命名实体识别研究*)

张晓艳　王　挺　陈火旺
(国防科技大学计算机学院　长沙410073)

摘　要　命名实体识别是文本信息处理的重要基础,已经逐步成为自然语言处理的一项关键技术。其基于规则、统计、机器学习的研究方法及成果,都推动了自然语言处理研究的发展,促进了自然语言研究与应用的紧密结合。本文回顾了命名实体识别技术的发展过程,分析了主要的方法和技术,并展望了未来的发展趋势。

关键词　命名实体识别(NER),隐马尔可夫模型(HMM),最大熵模型(ME)

Research on Named Entity Recognition

ZHANG Xiao-Yan　WANG Ting　CHEN Huo-Wang
(Department of Computer, National University of Defense Technology, Changsha 410073)

Abstract　Named Entity Recognition (NER), as the important foundation of Text Processing, has become a key technology of Natural Language Processing (NLP). It has made a great impetus to the research of NLP with its method and result. Moreover, NER has built a more effective connection between NLP theories and applications. This paper will review the history of the research, analyze the methods and technologies and give a view of the future of NER.

Keywords　Named entity recognition (NER), Hidden markov model (HMM), Maximum entropy model (ME)

1　引言

随着计算机的普及以及各种电子文本的广泛应用,海量的信息给人们的信息获取带来了严峻的挑战,人们迫切需要一些自动化工具帮助进行海量信息处理。信息抽取、信息检索、机器翻译、文摘生成等技术正是在这种背景下产生的。在这些技术中,一个共同而基础的问题就是命名实体识别(Named Entity Recognition)。命名实体识别作为这些研究中非常重要并且是必不可少的关键技术,越来越得到人们的重视和关注,时至今日已经发展成一个独立的研究分支,COLING2002就有专门的命名实体识别专题。

在一篇文章中,实体名字是基本的信息元素,往往指示了文章的主要内容。命名实体识别是对文本进行理解的前提工作。命名实体识别的质量会直接影响到后续的一系列工作,例如在信息抽取中如果没有先识别实体,根本就不可能识别实体关系;在文摘生成中,很多时候是对固定模式的填充,填充内容大都是"谁","干什么","什么时候","在哪里"等等,这正是命名实体的内容,因此从文章中获取这些内容就离不开命名实体识别;又如,在机器翻译中命名实体的翻译往往需要特殊处理。由此可见,命名实体识别已经越来越成为自然语言处理中的关键技术。因此本文就命名实体识别技术研究的历史和现状进行分析和介绍,并对未来的发展趋势作一展望。

本文首先回顾一下命名实体识别的发展和存在的问题,然后介绍命名实体识别的方法和技术,并分析了两个具体的命名实体识别系统,最后对命名实体识别的研究趋势进行了展望。

2　命名实体识别问题概述

命名实体识别最初是在MUC-6(Message Understanding Conference)上作为一个子任务提出的[1]。命名实体识别任务主要是要识别出文本中出现的专有名称和有意义的数量短语并加以归类。所谓的名字实体(Named Entity)主要包括实体(组织名、人名、地名)、时间表达式(日期、时间)、数字表达式(货币值、百分数)等。就整个的命名实体识别的研究结果而言,时间表达式和数字表达式的识别相对简单,其规则的设计、数据的统计训练等也比较容易。而对于实体中的组织名、人名、地名,因为其具有开放性和发展性的特点,而且构成规律有很大的随意性,所以其识别就可能会有较多的错选或漏选。现在大多数的命名实体识别的研究都集中于对这三种实体的识别技术的研究。

命名实体识别研究至今已经有近二十年的发展历史,已经成为自然语言处理领域的一项重要技术,并取得了很多成

*)本课题得到国家"863"高技术研究发展计划资助(2001AA114110)。张晓艳　硕士生,主要研究方向为自然语言处理;王　挺　博士,副教授,主要研究方向为自然语言处理、计算机软件;陈火旺　院士,教授,博士生导师,主要研究领域为人工智能、计算机软件。

总结　知识约简是知识发现的基础,也是Rough集理论的核心内容之一,计算所有约简已经被证明是NP完全问题。本文在分明矩阵法的基础上,给出了最小析取范式的判定定理,从而提出了计算所有属性约简的方法。理论分析和实验结果表明,该约简算法在效率上较现有的算法有显著提高。

参考文献

1　Pawlak Z. Rough sets. International Journal of Computer and Information Science, 1982, 11(5): 341～356
2　Wong S K M, Ziarko W. On optimal decision rules in decision tables. Bulletin of Polish Academy of Sciences, 1985(33): 693～696
3　Skowron A, Rauszer C. The discernibility matrics and function in information system. In: Slowinski R, ed. Intelligent Decision Support Handbook of Application and Advances of the Rough sets Theory. Dordreecht: Kluwer Academic Publishers, 1991. 331～362
4　刘清. Rough集及Rough推理. 北京:科学出版社,2001
5　张文修,吴伟志,梁吉业,李德玉. 粗糙集理论与方法. 北京:科学出版社,2001

果。其发展过程主要经历了基于规则的方法，基于统计的方法，混合方法(统计学习的方法)，下一节将按照其发展的过程详细介绍命名实体识别的技术研究。

命名实体识别在英语中已经取得了很好的研究成果，然而汉语的命名实体识别研究还处在不成熟的阶段。命名实体本身所具有的发展性和构词方式的随意性，以及各类词之间的共享性和制约性都对命名实体识别带来了一定的困难，而相对于其它语言中的命名实体识别问题，在汉语命名实体识别中又有一些特殊的难点：

·词在汉语中是个模糊的概念，没有明确的定义。即使人理解汉语也会出现边界歧义的情况，机器处理更加不可避免。分词仍然是中文信息处理的一个难题。边界模糊不仅存在于非实体词之间，也出现于实体词和非实体词之间。因此对于分词中的错误，相应地也会造成命名实体识别中的错误。另外，在命名实体识别时也会对分词结果作一些调整(主要是合并)。这样命名实体识别和分词相互交叉，使得汉语命名实体识别面临更多的问题。

·相比而言，汉语命名实体的生成规律以及结构更加复杂，尤其是缩略语的表示形式具有多样性，很难提取构成规则，因此不可能用一种识别模型应用于所有的命名实体。

·与西方语言比较，汉语缺少在命名实体识别中起重要作用的词形变换特征。英语中的这类信息能很好地指出实体的位置和边界，比如英语中的命名实体大都是以大写字母开头，而汉语并不具备这类显式的特征。我们要致力于在汉语中搜寻类似的各种有意义的潜在特征。

·汉语中除了一些比较特殊的字词外，命名实体也可以包含普通字词。事实上，几乎所有的中文字本身都可以作为一个词来使用，包括那些常用的人名用字和地名用字，这给命名实体带来了很大的困难。

·到目前为止，能用于汉语命名实体识别的开放型语料还很少，因此一方面需要开发大型命名实体标注语料库，另一方面研究不依赖大型命名实体标注文本库的算法也具有重要的意义。

评判一个命名实体是否被正确识别包括两个方面：一是实体的边界是否正确，二是实体的类型是否标注正确。前者称之为文本，后者称之为类型，由此可知文本正确，类型可能错误，反之，文本边界错误，而其包含了主要实体词且词类标记可能正确。因此对一个实体识别正确性的定义不是简单的，不同的系统侧重点不同，对其定义也可能不一样。但是一个命名实体的识别系统的识别结果不外乎下面几种情况：正确 correct(系统识别结果和标准结果相同)，丢失 missing(系统没识别而标准结果中有)，虚假 spurious(系统识别了但是标准结果中没有)。

衡量命名实体识别系统性能主要有两个评价指标：查全率和查准率。查全率是系统正确识别的结果占所有正确结果的比例；查准率是系统正确识别的结果占所有识别结果的比例。其计算公式如下：(在该公式中，把正确以外的所有情况都认为是错误的)

查全率＝count(正确)/(count(正确)＋ count(丢失))

查准率＝count(正确)/(count(正确)＋ count(虚假))

有时为了综合评价系统的性能，通常还计算查全率和查准率的加权几何平均值即 F 指数，这里查全率和查准率同等看待，权值为1，那么 F 的计算公式如下：

F ＝(2＊查准率＊查全率)/(查准率＋查全率)

介绍了命名实体识别研究存在的问题和评测标准之后，下面对命名实体识别的方法和技术进行分析和介绍。

3 命名实体识别方法

与大多数自然语言处理技术一样，命名实体识别的方法主要分为两大类：基于规则(rule-based)的方法和基于统计(statistic-based)的方法。

较早的命名实体识别方法多采用手工构造有限状态机的方法，以模式和字符串相匹配。典型的系统有用于英语命名实体识别的谢菲尔德大学的 LaSIE-II 系统[3]，爱丁堡大学的 LTG 系统[4]，这些系统主要是基于规则的。但是基于规则的方法缺乏鲁棒性和可移植性，对于每个新领域的文本都需要更新规则来保持最优性能，而这需要大量的专门知识和人力，代价往往非常大。

而基于统计的方法主要有 HMM 方法，ME 方法，决策树方法等等。在对这些方法的评价中，HMM 的性能是普遍认为比较好的，主要原因是它能较好地捕获命名实体的特征现象和位置，而且由于经典的 Viterbi 算法在求取最佳状态序列的高效性，使得 HMM 在该领域中的应用越来越频繁。但是，由于基于统计的方法获取的概率知识总是赶不上人类专家的专业知识的可靠性，而且有些知识获取必需专家的经验，因此基于统计系统的性能要比基于规则的系统性能偏低。

本节首先简要介绍一下基于规则的方法在命名实体识别中的应用，以及统计模型在命名实体识别中的应用，最后介绍统计和规则知识相结合应用的系统，即混合方法在命名实体识别中的具体应用。

3.1 基于规则和知识的方法

在基于规则的方法中，命名实体识别使用的不仅有各种命名实体的构成规则，还有实体本身和上下文的关系以及用词情况。比较简单的中文命名实体构成规则可以举例如下：

组织名→{[人名][组织名][地名][核心名]}＊[组织类型]〈指示词〉

人名 →〈姓氏〉〈名字〉

地名 →〈名字部分〉＊〈指示词〉

这是命名实体识别中最早使用的方法。这种方法的第一步要找出各种命名实体的构成规则，然后与单词(字)序列进行匹配。采取这种方法的系统主要有 NTU 系统[5]，FACILE 系统[6]，OKI 系统[7]等。

NTU 系统是一个基于规则的系统，该系统使用了不同类型的信息和模板，包括特征条件、统计信息、标题、标点符号、关键字、指示词和方向词等等。不同类型的命名实体使用的规则也不一样。具体实现中，NTU 系统，在识别人名时使用统计模型，识别地名和组织名则使用规则，在正式测试中 F-测量达到了79.61%。该系统忽略了那些低于阈值的人名，与规则不匹配的地名和组织名也没有被识别，这些是阻碍系统性能进一步提高的主要原因。另外，人名的构成规律以及组织名的嵌套和别名问题也是下一步需要解决的问题。

FACILE 系统也是一个基于规则的系统，与 NTU 不同的是，其规则形式支持上下文局部分析，并且当规则匹配发生冲突时，可通过赋给规则一个外部权值进行选择，该系统没有使用任何学习技术。该系统封闭测试结果的查全率达92%、查准率达93%，但是开放测试结果查全率下降了14个百分点，查准率下降了6个百分点。分析其原因主要是数据库条目不够充分，或者规则覆盖面有限，以及规则需要满足的条件不够完备。其中结果较差的是组织名识别，一方面因为组织名识别本身就是命名实体识别中比较困难的部分，另一方面可能是规则设计和领域联系比较紧密，导致可移植性不好。

OKI 系统是个基于规则的日语命名实体识别系统。该系

统分两个层次：首先通过一系列的名字列表生成串联规则(series rule)进行表层识别，接着在句子范围内使用结构模式规则，进行人名、地名、组织名的识别。该系统也使用了启发式信息和分析树，详细内容见参考文[7]。实验结果表明同样的方法用于英语时，效果不如日语，原因可能在于语言之间的差异，比如日语中对人名一直都有称谓，但在英语中一般仅在第一次出现时使用。

基于规则的系统往往和知识密不可分，上述几个系统都或多或少使用了一些知识，例如 NTU 中的指示词和方向词，FACILE 中的上下文知识，OKI 系统中称谓词的使用。命名实体识别中的这类知识对改善性能起到很大的作用，可以有效地减小搜索空间并增加命名实体识别的准确性。下面针对不同的实体类型，详细介绍汉语命名实体识别中各种有用的知识：

中文人名和译名：从词构成来说，中文人名的构成有一定的规律，主要有几种类型：[姓氏]＋名字、姓氏＋[职位|称谓]、阿＋名字(单字)、老(小)＋姓氏等。构成名字规则的各部分用字虽然具有任意性，但由统计结果来看还是表现出相对集中的特点(刘开瑛的书...)，因此可以建立姓氏数据库、名字数据库、职位和称谓数据库[8]。这些字词库都可以有效地缩小字词的搜索空间，减小命名实体识别的时间复杂性和空间复杂性。外国译名同样地也可以建立相应的译名用字数据库。从上下文角度看，动词、标点符号也是人名识别的有用知识。

地名：地名的主要形式是标志名＋[后缀]，因此建立地名后缀数据库是有意义的。因为大多数地名的使用还是相对固定的，所以建立一个包括国家、省、自治区、直辖市、市、县、镇，以及山脉、河流、湖泊、峡谷、海湾、岛屿等具有固定意义的地名库对地名的识别也有很大的帮助。另外文本中趋向动词的利用也有助于地名的识别[8]，比如去、往、到等。

组织名：组织名的识别是一个比较困难的问题，因为涉及到嵌套问题和别名问题。前面已经提到组织名的完全表示形式是：组织名→{[人名][组织名][地名][核心名]}*[组织类型]〈重要单词〉，由此可以看出组织名的识别应该放在前人名和地名识别之后。和地名识别类似，组织名的识别也需要建立一个后缀数据库。另外，因为一些知名组织的存在，比如国务院、中国人民银行，因此建立一个知名组织数据库可以加速组织名的识别。而在一些知识更新比较快的领域，比如计算机领域，新词或者生词出现时，往往会附带英文原词，这一特征对组织名中新词的识别很有帮助。

其它命名实体：时间表达式、数字表达式和数字的关系比较密切，那么数字集合，表示时间或数字的字词的集合会对识别有帮助，比如初、开始等。另外区间用语，节日，方向词等也是可利用的知识。

在所有类型的命名实体识别中，可以利用一些通用的小技术：局部 cache，记录当前文档中已经识别的命名实体，这不仅可以加快识别，也避免了同一个命名实体有不同的识别结果；映射关系，在已经识别的命名实体与其可能的别名之间作映射，这可以应用到人名和组织名两种实体的识别上；在训练语料中，统计各种命名实体的上下文字词，有助于确定命名实体的边界。

一般而言，当提取的规则能比较精确地反映语言现象时，基于规则的方法性能要优于基于统计的方法，但是这些规则往往依赖于具体语言、领域和文本风格，规则的设计主观性强，设计过程耗时且容易产生错误，需要富有经验的语言学家才能完成。另外，我们也看到，规则难以涵盖所有命名实体的知识，而且在语言之间的可移植性也不是很好。当从一种语料转移到另一种语料时，为保证不损失性能，往往还要花费很多工作进行规则的重新设计，性能和代价比不高。

3.2 基于统计的方法

与一般人工智能问题一样，规则知识的获取总是基于规则的方法的瓶颈。因此，人们越来越关注基于统计的方法。相比较而言，基于统计的方法利用原始或经过加工(人工标注)的语料进行训练，语料的加工(标注)也不一定需要非常广博的语言学知识，较小规模的语料也可以在可接受的时间和人力代价内完成。更有利的是，用统计方法实现的系统在移植到新的领域时可以不作或作较少的改动，只要利用新领域的语料进行训练即可。此外，由于统计方法对具体语言特性的依赖相对较少，因此基于统计的系统要移植到不同的自然语言也相对容易一些。

用于命名实体识别的的统计方法主要有 n 元模型、隐马尔可夫模型(HMM：hidden Markov model)[9,10]、最大熵模型(ME：maximum entropy model)[11]、决策树(decision tree)[12]、基于转换的学习方法[13]、推进方法[14]、表决感知器方法[14]、以及条件马尔科夫模型[15]等等。其中评价性能最好的是 HMM 模型，而最大熵模型因其自身的特点仍然是当前的主要研究方向。下面以这两个有代表性的统计模型为例，介绍一下统计模型在命名实体识别中的应用。

3.2.1 HMM 模型 HMM 模型一直在统计模型中占据着非常重要的地位，广泛应用于语音识别，词性标注等领域，这里简述一下 HMM 模型在命名实体识别中的应用：

一个隐马尔可夫模型(HMM)是一个五元组：

$(\Omega_X, \Omega_O, A, B, \pi)$

其中：$\Omega_X = \{q_1, \cdots, q_N\}$，状态的有限集合，在命名实体识别中就是 N 个词类标注，这些类型不仅包括各种命名实体类型，也包括非命名实体的各种类型。

$\Omega_O = \{v_1, \cdots, v_M\}$，观察值的有限集合，在命名实体识别中就是所有能看到的词集合，即语料生成的词典。要说明的是对应于每个状态的观察值都是个集合，对应于每个观察值的状态也不是唯一的。M 指所有可能的观察值的数目。

$A = \{a_{ij}\}, a_{ij} = p(X_{t+1} = q_j | X_t = q_i)$：转移概率，在命名实体识别中是指词类标注之间的转移概率。

$B = \{b_{ik}\}, b_{ik} = p(O_t = v_k | X_t = q_i)$：输出概率，在命名实体识别中是指词相对于标注的概率分布矩阵。

$\pi = \{\pi_i\}, \pi_i = p(X_1 = q_i)$：初始状态分布，在命名实体识别中是指一个句子第一个词类标注的概率分布。

在这里，HMM 模型的使用就是在给定观察值序列的条件下，对观察值所对应的可能的状态序列的遍历过程。在 HMM 框架下，命名实体识别已经成为了词性标注的一部分，命名实体识别的任务就是给定观察值序列(即单词序列，亦即句子)，试图找到它的最佳状态序列(即该句的标记序列)。著名的 Viterbi 算法[16]可以用来找到一个句子的最为可能的标记序列。

在众多的统计模型中，HMM 模型的评价性能比较好，其主要原因是它能较好地捕获所需要的状态转移信息，而且由于经典的 Viterbi 算法在求取最佳状态序列的高效性，使得 HMM 在该领域中的应用越来越频繁。然而，由于 HMM 模型是建立在三个假设之上：

1)马尔科夫假设(状态构成一阶马尔科夫链)，形式化表述为：$p(X_i | X_{i-1} \cdots X_1) = p(X_i | X_{i-1})$；

2)不动性假设(状态与具体时间无关)，形式化表述为：$p(X_{t+1} = q_j | X_t = q_i) = p(X_{s+1} = q_j | X_s = q_i)$ 对任意时间 s, t 成立，q_i，q_j 指状态；

3)输出独立性假设(输出仅与当前状态有关),形式化表述为:$p(O_1,\cdots,O_T|\ X_1,\cdots,X_T)=\Pi\ p(O_t|\ X_t)$;

所以,在实际应用中,上述三个假设是否符合系统实际,将会影响到系统的最终性能。

3.2.2 最大熵模型 最大熵模型是一种广泛应用于自然语言处理中的概率估计方法。它可以综合观察到的各种相关或不相关的概率知识,具有较强的知识表达能力,对文本分类、数据挖掘、词性标注等许多问题的处理结果都取得了很好的结果。

最大熵原理的基本思想是:给定训练数据及训练样本,选择一个与所有的训练数据一致的模型。例如在汉语中,对于一种命名实体,如果在训练数据中该命名实体前驱词为形容词的概率是50%,而为冠词的概率是30%,则最大熵模型在这些情况下的概率应该与训练语料中的相应概率分布一致,即也是50%和30%。换言之就是给定一些事实集,选择一种模型与现有事实一致,而对于未知事件尽可能使其分布均匀,保持对未知事件的未知状态。最大熵原理的形式化表述是:

假设存在 n 个特征 $f_i(i=1,2,\cdots,n)$,则满足所有约束的模型的集合如下:

$C=\{p\in P|p(f_i)=p'(f_i),i\in\{1,2,\cdots,n\}\}$

p' 指样本经验分布,P 指所有概率模型的集合。

由上式可以看到满足约束条件的模型并不唯一,而目标模型是在约束集下具有最均匀分布的模型,即在概率分布集合 C 中选择具有最大熵的模型。

最大熵模型在命名实体识别中的应用可表述如下:

1)特征函数的生成,在命名实体识别中,特征函数一般是个二值函数,这些函数包括词法特征、单词特征、先验特征、词典特征以及复合特征。

2)特征函数的选取,即选取对模型具有表征意义的特征。应用领域不同,观察角度不一样都会使特征函数的选取标准改变。互信息以及在训练语料中出现的频次都可以是选取的标准。在命名实体中通常是由特征函数出现的频次进行选择,需要注意的是阈值大小的确定问题。

3)参数估计,这一点是命名实体识别应用中非常重要的一步,它建立了特征和概率模型之间的联系。特征参数估计方法有传统的 GIS(General Iterative Scaling)和 IIS(Improved Iterative Scaling)迭代算法,也有为降低计算量而改进的使用 Z-测试的特征选取算法[17],这种方法把第2步和第3步结合到一起。

由上面可以看出,在命名实体识别中,构建最大熵模型应集中在模型的特征归纳上,即为模型选择具有表征意义的特征。

单独使用最大熵模型进行命名实体识别的效果并不很好,但是如果结合了其他知识,例如外部系统特征(External System Feature),即把其他系统对命名实体的识别结果也作为一个特征来考虑(在下一节介绍的第二个系统将会用到该技术)往往会取得非常好的效果。该模型的一个优势是,理论上在避免碎片化的同时可以集成任意知识源,不管这些知识是相关的或无关的,类似的或迥异的。另外最大熵模型结构紧凑,具有较好的通用性。最大熵模型的主要缺点是训练时间复杂性非常高,有时甚至导致训练代价难以承受。另外由于需要明确的归一化计算,导致 CPU 开销比较大。

4 命名实体识别系统

上一节主要描述了命名实体识别所使用的主要方法和技术,总结了可能用到的各种规则知识,下面主要介绍这些方法和知识相结合而实现的一些系统。

命名实体识别在英语中已经取得了很大的成功,最好的命名实体识别系统是 MUC7上的 Mikheev 等人开发的系统[18],其查准率达到95%,查全率达到92%。然而汉语的命名实体识别仍然处在未成熟阶段,正如在第2节中所描述的其主要原因在于语言的差异。到目前为止,已有的实验效果比较好的汉语命名实体识别系统主要有:1、NTU 系统,在识别人名时使用统计模型,识别地名和组织名是使用规则,在正式测试中 F-测量达到了79.61%;2、Shihong Yu, Shuanhu Bai and Paul Wu 等人开发的系统[19],使用了上下文模型和形态模型。但是该系统需要词性信息,语义标记和命名实体列表,该系统的 F-测量达到了86.38%;3、CHUA 等人开发的系统[20],该系统把基于模板的规则和决策树相结合,在 MET-2测试数据上 F-测量达到了91%。该系统也使用知网从语义上对相关词语分组。4、Jian Sun 等开发的系统[21],是一个用于汉语命名实体识别的基于分类的语言模型,在 MET-2测试数据上 F-测量达到81.79%,在 IEER 测试数据上 F-测量达到78.75%。但是该模型比较依赖统计信息,必须先在大规模的标记语料库上进行训练。

就命名实体识别方法而言,基于规则的方法主观性强,可移植性不好,而且歧义是语言的一个固有特点,此外规则很难覆盖所有的语言现象,语言处理需要机器具有学习能力。另一方面,人类语言的运用并不纯粹是一个随机过程,单单使用基于统计的方法将使状态搜索空间非常庞大,借助规则知识及早剪枝是一个比较有效的方法。所以目前几乎没有单纯使用统计模型而不使用规则知识的命名实体识别系统,在很多情况下使用混合方法,即统计模型+规则知识进行识别。下面我们就介绍两个使用混合方法实现的命名实体识别系统:Class-based LM 和 MENE 系统。

1. **Class-based LM**[21]

该系统是一个基于类的汉语命名实体识别模型,识别的命名实体包括人名、地名、组织名,每种命名实体分别作为一个词类。其识别过程是首先识别人名和地名,在此基础上再识别组织名。实现的语言模型主要包含两个子模型:1) 一组实体模型,每个实体模型用来估计一个中文字符串在给定实体类型下的生成概率;2) 上下文模型,用来估计一个类序列的生成概率。该模型结合 HMM 统计模型和各种有用的命名实体识别的知识,提供了一个统计框架把分词和命名实体的识别统一起来。实验结果表明该系统的性能达到了汉语命名实体识别系统中的较高水平。

在实体模型的使用中,人名识别使用了以字为单位的 N 元模型,地名使用了以词为单位的三元模型,而组织名使用了更为复杂的以类为单位的三元模型,由于在组织名的识别中人名地名作为一个词类来看待,因此组织名的识别要在人名和地名识别之后。

一般地,该系统执行包括三个步骤:候选词产生(即分词),候选实体词产生(即实体模型调用),Viterbi 算法搜索。其中,最主要的是第二步,又包括两部分工作:根据命名实体的语法结构产生候选词;使用相应的实体模型给每个候选词赋一个概率值。这里使用了两种知识:内部知识和上下文知识。内部知识用于触发生成实体候选词,比如中国姓氏表,译名用字列表。上下文知识用于计算生成概率,比如人物头衔列表,说话动作列表,上下文单词列表。而组织名候选词的产生和概率计算是利用组织名关键字列表。

根据测试,该系统的查准率分别是:人名 79.78%、地名86.02%、组织名76.79%,查全率分别是:人名 89.29%、地名

84.87%、组织名 59.75%，可以发现组织名的查全率要偏低一些，这可能是因为组织名结构比较复杂，而且存在着大量的嵌套和别名现象，因此需要一个更加精确的组织名模型。

Class-based LM 中并没有使用命名实体词典，没有考虑全局信息，组织名和缩略语、别名之间的映射关系也没有利用，而这些知识都有助于提高命名实体识别的准确性。

2. MENE 系统[2]

该系统是一个典型的使用最大熵统计模型的英语命名实体识别系统。系统使用的特征函数包括：二值特征（binary features），比如首字母大写；词汇特征，系统考虑了当前位置前后两个词的特征；区域特征，比如序，引言；分类词典特征，例如名字词典，社团词典；外部系统特征，即把外部系统对当前词及上下文的命名实体识别结果作为 MENE 系统的一个特征，这些系统包括 Proteus[22]，Manitoba 大学的系统[23]，IsoQuest 公司的系统[24]；复合特征；参考结论特征等。从上一节最大熵模型的讨论可知，该模型本身就融合了统计和规则知识。特征选取的标准是简单的出现频次，该系统设定阈值为 3。通过 IIS 迭代算法得到特征函数和统计模型之间的联系参数。但是在该系统中用最大熵模型得到的所有概率都是候选的，等到最后一步使用 HMM 中的著名的 Viterbi 算法来确定最佳路径，即最可能的标注类型序列。

值得一提的是，该系统使用了外部系统特征，把外部系统的输出结果作为一个参考特征，试验结果表明该特征大大提高了系统性能，甚至超过了基于 HMM 的系统，从而很好地体现了最大熵模型知识源的任意性。

该系统还针对不同特征的重要性作了比较：词典特征，外部系统特征，参考结论特征都对提高性能有很大帮助。

MENE 系统还可以利用前后缀特征和英语语法信息触发使用最大熵模型，从而进一步提高系统的速度和精确性。但是最大熵模型时间复杂性仍是需要进一步研究的问题。

5 命名实体识别研究的发展趋势

从语言处理的层次来看，现有的命名实体识别研究大都停留在文本表层知识的分析和利用，很少更进一步涉及到句法或语义一级。而句法知识和词汇的语义知识都是自然语言处理比较基本和重要的知识，随着命名实体识别研究的更进一步深入，这些知识必然会得到进一步研究和利用。

从使用方法的发展过程来看，基于规则的方法领域性较强，而且需要语言专家的参与，主观意味比较重，缺乏鲁棒性和可移植性；而统计的方法虽然具有一定的客观性，但是人类语言的使用不是一个单纯的随机过程，严重的数据稀疏和系统处理能力的限制也使得统计模型适用的范围很有限，使用统计的方法搜索空间往往非常大而导致过大的开销，因此混合方法是比较理想的方法，统计模型和规则知识结合使用将具有较好的可训练性和可适应性，而且保持性能所花费的代价要比基于规则的系统低得多，因此更具吸引力。这也是命名实体识别未来的发展趋势。

结束语 本文首先介绍了汉语命名实体识别研究中存在的问题以及命名实体识别系统的评测标准，然后从命名实体识别研究整体发展过程来看，着重分析了两种主要方法：基于规则的方法和基于统计的方法，并对研究中使用到的规则知识进行了概括总结。通过分析两个具体的命名实体识别系统对统计模型和规则知识的结合使用进行了介绍。总之，命名实体的识别是自然语言处理中的重要技术，而语言知识和统计方法的有效结合，是命名实体的识别取得成功的必要条件。

参考文献

1 Sundheim B M. Named entity task definition, version 2.1. In: Proc. of the Sixth Message Understanding Conf. 1995. 319～332

2 Borthwick A. A Maximum Entropy Approach to Named Entity Recognition: [Ph. D]. New York University. Department of Computer Science, Courant Institute 1999

3 Humphreys K, Gaizauskas R, Azzam S, et al. Description of the LaSIE-II system as used for MUC-7. In:Proc. of the 7th Message Understanding Conference (MUC-7), 1998

4 URL http://www.ltg.ed.ac.uk

5 Chen H H, Ding Y W, Tsai S C, et al. Description of the NTU System Used for MET2. In: Proc. of 7th Message Understanding Conference, 1998

6 Black W J, Rinaldi F, Mowatt D. Facile: Description of the NE System Used For MUC-7. In:Proc. of 7th Message Understanding Conf. 1998

7 Fukumoto J, Shimohata M, Masui F, Sasaki M. Oki Electric Industry: Description of the Oki System as Used for MET-2. In: Proc. of 7th Message Understanding Conf. 1998

8 Wu Youzheng, Zhao Jun, Xu Bo. Chinese Named Entity Recognition Combining a Statistical Model with Human Knowledge. The Workshop on Multilingual and Mixed-language Named Entity Recognition: Combining Statistical and Symbolic Models (ACL 2003), Sapporo, Japan, 2003. 65～72

9 Sun Jian, et al. Chinese Named Entity Identification Using Class-based Language Model. In: Proc. of the 19th Intl. Conf. on Computational Linguistics 2002

10 Zhou GuoDong, Su Jian. Named Entity Recognition using an HMM-based Chunk Tagger. In:Proc. of the 40th Annual Meeting of the ACL, Philadelphia, PA 2002. 473～480

11 Ratnaparkhi A. A simple introduction to maximum entropy models for natural language processing: [Technical Report 97-08]. Institute for Research in Cognitive Science, University of Pennsylvania, 1997

12 Sekine S, Grishman R, Shinou H. A decision tree method for finding and classifying names in Japanese texts. In: Proc. of the Sixth Workshop on Very Large Corpora, Montreal, Canada, 1998

13 Brill E. Transform-based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-speech Tagging. Computational Linguistics, 1995. 21(4): 543～565

14 Collins M. Ranking Algorithms for Named-Entity Extraction: Boosting and the Voted Perceptron. In:Proc. of the 40th Annual Meeting of the ACL, Philadelphia, July, 2002. 489～496

15 Jansche M. Named Entity Extraction with Conditional Markov Models and Classifiers. The 6th Conf. on Natural Language Learning, 2002

16 张宏林. Visual C++. 数字图像模式识别技术及工程实践, 58～93

17 李涓子, 黄昌宁. 语言模型中一种改进的最大熵方法及其应用. 软件学报, 1999

18 Mikheev A, Grover C, Moens M. Description of the LTG system used for MUC-7. In:Chinchor[1998]. URL http://www.muc.saic.com/proceedings/muc_7_toc.html

19 Yu et al. Description of the Kent Ridge Digital Labs System Used for MUC-7. In: Proc. of the Seventh Message Understanding Conf. 1998

20 Chua Tat-Seng, et al. Learning Pattern Rules for Chinese Named Entity Extraction. In: Proc. of AAAI'02, 2002

21 Sun Jian, Zhou Ming, Gao Jianfeng. A Class-based Language Model Approach to Chinese Named Entity Identification. Computational Linguistics and Chinese Language Processing, 2003

22 Grishman R. The NYU system for MUC-6 or where's the syntax? In:Proc. of the Sixth Message Understanding Conf. (November 1995), Morgan Kaugmann

23 Lin D. Using collocation statistics in information extraction. In: Proc. of the Seventh Message Understanding Conf. (MUC-7), 1998

24 Krupka G R, Hausman K. IsoQuest: Description of the NetOwl (tm) extractor system as used in MUC-7. In:Proc. of the Seventh Message Understanding Conference (MUC-7), 1998