HMM 与自动规则提取相结合的中文命名实体识别

廖先桃 于海滨 秦兵 刘挺
（哈尔滨工业大学信息检索研究室，哈尔滨，150001）
E-mail: {taozi, yhb, qinb, tliu}@ir.hit.edu.cn

摘要： 本文实现的中文命名实体识别系统采用了隐马尔可夫模型（Hidden Markov Model, HMM）与自动规则提取相结合的方法。整个识别过程可以分为两个步骤，首先使用 HMM 识别，然后再利用自动提取的规则对识别结果进行修正。二者的有机结合使得系统性能有了很大的提高。实验表明，中文命名实体识别总的精确率、召回率和 F 值分别达到了 86.93%，83.69%，85.28%。

关键词: 中文命名实体; 隐马尔可夫模型

引言

命名实体（Named Entity, NE）是指被命名的唯一确定的最小信息单位，包括人名、地名、机构名、专有名词、时间表达式、数字表达式等，是构成句子的重要成分。它的识别是信息抽取（Information Extraction）、自动问答（Question Answering）、机器翻译（Machine Translation）等技术的重要基础。近年来，NE 识别系统的评测受到很多会议的关注，如 MUC-6，MUC-7，Coling2002，Coling2003 等[1]。美国国家标准技术研究所（NIST）组织的自动内容抽取（Automatic Content Extraction，ACE）评测会议[1]也将 NE 识别作为其评测的两大任务之一，可见其重要程度。

NE 的识别方法主要有两种，一种是基于规则的方法，它的识别系统比较简单，但是必须人工制定规则，费时费力，且系统的健壮性和移植性不好。现在的 NE 识别更倾向于基于统计的机器学习方法，它的健壮性和灵活性都比规则的方法好，且代价小，但是需要大规模的语料进行训练。具有代表性的机器学习算法有 HMM[2]，Maximum Entropy[3]，Memory-based learning[4], Support Vector Machine[5]等。

目前，英文 NE 的识别技术已经达到了较高水平。然而，对中文来说，因为一方面它没有空格标志词语边界，另一方面也没有词语的明确定义，所以和英文相比，中文 NE 的识别更加困难。常用手动书写规则并结合词频信息的方法，目前也开始使用机器学习算法，如 HMM[6][7]，class-based language model[7][8], RRM type methods[7][9]等。

本文实现的中文 NE 识别系统采用了统计和规则相结合的方法，首先以词性为观察值建立一个简单的基于 HMM 的 NE 识别系统，然后再利用自动获取的规则对 HMM 的识别结果进行修正。这种方法可以有效的发挥二者的优势，充分利用 NE 的内部结构特征和上下文信息，克服了二者简单结合所带来的识别结果不易融合的缺点，提高了系统对中文 NE 的整体识别效果。

1 系统概述

我们的系统包括两个部分，一是基于 HMM 的 NE 识别部分，二是规则修正部分。系统的结构图如图 1 所示。基于 HMM 的 NE 识别部分的输入是已分词且标注了词性的句子，它的识别结果又作为规则修正部分的输入，整个系统的输出是标注了 NE 标记的句子。下面我们就将对系统做详细介绍。

[1] http://www.itl.nist.gov/iad/894.01/tests/ace/

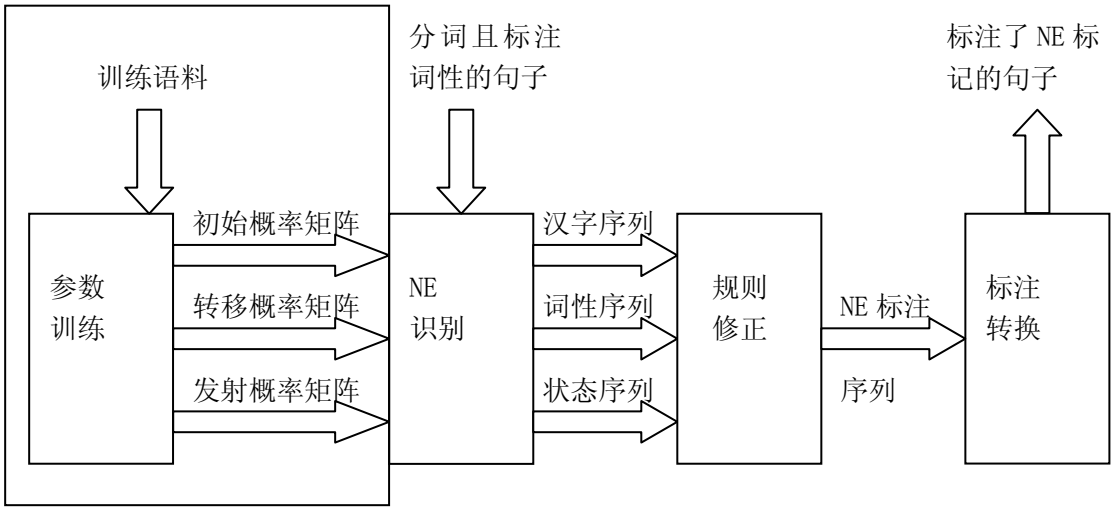

图 1 HMM 与自动规则提取相结合的中文 NE 识别系统

2 基于 HMM 的 NE 识别

2.1 训练

HMM[10]是一种重要的统计自然语言模型，已被广泛应用于语音识别，音字转换，词性标注等领域。模型描述了一个双重的随机过程，其中，状态之间的转换过程是不可观察（隐藏）的,它对应了转移概率矩阵。而可观察的事件的随机过程是隐藏的状态转换过程的随机函数，它对应了发射概率矩阵。HMM 可以表示为一个五元组{S,O,P,A,B}。其中，

S：$\{S_1,\cdots,S_n\}$表示状态的集合；

O：$\{O_1,\cdots,O_m\}$表示观察值的集合；

P = $\{ p_i \}$表示状态的初始概率；

A = $\{ a_{ij} \}$表示从状态 S_i 到状态 S_j 的转移概率矩阵；

B = $\{ b_{jk} \}$表示从状态 S_j 观察到 O_k 的发射概率矩阵。

在我们的基于 HMM 的 NE 识别系统中，主要利用词性作为观察值，共 28 种。系统识别的 NE 有四类，包括人名（Nh）、地名（Ns）、机构名（Ni）和专有名词（Nz）。每一种类别根据它的组成部分在 NE 中出现位置的不同又可以分为 NE 开头(B-NE)、NE 内部(I-NE)、NE 结尾(E-NE)以及独立 NE(S-NE)，再加上不属于任何 NE 类别的“其它”类型（O），共 17 种。这 17 种即为 HMM 中的 17 种状态。

对于某个给定的词性序列$W = w_1, w_2,\cdots, w_m$，NE 识别的目的是要找到一个最优的 NE 标注序列$T = t_1, t_2,\cdots, t_m$使得条件概率$P(T|W)$达到最大。由贝叶斯公式可得：

$$P(T|W) = \frac{P(W,T)}{P(W)} = \frac{P(T)P(W|T)}{P(W)} \tag{1}$$

假设转移概率只与前一个状态有关，输出观察值的概率只与当前状态有关，则有：

$$P(T) = \prod_{i=1}^{m} P(t_i|t_{i-1}) \tag{2}$$

$$P(W|T) = \prod_{i=1}^{m} P(w_i|t_i) \tag{3}$$

其中$P(t_i|t_{i-1})$表示状态t_{i-1}到状态t_i的转移概率。$P(w_i|t_i)$表示在状态t_i出现的条件下观

察到w_i的发射概率。而对于一个给定的词性序列W来说，$P(W)$的值是确定的，可以不考虑。于是系统的最终输出T^*可以表示为：

$$T^* = \arg \max_T (\sum_{i=1}^{m} (\log P(t_i | t_{i-1}) + \log P(w_i | t_i))) \quad (4)$$

2.2 识别

基于 HMM 的 NE 识别过程就是利用训练参数为当前输入的词性序列标注最优的状态序列的过程。即在给定模型$\lambda = \{A, B, P\}$和观察值序列W的条件下找出概率最大的状态序列：

$$Q^* = \arg \max_Q P(Q | W, \lambda) \quad (5)$$

本文采用动态规划的韦特比算法[14]求解最佳的状态序列，其定义为：

$$\delta_s(i) = \max_{t_1, t_2, \cdots, t_{s-1}} P(t_1 t_2 \cdots t_s = i, w_1 w_2 \cdots w_s | \lambda) \quad (6)$$

归纳后有：

$$\delta_{s+1}(j) = [\max_i \delta_s(i) a_{ij}] \cdot b_j(w_{s+1}) \quad (7)$$

也把它称为局部最优函数。

动态规划的思想是把问题分解，先解决最基本的子问题，再逐步外推寻找更大的子问题的最优解，经有限步之后将达到整个问题的最优解。

利用 HMM 进行 NE 识别，系统实现简单，效率也很高。但是由于系统训练时只考虑了词性序列，这就使得训练得到的参数更利于使某个出现的词性标注为它的常见状态。例如，当输入语句为“北京/ns 大学/n”时，因为训练语料中词性/ns（地名）被标为独立地名（S-Ns）的概率更大，所以经 HMM 识别后的结果为“北京/ns/S-Ns 大学/n/0”，这就产生了错误。另外，仅用词性作为观察值也使系统无法利用丰富的词汇信息。比如，“公司”、“研究所”等。同时，系统也不能很好的描述 NE 的内部结构，因此复杂 NE 的召回率很低。

3 规则的自动获取

为了解决上面提到的基于 HMM 的 NE 识别系统遇到的问题，我们提出使用自动提取的规则对 HMM 的识别结果进行修正。这样一来，我们就可以将词汇信息作为特征加入到系统中帮助识别。但规则的运用必须满足前提条件，即修正要以 HMM 的识别结果为基础。

我们提取的规则主要包括 NE 的内部结构特征和外部特征。前者用于复合 NE（主要指由多个词构成的地名、机构名和专有名词）的判断，后者用于独立 NE（主要指人名、独立地名、独立机构名和独立专有名词）的确认。

3.1 NE 内部结构特征的自动提取

从词的角度看，NE 是由一个或多个词组成的，具有一定的结构。然而 HMM 却很难描述，因此我们引入复合 NE 的内部结构特征。它由 NE 的内部词性序列和特征词两部分构成。NE 的内部词性序列是指从语料中自动提取并满足一定阈值的构成复合 NE 的词性串，它体现了 NE 的内部结构。设某个复合 NE 由 n 个词组成，那么它的内部词性序列就是指这 n 个词所对应的词性。特征词是指诸如“公司”、“研究所”等可以明确指示 NE 类别的词语，同样从语料中自动提取。规则的组成形式为：

（1） POS ＋ Pro ＋ Nh ｜ Ns ｜ Ni｜ Nz

（2） 特征词 ＋ Nh ｜ Ns ｜ Ni ｜ Nz

其中 POS 是指复合 NE 的词性串序列。Pro 是该词性串序列出现的概率，由 POS 出现的次数除以词性串序列总数得到。Nh | Ns | Ni | Nz 为 POS 或特征词对应的 NE 类别。

为了满足使用规则的前提条件，我们规定只有 NE 的内部词性序列和特征词对应同一种 NE 类别才表示规则匹配，此时才能修改 HMM 的识别结果。这与传统的规则匹配方法不同。首先，传统的规则方法是独立运行的，它与 HMM 一样直接对原始语料进行处理，只要规则匹配就表示识别成功。但是，这种做法存在一个明显的不足，即当同一个序列经 HMM 和规则处理之后，被识别为不同类别的 NE，此时识别结果无法取舍。而我们的系统是将规则作为 HMM 识别结果的修正部分，这样做一方面保证了 HMM 的高效性，减小了规则的负面影响，另一方面也减少了规则匹配的次数，提高了规则匹配的精确率和系统运行的效率。其次，传统的规则方法常常是由人手动书写规则，需要耗费大量的人力物力，而我们的规则从训练语料中自动提取，减少了人的参与，同时还加入了概率统计信息，更好的解决了规则冲突的问题。例如下面的例子。

例 1：“哈尔滨工业大学”经 HMM 识别后的结果为：

哈尔滨/ns/S-Ns 工业/n/0 大学/n/0

根据词性序列规则 /ns /n /n *0.07691 #Ni 和特征词 大学#Ni，我们可以将结果修正为：哈尔滨/ns/B-Ni 工业/n/I-Ni 大学/n/E-Ni

例 2：“中国康复研究中心”经 HMM 识别后的结果为：

中国/ns/S-Ns 康复/v/0 研究/v/0 中心/n/0

它同时满足规则 /ns /v /v /n *0.000113131 #Ns、中心#Ns 和 /ns /v /v /n *0.00081077 #Ni、中心#Ni，此时根据概率，我们可以将结果修正为：

中国/ns/B-Ni 康复/v/I-Ni 研究/v/I-Ni 中心/n/E-Ni

可见，NE 的内部结构特征对复合 NE 的识别很有效。

3.2 NE 外部特征的自动提取

由于 NE 是处在某个具体的语言环境中的，因此 NE 类别的判定必然受到上下文环境的影响。为此，我们从训练语料中提取出 NE 的上下文信息作为外部特征帮助识别。但是，它的作用又和一般的外部特征不同，它是专用于确认 HMM 系统识别出的独立 NE 的。NE 类别不同，外部特征的定义也有差异。

⑴人名

人名的结构相对简单，因此对人名的识别主要集中在对它的确认上。而人名的前边或后边又常有一些指示词，如“国务卿”、“教授”、“说”等，所以当出现这类表示头衔的词或动词时，就可以确定当前待识别的 NE 是人名。

⑵地名、机构名、专有名词

由于这三类 NE 的结构比较复杂，因此在 HMM 识别之后，我们首先要判断识别出的 NE 是否属于独立 NE，其次再考查它们是否是复合型 NE。

判断独立 NE 时，我们使用这三类 NE 的后缀词。选取依据主要来自 NE 的内部结构特征，即挑选那些出现在词性/ns、/ni 和/nz 之后，且不能和它们组合成为复合 NE 的词性所对应的词语。具体而言：

[1]地名：选取/ns 后边词性为/p，/u，/wp，/b，/v 对应的词；

[2]机构名：选取/ni 后边词性为/p，/wp 对应的词；

[3]专名：选取/nz 后边词性为/p，/wp 对应的词。

判断复合 NE 则利用 NE 的内部结构特征。

3.3 自动提取规则的应用

规则的应用分为两步，一是确认独立 NE，即利用 NE 的外部特征确认 HMM 识别的独立 NE。二是识别复合 NE，即运用 NE 的内部结构特征对 HMM 的识别结果进行修正。最后将识别后的

NE 标记序列输出。

4 实验结果及其分析

我们使用北大富士通语料作为训练集和测试集，分别为 240994 句和 10000 句。使用的词性集是 863 词性标注集，共 28 种。

从语料中自动提取的 NE 的内部结构特征包括：词性串序列 413 条，特征词 2301 个，其中地名特征词 307 个，机构名特征词 1835 个，专有名词特征词 159 个。NE 的外部特征包括：人名前缀词 1565 个，人名后缀词 2909 个，地名后缀词 339 个，机构名后缀词 39 个，专有名词后缀词 40 个。评测指标选用准确率（P）、召回率（R）和 F 值。具体定义如下：

$$P = \frac{\text{系统标注正确的 NE总数}}{\text{系统标出的 NE总数}}$$

$$R = \frac{\text{系统标注正确的 NE总数}}{\text{测试集中出现的 NE总数}}$$

$$F = \frac{2PR}{P + R}$$

为了比较单独使用 HMM 识别 NE 与利用规则修正后的识别效果的差异，我们做了两组对比实验。实验结果见表 1。

表 1 实验对比结果

NE 类型	NE 总数	模型	准确率（P）	召回率（R）	F 值
人名	3,104	HMM	93.30%	93.33%	93.32%
		HMM + rule	93.86%	92.53%	93.19%
地名	3,725	HMM	72.31%	89.72%	80.08%
		HMM + rule	86.69%	85.83%	86.25%
机构名	1,906	HMM	76.47%	25.92%	38.71%
		HMM + rule	77.20%	65.90%	71.10%
专有名词	437	HMM	59.27%	81.24%	68.53%
		HMM + rule	77.14%	80.32%	78.70%
总的结果	9,172	HMM	79.00%	77.28%	78.13%
		HMM + rule	86.93%	83.69%	85.28%

从实验结果来看，经过规则修正，系统识别 NE 的总的 F 值提高了 7.15 个百分点。结构比较复杂的地名、机构名和专有名词的识别率都有显著的提高，而相对简单的人名的识别结果变化不大。这说明规则能很好的体现复杂 NE 的内部结构特征，并充分利用 NE 的外部特征，它克服了单独运用 HMM 不能很好体现复合 NE 结构特征，而单独运用规则时间开销大、易发生冲突的缺点，使 HMM 和规则有机结合，发挥了最大的优势。

5 结论

本文采用 HMM 与自动规则提取相结合的方法，将中文命名实体的识别过程分解为识别和修正两个步骤，有效的利用了 NE 的内部结构特征以及上下文环境帮助识别，实现简单，且极大的改善了系统的性能。同时所用的规则全部从训练语料中自动提取，减少了人工的参与。但是由于系统训练时只考虑了词性，没有充分利用词汇、语义等信息，且规则的移植性不太好，因此势必制约系统效率的进一步提高。下一步我们将对系统存在的问题做改进，加入更多的特征，同时尝试新的方法，力争使系统的性能有更大的提高。

参考文献：

[1] Honglei Guo, Jianmin Jiang, Gang Hu,Tong Zhang.Chinese Named Entity Recognition Based on Multilevel Linguistic Features. IJCNLP-04, Hailan, China, March 22-24,2004. 294~231.
[2] Daniel M.B., Richard L.S., Ralph M.W. An algorithm that learns what's in a name. Machine Learning, 1999,34(1-3):211~231.
[3] Andrew B. A Maximum Entropy Approach to Named Entity Recognition. Ph.D. thesis, New York University, 1999.
[4] Fien D.M., Walter D. Memory-based named entity recognition using unannotated data.In Proceedings of CoNLL-2003, 2003. 208~211.
[5] Hideki I., Hideto K. Efficient support vector classifiers for named entity recognition.In Proceedings of Coling-2002, 2002.
[6] Yu S.H.,Bai S.H.,Wu P. Description of the kent ridge digital labs system used for muc-7. In Proceedings of the Seventh Message Understanding Conference, 1998.
[7] Hongyan Jing,Radu Florian, Xiaoqiang Luo, Tong Zhang, Abraham Ittycheriah. Howtiget-achinesename (entity):Segmentation and combination issues. In EMNLP 2003, 2003.
[8] Jian Sun, Jianfeng Gao, Lei Zhang, Ming Zhou, Changning Huang. Chinese named entity identification using class-based language model. COLING 2002.Taipei, Taiwan, August 24-25, 2002.
[9] Jianmin Jiang, Honglei Guo, Gang Hu, Tong Zhang. Chinese named entity recognition by regularized winnow algorithm. In Proceedings of 20th International Conference on Computer Processing of Oriental Languages, 2003.
[10] 翁富良,王野翊.计算语言学导论.中国社会科学出版社,1998,124~136.

作者简介： 廖先桃（1981-），女，四川什邡人，哈尔滨工业大学计算机系本科四年级学生。

HMM combined with automatic rules-extracting for Chinese Named Entity recognition

Xiantao Liao Haibin Yu Bing Qin Ting Liu

(Information Retrieval Laboratory of Harbin Institute of Technology, Harbin, 150001, China)

E-mail: {taozi, yhb, qinb, tliu}@ir.hit.edu.cn

Abstract: This paper presents a Chinese named entity recognition system that integrates the Hidden Markov Model (HMM) and rules which are automatic extracted from the training corpus. The whole process of recognizing can be divided into two steps. First use the HMM for recognition, and then make use of rules to amend the result of the HMM. System's efficiency has been much improved as a result of the combination of the two models. The precision, recall and F-score of the Chinese named entity recognition are 86.93%, 83.69%, 85.28%, respectively.

Key Words: Chinese named entity; HMM;