大楼、上班、废墟等词搭配时，可以确定指的就是美国世贸大厦，当“世贸”和组织、加入、成员、上告等共现时，可以确定指的是世界贸易组织，但语料中的“世贸”会全部都标记为组织名。因此，有必要加强组织名的前后界特征和规则的统计和总结，包括领域分布和搭配信息，以尽可能地避免识别错误。因此，我们下一步将在构建组织名的系列知识库的过程中，对组织名的前后界特征、领域分布进行考察研究。

参考文献

[1]黄昌宁，赵海. 中文分词十年回顾[J]. 中文信息学报，2007(03).

[2]杨尔弘. 突发事件信息提取研究[D]. 北京语言大学博士论文，2005.

[3]周俊生，戴新宇，尹存燕，陈家骏. 基于层叠条件随机场模型的中文机构名自动识别[J]. 电子学报，2006(05).

[4]Borthwieh. ExPloiting Diverse Knowledge Sources via Maximum EntroPy in Named Entity Recognition. WVLC98.

[5]Erik Peterson A Chinese Named Entity Extraction System http://epsilon3.aeoraelow. n. edu/Peler-see/Chinesee. html，1999

[6]Hongqiao Li& ChangNing Huang. The use of SVM for Chinese new word identifieation Proeeedings of the 1st International Joint Conference on Natural Language Proeessing(IJCNLP2004)2004

[7] Hsin-His Chen. Description of the NTU System Used for METZ In Proeeedings of 7th Message Understanding Conference，1998.

[8]Huaping Zhang& Qun Liu. Chinese Named Entity Recognition Using role Model ComPutational Linguistics and Chinese Language Proeessing，Vol8，No3，2003.

[9]Keh Jiann Chen& Chao an Chen knowledge Extraction for Ideniification of Chinese Organization Names Proeeedings of ACLwork shop on Chinese LanguageProcessing2000

[10]Michael Collins& Yoram Singer Unsupervised models for named entity classification Proeeedings of the 1999 SIGDAI Conference on EmPirieal

[11]Methods in Natural Language Proeessing and Very Large Corpora Col-