DOI:10.11992/tis.201606009
网络出版地址：http://www.cnki.net/kcms/detail/23.1538.TP.20160808.0830.010.html

融合实体特性识别越南语复杂命名实体的混合方法

刘艳超[1]，郭剑毅[1,2]，余正涛[1,2]，周兰江[1,2]，严馨[1,2]，陈秀琴[3]
（1.昆明理工大学 信息工程与自动化学院，云南 昆明 650500；2.昆明理工大学 智能信息处理重点实验室，云南 昆明 650500；3. 昆明理工大学 国际教育学院，云南 昆明 650093）

摘 要：命名实体识别是自然语言处理过程中的基础任务。本文针对越南语的复杂命名实体难识别及 *F* 值不够高的问题，提出了一种结合实体库的越南语命名实体识别混合方法。首先，本文根据越南语的语言和实体特点，选取有效的局部特征和全局特征，应用最大熵模型进行越南语命名实体识别；其次，根据本文制定的命名实体的规则进行越南语命名实体识别；然后，结合两者的识别结果，以规则为主，统计为辅原则；最后经过人工校对，把获取到的正确标记的实体加入到实体库，动态扩增实体库，为规则制定和特征选取提供丰富的语料和依据。实验表明，该方法能够有效地结合规则与统计的方法优点，互相弥补不足，明显提高了识别的正确率、召回率和 *F* 值。

关键词：越南语；实体库构建；实体识别；最大熵；规则；实体特点；全局特征；局部特征
中图分类号：TP391 **文献标志码：**A **文章编号：**1673-4785(2016)04-0503-10

中文引用格式：刘艳超，郭剑毅，余正涛, 等. 融合实体特性识别越南语复杂命名实体的混合方法［J］. 智能系统学报， 2016, 11(4): 503-512.
英文引用格式：LIU Yanchao, GUO Jianyi, YU Zhengtao，et al. A hybrid method to recognize complex vietnamese named entity incorporating entity properties［J］. CAAI Transactions on Intelligent Systems, 2016, 11(4): 503-512.

A hybrid method to recognize vietnamese complex named entity incorporating entity properties

LIU Yanchao[1], GUO Jianyi[1,2], YU Zhengtao[1,2], ZHOU Lanjiang[1,2], YAN Xin[1,2], CHEN Xiuqin[1,2]
(1.School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China;2.Key Laboratory of Pattern recognition and Intelligent computing of Yunnan College, Kunming 650500, China; 3.The School of International Educaton, Kunming University of Science and Technology, Kunming, 650093, China)

Abstract：NER (named entity recognition) is the basic task in natural language processing. Aimed at the problems of low F values and the difficulty with complex Vietnamese named entity recognition, a hybrid method incorporating entity properties is proposed. Firstly, according to the Vietnamese language and entity characteristics, local and global features were selected and a maximum entropy model built to recognize Vietnamese named entities. Secondly, according to the named entity rules obtained, the Vietnamese entity was recognized. Then, combining the recognition results, this paper uses the rule as the main principle and statistics as the supplementary principle. Finally, the obtained correct entity was added to the entity corpus after manual correction, dynamically expanding the entity corpus, which provided a rich corpus and a basis for determining rules and selecting features. Experimental results show that the method can effectively take advantage of rules and statistics, and that recognition accuracy, recall, and *F* are all significantly improved.

Keywords: vietnamese; entity library construction; entity recognition; maximum entropy; rules set; entity characters；lobal features; local features

命名实体识别的任务是识别待处理文本中的人名、地名、机构名、数字、时间、货币和百分号这 7 种命名实体。其中，人名、地名、组织机构名最难识别，同时也是最重要的 3 类实体；虽然数字、时间、货币和百分号这些实体相对简单，但是对上

收稿日期：2016-06-02. 网络出版时间：2016-08-08.
基金项目：国家自然科学基金项目(61262041,61472168,61562052)；云南省自然科学基金重点项目(2013FA030).
通信作者：郭剑毅. E-mail: gjade86@hotmail.com.

层分析都有重要意义。命名实体识别属于自然语言处理的基础研究领域,是组块分析[1]、数据挖掘、信息抽取[2]、信息检索[3]、句法分析[4]、语义分析[5]、自动文摘[6]、问答系统[7]和机器翻译[8]等自然语言处理过程中的重要基础,同时也是重要的预处理过程。

越南语命名实体识别是很困难的一项任务。原因包括：1）实体复杂。越南国家受多文化的影响，在实体命名方面显示出命名实体的多样性和复杂性；越南地名命名广泛，主要分为基本地名和复合地名；越南语实体拼写多样化，比如：东京（Đông Kinh, Tôkiô, Tô-ky-ô, Tô-ki-ô),胡志明(tphcm, hồ chí minh, hochimin.)等；地名中同时含有数字出现，比如第 1 坊 h, tp hcm.（“phường 1”），3 号国道（“quốc lộ số 3”），同时越南语和其他语言一样都存在外来词现象等；2）越南语有其独特的语言特点。越南语是孤立语，没有丰富的形态变化；越南语词是由一个或多个词素构成；越南人名和中国人名类似，唯一不同在于人名存在垫字，例如“Nguyễn Thị Tuyết”阮氏雪，常见的垫字有“文”（Văn）、“妙”（Diệu）、“女”（Nữ）、“玉”（ngọc）、“氏”（Thị）等；越南地名各音节首字母大写；比如：昆明（Côn Minh）、云南（Vân Nam）；非汉越外国地名，首字母大写，音节内部使用“-”连接，比如：Oen-linh-tợn；越南语机构、团体名称一般第一个音节首字母大写（词组除外）等。以上问题给越南语命名实体识别带来极大的困难与挑战。

1　相关研究

对于英语和汉语等语言，命名实体的研究都取得了较好的研究成果。目前命名实体的研究，主要有以下几类方法：1)基于规则的方法, R.Alfred 等[9]根据马来西亚语的语言特点，制定马来西亚语命名实体识别的规则集合；如李楠等[10]根据中文化学领域中实体特点，制定中文化学领域的命名实体识别的规则集合，并引用启发式信息；Elsebai 等[11]根据阿拉伯语命名实体特点，制定阿拉伯语命名实体识别的规则集合,进行识别实体。2)基于统计的方法；S.Zhao 等[12]结合印第安语言特点提出基于隐马尔可夫模型的命名实体识别方法；I.Ahmed 等[13]使用最大熵模型进行命名实体识别,取得了很好的效果；张玥杰等[14]提出以最大熵模型作为框架，结合中文实体特点,融合全局特征和局部特征识别命名实体，取得了很好的效果,正确率达到 87.29%；Y.Benajiba 等[15]结合阿拉伯语语言特点，提出基于支持向量机的命名实体识别方法，F 值达到 82.71%。3)基于混合的方法，潘正高等[16]结合中文命名实体的特点，采用规则与统计相结合的方法进行中文命名实体识别，互相弥补不足，取得了很好的效果； Y.H.Cai 等[17]针对中文组织机构名识别中的标注语料匮乏问题，提出一种基于协同训练机制的机构名识别方法，主要将条件随机场、支持向量机和记忆学习方法组合成一个分类体系，实验表明，混合方法能有效地互相弥补不足；如 S.Biswas 等[18]主要提出一种基于隐马尔可夫模型和最大熵模型的结合，同时根据语言特点，制定规则集识别命名实体，取得了很好的效果；M.A.Meselhi 等[19]提出一种新的混合方法，把规则和统计相结合提高命名实体识别的正确率，实验表明该方法取得的效果要高于单独使用规则或者统计分析的正确率。4)其他方法，尹继豪等[20]针对中文机构名称自动识别提出了简化的一体化 N 最佳层叠模型,该模型实现了汉语切分、词性标注、组块分析和机构名实体识别，同时加入启发信息和机构名称缩写处理,命名实体识别效果显著提高。目前，在越南语实体识别方面有部分研究：V.H.Nguyen 等[21]首先规范越南语的微博内容，然后在支持向量机模型中融入特征进行只针对越南语微博语料进行实体识别，该研究有一定的局限性；R.C.Sam 等[22]为了解决大规模标记训练语料不足的问题，提出半监督学习方法实现越南语文本的命名实体识别，并结合指代词与模糊启发式信息；闫丹辉等[23]结合越南语实体特点，提出了基于规则的越南语的命名实体识别，由于语言的多样性和复杂性，该方法所制定的规则集合难以覆盖完全且工作量很大，难以识别新实体、外来实体和缩写实体等；同时该工作只针对人名、地名、组织机构名进行识别，并没有对数字、百分号、时间和货币做出识别，但是这些实体对于文本分析等应用十分重要；潘清清等[24]采用条件随机场模型对越南语的命名实体识别，该方法的局限性在于：单一的模板识别多种类型实体，所选取的特征只有词、词性以及上下文

信息，没有充分结合越南语的语言和实体特点；所选的窗口大小不能满足复杂实体（如长组织机构名等）的识别需求；对语料的选取和规模都有所要求。另外，上述研究的 F 值不高且未充分利用语言和实体特点、受语料规模和类型限制等，单独使用规则或者统计方法已不能解决上述问题。

因此，本文提出一种融合实体特性识别越南语复杂命名实体的混合方法。其主要思想是：首先用人工标记的方法构建一定规模的实体库，包含常用人名、地名、组织机构名、人名姓氏等；其次根据越南语语言特点和实体库中实体特点，制定出识别越南语命名实体识别的规则集合以及选取越南语命名实体识别所用的局部特征和全局特征，使用最大熵模型统计分析，得到越南语命名实体最大熵模型；然后将测试语料分别使用规则集合和最大熵模型分别进行规则和统计分析实体识别，将得到的实体标记结果进行去重、组合等操作进行综合，得到越南语命名实体识别结果；最后人工校对实体识别结果，将正确识别结果加入到实体库中，方便尽可能地制定更全的规则和抽取更有效的特征。实验表明，该方法能够有效地克服了以上越南语实体识别研究的不足；明显地提高了正确率、召回率和 F 值。因此，该方法是有效可行的。

1　命名实体识别框架

本文提出了一种融合实体特性的越南语复杂命名实体识别的混合方法，该方法能够有效地克服单独使用统计分析或规则集合进行命名实体识别的缺点，并融合越南语语言和实体库中实体特点，原理及流程如图 1 所示。

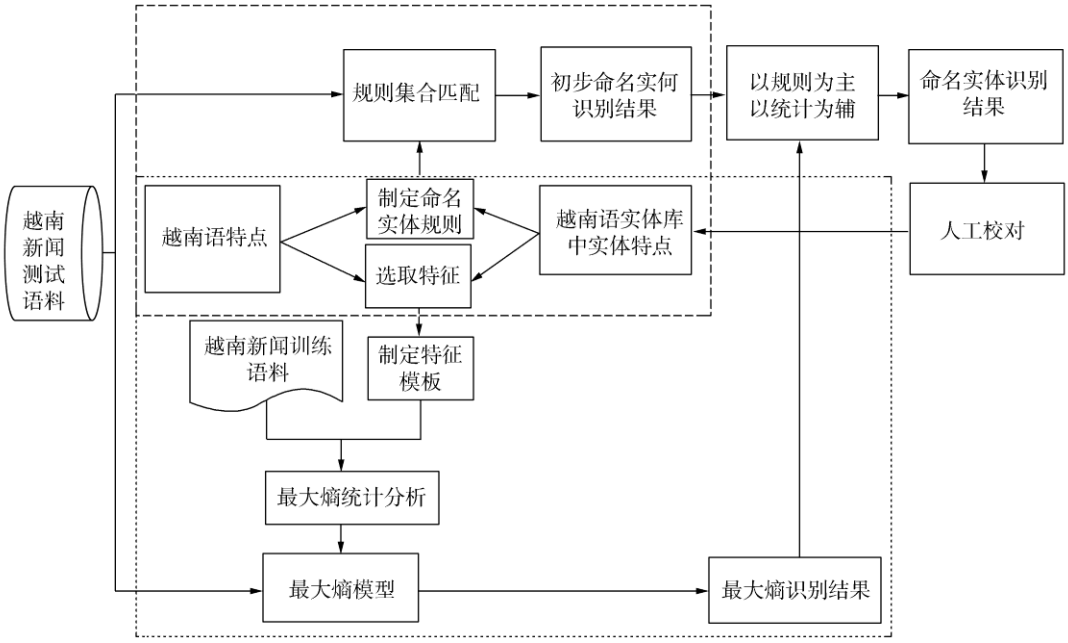

图 1　本文越南语命名实体识别框架

Fig.1　The proposed framework for vietnamese named entity recognition

图 1 中，越南语命名实体库构建主要来源于中越交流圈中新闻、经济、政治等语料进行人工标记以及从维基百科抽取得到，越南语语料来源于微信中越交流圈中收集得到。首先构建越南语命名实体规则集合：根据越南语的语言特点和越南语实体库中实体特点，制定越南语命名实体规则集合；其次构建越南语最大熵模型的命名实体识别模型：根据越南语的语言特点和越南语实体库中实体特点，抽取命名实体识别的特征，构建语料训练格式，使用最大熵统计分析进行建模，得到最大熵模型；然后对越南语实体语料进行测试：将测试语料放入已得到规则集合和最大熵模型进行命名实体识别，分别得到规则识别结果和统计识别结果，将两者得到的结果进行去重等操作，如果两者识别结果不一致，

以规则识别为主；最后将识别得到正确的实体加入到实体库中，方便尽可能地制定更全的规则和抽取更有效的特征。

3 规则集合制定

本文人名、地名、组织机构名所使用的规则以文献[23]制定的规则为基础，已取得很好的效果。另外，其他实体（时间、数字、百分号、货币）相对人名、地名、组织机构名较为简单，识别正确率也高。对于除人名、地名、组织机构名以外的实体本文采用正则表达式和模式匹配进行识别。部分规则表达式如表 1 所示。

表 1 部分规则集合

Table1 Partial rules set

编号	实体规则	例子
1	(0*[1-9]|1[012])/(0*[1-9]|[12][0-9]|3[01])/\d\d	2016/01/01
2	（một|hai|ba|bốn|năm|sáu|bảy|tám|chín|chuc|trâm|nghìn|triệu|tỷ）lần $	Hai lần $
3	^[0-9]+/[0-9]+$	21/01
4	Ngày+Num+tháng+Num+năm+Num	Ngày 01 tháng 1 năm 2015
5	Num + phần trăm	tám mươi phần trăm（80%）
6	(N|n)ăm +Num	năm 2015
7	^[0-9]%$	20%
8	(T|t)háng +Num	tháng 1
9	^([0-9]|[A-Z])+$	10COM
10	(N|n)gày +Num	Ngày 01
11	((\d{4}|\d{2})/((0?([1-9]))|(1[1|2]))/((0?[1-9])|([12]([1-9]))|(3[0|1]))/\1	21/01/2015-22/01/2015
12	Num+giờ+(sáng|trưa|chiều|tối.....)	5 giờ sáng
13	10 giờ 20	10:20
14	(Gấp+Num)|(Num+lần)|(Gấp+Num+lần)	5 lần ; hai lần
15	(T|t)hử+(hai|ba|tư|năm|sáu|bảy...)(C|c)hủ +nhật$	Thử hai
16	^[0-9]+(.[0-9]{n})?$	22.222..
17	một|hai|ba|bốn|năm|sáu|bảy|tám|chín|chuc|trâm|nghìn|triệu|tỷ...	Một tỷ
18	Num+RUR|GBP|CAD|USD|CHF|NZD|JPY|THB|SEK|NOK|DKK|AUD|HKD|SGD|CNY|KRW|MYR|LAK...	2.1 USD
19	Num+giờ+Num+ phút +Num+ giây	3 giờ 4 phút 5 giây
20	Num+ phần Num	1/2
21	(0*[1-9]|1[012])-(0*[1-9]|[12][0-9]|3[01])-\d\d	2016-01-01
22	Num + tỉ|tỷ|triệu|ngàn|nghìn|trăm|chục...	10 tỷ
23	Num+(thước|mét|cây số|cân|tấn...)	1 mét
...		

4 最大熵模型(ME)构建

4.1 最大熵理论

最大熵原理最早由 E.T.Jaynes 于 1957 年提出，1996 年被应用于自然语言处理中。目前，最大熵广泛运用于歧义消解、句法分析、语义分析和上层机器翻译中。

最大熵模型是最大熵分类器的理论基础，该模型基本思想就是为所有已知的因素构建模型，并把未知因素排除在外。它的一个最显著的特点就是，能有效整合多种约束信息，对于越南语命名实体识别具有很好的适用性；同时降低了搜索空间并提高了处理效率。基于最大熵模型的优点，本文采用最大熵模型对越南语命名实体进行建模。

在确定一个词是否为实体过程中，会涉及各种因素，假设 $\boldsymbol{x}$ 就是一个由这些因素构成的向量，变量 y 的值为 1（属于命名实体有效特征）或者 0（不属于命名实体有效特征）。$P(\boldsymbol{x}|y)$是指模型对某个词是否为实体的概率。这个概率可以用上述思想来估计。最大熵模型要求 $P(y|\boldsymbol{x})$在满足一定约束的条件下，必须使得式(1)的熵取得最大值:

$$H(p)=-\sum_{x,y}p(y|x)\log p(y|x)$$

式中的约束条件实际上就是指所有已知的特征:

$$f_i(x,y)=\begin{cases}1 & f(x,y)\ \text{满足一定条件},\\ 0 & \end{cases}$$

$$i=1,2,\cdots,\ n,$$

称 $p^*(y\mid \boldsymbol{x})=\frac{1}{z(\boldsymbol{x})}\exp(\sum_i \lambda_i f_i(\boldsymbol{x},y))$ 为最大熵模型的特征。n 为所有特征的总数。可以看到这些特征描述了向量 $\boldsymbol{x}$ 与变量 y 之间的联系。最终概率输出:

$$p^*(y\mid \boldsymbol{x})=\frac{1}{z(\boldsymbol{x})}\exp(\sum_i \lambda_i f_i(\boldsymbol{x},y))$$

式中：λ_i 是每个向量的权重，且 $z(\boldsymbol{x})$ 表示为

$$z(\boldsymbol{x})=\sum_y \exp(\sum_i \lambda_i f_i(\boldsymbol{x},y))$$

4.2　特征的选取

对于统计模型来说，特征的选取直接决定模型的好坏，对于最大熵模型来说，好处在于选择特征的灵活性，但也要保证选择的特征能反映不同实体类型之间的差异。根据对现有的越南语语言特点和实体库中实体的特点进行分析，本文主要选取局部特征和全局特征作为本文的有效特征。

4.2.1　全局特征

本文所选取的全局特征，针对所有的实体类型进行选取：

1）词上下文信息特征：本文选取词以及上下文信息做为本文的特征，词字符包含丰富形态信息。例如：“河南省”翻译成“tỉnh Hà_Nam”；“阮生雄”翻译为“Nguyễn_Sinh_Hùng”；“1987 年，北京故宫被列入《世界遗产名录》。”翻译为“Năm 1987, Cố_Cung Bắc_Kinh được đưa_vào 《Danh_mục di_sản thế_giới》 .”其中，对于“Bắc_Kinh”（北京）做为当前词，本文选取词的上下文信息为：前一个词是“Cố_Cung”；前第 2 个词是“，”；后一个词是“được”;后第 2 个词是“đưa_vào”作为有效特征。

2）词性上下文信息的特征：本文选取词性以及上下文作为本文的词性特征，词性能够有效地判断词在句子中所起的角色，同时也影响当前词及周围词的大致信息。例如：“chế_biến//N thủy_sản//N xuất_khẩu//N”中，词性顺序为“N N N”构成了一个组织机构名；“Phương_Thị_Thanh//Np”中“Np”表示人名的名词；“1//M”其中词性“M”在识别数字时，起到了很明显的作用；“十亿”翻译为：“một//M tỷ//M”等；在越南语的句子中，句子中的动词、形容词、副词等不可能成为实体的标志，这样可以减小搜索范围，同时也降低了识别错误率，提高处理效率。因此，本文选取词性和词性前后两个词性作为本文的特征。

3）组块上下文信息特征：用组块技术处理命名实体识别技术是可行的[1]，因为名词性组块的定义和命名实体名称结构有很强的相似性，所以只考虑越南语的名词性组块、时间组块、数词组块等来分析越南语的命名实体识别问题是可行的，其他类型组块（形容词组块、副词组块等）不可能成为实体，这样可以减少识别范围和模型搜索范围。本文选取组块以及上下文信息特征，组块标记能够有效地帮助识别实体的边界和类型。首先，“阮芳去学校。”翻译为“Nguyễn_Minh_Phương //B-NP ”Đi //B-VP Đến //B-PP Trường_học //B-NP . //O”,在句子中“Nguyễn_Minh_Phương”是一个名词组块,确定了人名实体边界，同时也确定了名词组块的实体类型;“... Một //B-MP tỷ//I-MP ...”可以确定数字的类型和数字的边界等；组块的标记有利于命名实体边界和类型的识别，同时对组块的长度可以有效地辅助识别实体，组织机构名往往比较长；时间、数字、百分号、人名、地名往往组块长度较短。因此，本文选取当前组块标记、前后两个词的组块标记和组块长度作为本文的有效特征。

4.2.2　局部特征

由于实体类型不一样，所选取的实体特征不一样，本文根据越南语语言特点和实体特点进行选取各种实体类型特征：

1）词素个数信息特征：本文选取词素个数信息作为本文的有效特征，本特征主要针对越南语人名选取的特征,如表 2 所示。

据整理与收集的数据统计分析，越南语的人名主要以 2、3、4 个词素组成。主要受垫字影响，垫字可以省略也可以不省略，比如“Tình”、“Thị”、“Khắc”等。对于其他越南语的构词，主要是 1 个词素和 2 个词素为主，其他词素的个数比例很小，而人名的词素个数主要集中在 3、2、4 为主，因此，越南语词素的个数对于越南语的识别是有效的，本文选取当前词词素个数作为本文的有效特征，其他词素个数不再考虑。

表 2 越南人名词素个数比例

Table2　Morpheme number proportion of vietnamese names

词素个数	频数	比例	举例
1	10	0.000 5	Vượng
2	1 906	0.092 5	Võ Thanh
3	16 811	0.816 0	Vũ Bảo Trân
4	1 561	0.075 7	Hà Thị Vô Danh
5	318	0.015 4	CH TÔN NỮ THANH TRÚC
>=6	6	0.000 3	CÔNG TẰN TÔN NỮ THI CÔ SONG

2）指示词信息特征：本文选取指示词作为本文的有效的特征。指示词能为实体识别提供一定的启发信息，此特征已广泛应用于英文和中文的命名实体当中,指示词往往与实体紧挨。比如指示词“公司”（công ty）、“学校”（trường học）、“大学”（trường đại học）、“先生”（Ông）、“夫人”（bà）、“叔叔”（bác）、省（tỉnh）、县（huyện）、到（đến）、去（đi）、在（tại）等；在识别百分比时可以用“%”作为指示词特征，识别时间时可以用“年”、“月”、“日”等做为指示词。因此，本文选取指示词作为有效特征。

3）首词素是否存在姓氏库信息特征：本文选取首音节是否在姓氏库中存在来判断该越南语词是否是人名实体，越南语人名和中文人名一样，首音节是姓氏，很有可能构成的是人名。本文统计了越南语的人名姓氏库，判断第 1 个词素是否存在于姓氏库中，这样可以减小判断范围，有利于越南语人名的识别。

4）首字母是否为大写信息特征：本文选取越南语词中第 1 个词素的首字母是否大写，在越南语正式的书写中人名和地名的首字母是大写。因此选取首字母是否为大写来区别实体词语非实体词。例如：“北京市”翻译成“Bắc Kinh”；“福建省”翻译成“Phúc Kiến”；人名中“Võ Thanh”，“Hà Thị Vô Danh”等，因此，该特征可以作为本文的特征。

5）其他词素的首字母是否为大写信息特征：本文选取除了首字母以外，其他词素第 1 个字母是否大写，因为对于人名和地名来说，每个词素的首字母都是大写，而对于组织机构名来说并非全部大写。例如：组织机构名“CN Công_ty Du_lịch Hà_Tây tại Hải_Phòng” 中的“tại”词素的首字母为小写；地名中“A Lưới”所有的词素为大写；人名中“Đinh Vũ Nhật Long”所有的词素的首字母均为大写。

6）命名实体字典信息特征：其目的在于有效利用越南语命名实体的相关字典信息，从而弥补训练语料资源受限的不足。其中人名字典分为“越南语姓氏表”、“越南语人名用字表”、“越南语垫字用字表”；地名词典涉及到“常用地名表”和“缩写地名表”；组织机构名字典涉及到“常用机构名表”和“缩写机构名表”；时间字典表涉及到“常用时间表达方式表”。

以上结合越南语的语言和实体特点，选取相应的特征，有效地利用局部特征和全局特征做为本文的特征，作为最大熵模型中必选特征。

5　实体库构建

越南语实体库构建可以有效地分析出越南语实体特点，根据实体特点制定实体识别规则和特征模板。实体库中实体主要来源于新闻、经济、政治等网页识别、中越交流圈平行语料中抽取和维基百科中收集和整理，并经越南语言专家核对得到。实体库实体共收集 139 个常用姓氏；31 251 个常用人名；20 323 个常用地名；6 698 个常用组织机构名；18 个常用货币名称；其中百分比、数

字和时间有固定表达方式，本文采用规则匹配。

6　举例

对于越南语句子“Ngày 24 tháng 11, tại hội nghị, Chủ tịch Tập Cận Bình tuyên bố thực thi toàn diện chiến lược cải cách phát triển quân đội, kiên định bất di bất dịch đi con đường phát triển quân đội đặc sắc Trung Quốc.” (习近平主席 11 月 24 日在会上宣布，全面实施改革强军战略，坚定不移走中国特色强军之路。),其中人名实体“Tập Cận Bình”对人名实体建最大熵模型特征，如表 3 所示。。

表 3 命名实体识别特征选取示例

Table3　Sample of selecting named entity recognition features

特征	特征值	含义
词特征及词上下文信息特征	Tập Cận Bình	当前词
	Chủ tịch	当前词的前第 1 个
	,	当前词的前第 2 个
	tuyên_bố	当前词的后第 1 个
	thực	当前词的后第 2 个
组块特征及上下文信息特征	I-NP	当前组块标记
	B-NP	当前组块标记的前第 1 个
	CH	当前组块标记的前第 2 个
	B-VP	当前组块标记的后第 1 个
	O	当前组块标记的后第 2 个
组块长度	2	当前词所在组块长度
词性特征及词性上下文信息特征	Np	当前词性
	词性 N	当前词性的前第 1 个
	Mark	当前词性的前第 2 个
	V	当前词性的前后 1 个
	ADJ	当前词性的后第 2 个
姓氏特征	1	姓氏是否在姓氏库中
指示词特征	Chủ tịch	指示词字符
词素个数	3	当前词的词素个数
大写特征	1	当前词的首字母大写
	1	其它词素的首字母是否为大写信息特征
是否存在实体字典中	1	判断当前词是否存在实体库字典中

大熵模型的训练文件的格式如图 2 所示。

图 2 的训练文件中列与列之间用制表符“\t”分开。图中第 1 列表示各类实体的标记符号，第 2 列表示当前词的字符，第 3 列表示当前词的词性，第 4~7 列表示当前词的上下文信息特征，第 8~11 列表示当前词的词性的上下文信息特征，第 12~16 列表示当前组块标记以及组块上下文信息，第 17 列表示当前词的词素个数,第 18 列表示姓氏是否在姓氏列表中，第 19 列表示指示词特征，第 20 列表示当前词的首字母是否大写,第 21 列表示除了第 1 个词素之外,其他的词素首字母是否大写，第 22 列表示组块的长度，第 23 列表示该实体是

否包含在实体列表中。

7　实验与分析

7.1　实验数据

实验数据语料来源于中越交流圈中越南新闻、经济、政治等网页，包含大量的命名实体和维基百科抽取得到；通过爬取获得的文本语料，对文本语料进行预处理；经过越南语专家人工标记命名实体语料，形成 140 392 词级规模的命名实体语料。

```
B-TIM   Ngày    N   NULL    NULL    24  tháng   NULL    NULL    M   N   B-TP    NULL    NULL    I-TP    I-TP    1   0   24  1   0   4   1
I-TIM   24  M   Ngày    NULL    tháng   11  N   NULL    N   M   I-TP    B-TP    NULL    I-TP    I-TP    1   0   Ngày    0   0   4   1
I-TIM   tháng   N   24  Ngày    11  ,   M   N   M   Mark    I-TP    I-TP    B-TP    I-TP    CH  1   0   11  0   0   4   1
I-TIM   11  M   tháng   24  ,   tại N   M   Mark    P   I-TP    I-TP    I-TP    CH  O   1   0   tháng   0   0   4   1
O   ,   Mark    11  tháng   tại hội_nghị    M   N   P   N   O   I-TP    I-TP    O   B-NP    1   0   O   0   0   1   0
O   tại P   ,   11  hội_nghị    ,   Mark    M   N   Mark    O   CH  I-TP    B-NP    O   1   0   O   O   0   0   1   0
O   hội_nghị    N   tại ,   ,   Chủ_tịch    P   Mark    Mark    N   B-NP    O   CH  O   B-NP    2   0   O   O   0   0   0   0
O   ,   Mark    hội_nghị    tại Chủ_tịch    Tập_Cận_Bình    N   P   N   Np  O   B-NP    O   B-NP    I-NP    1   0   O   O   0   1   0
O   Chủ_tịch    N   ,   hội_nghị    Tập_Cận_Bình    tuyên_bố    Mark    N   Np  V   B-NP    O   B-NP    I-NP    O   2   0   O   1   0   2
B-PER   Tập_Cận_Bình    Np  Chủ_tịch    ,   tuyên_bố    thực_thi    N   Mark    V   ADJ I-NP    B-NP    O   O   O   3   1   Chủ_tịch    1
O   tuyên_bố    V   Tập_Cận_Bình    Chủ_tịch    thực_thi    toàn_diện   Np  N   ADJ ADJ O   I-NP    B-NP O O    2   0   O   0   0   1   0
O   thực_thi    ADJ tuyên_bố    Tập_Cận_Bình    toàn_diện   chiến_lược  V   Np  ADJ N   O   O   I-NP O B-NP 2   0   O   0   0   1   0
O   toàn_diện   ADJ thực_thi    tuyên_bố    chiến_lược  cải_cách    ADJ V   N   V   O   O   O B-NP  O   2   0   O   0   0   1   0
O   chiến_lược  N   toàn_diện   thực_thi    cải_cách    phát_triển  ADJ ADJ V   ADJ B-NP    O   O   O   O   2   0   O   0   0   1   0
O   cải_cách    V   chiến_lược  toàn_diện   phát_triển  quân_đội    N   ADJ ADJ N   O   B-NP    O   O   B-NP    2   0   O   0   0   1   0
O   phát_triển  ADJ cải_cách    chiến_lược  quân_đội    ,   V   N   N   Mark    O   O   B-NP    B-NP    O   2   0   O   0   0   1   0
O   quân_đội    N   phát_triển  cải_cách    ,   kiên_định  ADJ V   Mark    ADV B-NP    O   O   O   O   2   0   O   0   0   1   0
O   ,   Mark    quân_đội    phát_triển  kiên_định   bất_di_bất_dịch N   ADJ ADV ADJ O   B-NP    O   O   O   1   0   O   0   0   0   0
O   kiên_định   ADV ,   quân_đội    bất_di_bất_dịch di  Mark    N   ADJ V   O   O   B-NP    O   O   2   0   O   0   0   1   0
O   bất_di_bất_dịch ADJ kiên_định   ,   di  con_đường   ADV Mark    V   N   O   O   O   O   B-NP    4   0   O   0   0   1   0
O   di  V   bất_di_bất_dịch kiên_định   con_đường   phát_triển  ADJ ADV N   ADJ O   O   O   B-NP    I-NP    1   0   O   0   0   0   1   0
O   con_đường   N   di  bất_di_bất_dịch phát_triển  quân_đội    V   ADJ ADJ N   B-NP    O   O   I-NP    I-NP    2   0   O   0   0   0   6   0
O   phát_triển  ADJ con_đường   di  quân_đội    đặc_sắc N   V   N   N   I-NP    B-NP    O   I-NP    I-NP    2   0   O   0   0   0   6   0
O   quân_đội N   phát_triển  con_đường   đặc_sắc Trung_Quốc  ADJ N   N   Np  I-NP    I-NP    B-NP    I-NP    I-NP    2   0   O   0   0   0   6
O   đặc_sắc N   quân_đội    phát_triển  Trung_Quốc  .   N   ADJ Np  Mark    I-NP    I-NP    I-NP    I-NP    O   2   0   O   0   0   6   0
B-LOC   Trung_Quốc  Np  đặc_sắc quân_đội    .   NULL    N   N   Mark    NULL    I-NP    I-NP    I-NP    O   O   2   0   đặc_sắc 1   1   6
O   .   Mark    Trung_Quốc  đặc_sắc NULL    NULL    Np  N   NULL    NULL    O   I-NP    I-NP    O   O   1   0   O   0   0   0   0
```

图 2　最大熵模型训练文件

Fig.2　Maximum entropy training file

7.2　实验的评测标准

为了评估本文方法识别命名实体的效果，实验将采用统一的评价标准：正确率、召回率、F 值作为本文评价标准，衡量本文提出的方法的性能。

$$P=\frac{\text{正确识别的实体个数}}{\text{识别出来的实体个数}}$$

$$R=\frac{\text{正确识别的实体个数}}{\text{所有的实体个数}}$$

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

7.3　实验建立

本文为了验证融入实体库中实体特点和越南语言特点的混合方法的性能，主要以下面 3 组实验进行验证本文方法的有效性。

实验 1　为了评估本文方法的性能，我们将 140 392 个词级语料分为 5 份，其中一份做测试语料，另外 4 份作为训练语料，做 5 倍交叉验证实验，然后求平均准确率，作为本文方法的测评结果。实验结果如图 3 所示。

从图 3 中可以看到，Fold5 正确率达到局部最高为 96.14%，为了更准确评估本文方法的可信度和准确性，用平均准确率来评价本文方法，平均准确率为 94.53%。

实验 2　为了验证开放测试和封闭测试对于本文方法的影响，本文在开放和封闭语料上进行测试，实验结果如图 4 所示。

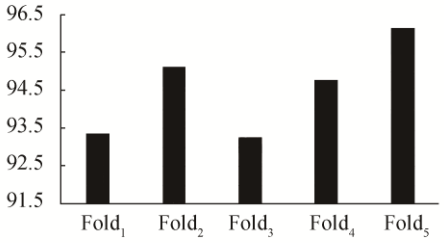

图 3　5 倍交叉验证

Fig.3　5-fold cross-validation

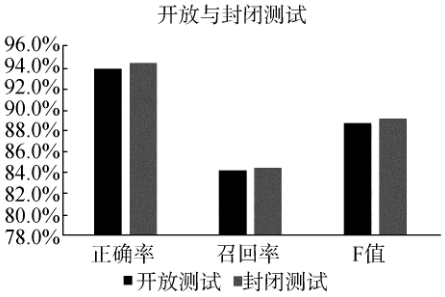

图 4 开放测试与封闭测试

Fig. 4 Open and close testing

从图 4 中可以看到，本文的方法进行了开放测试和封闭测试，实验表明封闭测试正确率比开放测试正确率高 0.66%，封闭测试 F 值比开放测试高

0.49%。因此，本文方法在封闭测试集上效果好于开放测试。

实验 3　为了证明本文方法的效果，本文与文献[23]中规则方法和文献[24]中条件随机场模型进行对比，实验结果如图 5 所示。

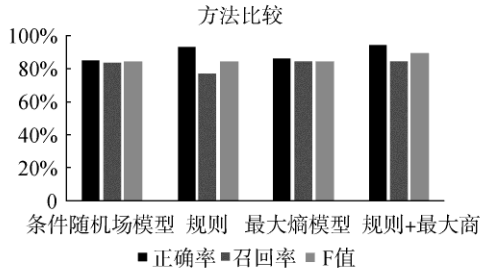

图 5 方法对比

Fig.5　Method comparison

从图 5 可以看到，本文方法（规则和最大熵模型的平均正确率）要比文献[24]中条件随机场模型高 9.69%；比最大熵模型方法高 8.18%;比文献[23]平均正确率高 1.53%；同时 *F* 值也得到提高；其中规则的召回率达到局部最低为 77%，本文召回率高于条件随机场、规则和最大熵模型。因此，本文方法有效可行。

8　结束语

本文根据越南语的语言与实体特点，选取有效的全局特征和局部特征，并借鉴现有的方法，提出融入实体库中实体和越南语语言特点的混合方法进行越南语的命名实体识别。本文在现有的规则基础之上，补充一些重要规则，进行越南语的命名实体识别；同时结合最大熵模型识别实体。由于最大熵能有效地整合多种约束信息和只用考虑特征的选取，因此本文选取最大熵模型进行训练模型，得到最大熵识别识别模型；以规则为主，统计为辅原则，综合实体识别结果；最后，经过人工校对识别结果，将识别正确结果的实体加到实体库中，动态扩展实体库，为规则制定和特征选取提供丰富的语料和依据。通过实验表明本文方法能有效地识别越南语命名实体识别，并且正确率、召回率、*F* 值均有提高，因此，本文方法有效可行。

本文的下一步工作，将解决最大熵模型数据稀疏等问题和结合更多语言特点来识别复杂实体。

参考文献:

[1] ZHOU Guodong, SU Jian. Named entity recognition using an HMM-based chunk tagger[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002: 473-480.

[2] SUNNY T A, SUNDAR G N. An efficient information extraction model for personal named entity[J]. International journal of computer trends and technology, 2013, 4(3): 446-449.

[3] VIRGA P, KHUDANPUR S. Transliteration of proper names in cross-lingual information retrieval[C]//Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition-Volume 15. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003: 57-64.

[4] 尹凌, 姚天昉, 张冬茉, 等. 一种基于混合分析的汉语文本句法语义分析方法[J]. 中文信息学报, 2002, 16(4): 45-51.
YIN Ling, YAO Tianfang, ZHANG Dongmo, et al. A hybrid analysis based Chinese text syntactic and semantic analysis method[J]. Journal of Chinese information processing, 2002, 16(4): 45-51.

[5] 于江德, 樊孝忠, 庞文博. 事件信息抽取中语义角色标注研究[J]. 计算机科学, 2008, 35(3): 155-157.
YU Jiangde, FAN Xiaozhong, PANG Wenbo. Research on semantic role labeling for event information extraction[J]. Computer science, 2008, 35(3): 155-157.

[6] 于海滨, 秦兵, 刘挺, 等. 命名实体识别和指代消解在文摘系统中的应用[J]. 计算机应用研究, 2006, 23(4): 180-182, 195.
YU Haibin, QIN Bing, LIU Ting, et al. Application of named entity and coreference resolution to summarization system[J]. Application research of computers, 2006, 23(4): 180-182, 195.

[7] LU Yonghe, LIANG Minghui. Answer extraction model based on named entity recognition[J]. Applied mechanics & materials, 2014, 571-572: 339-344.

[8] BABYCH B, HARTLEY A. Improving machine translation quality with automatic named entity recognition[C]//Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003: 1-8.

[9] ALFRED R, LEONG L C, ON C K, et al. Malay named entity recognition based on rule-based approach[J]. International journal of machine learning and computing, 2014, 4(3): 300-306.

[10] 李楠, 郑荣廷, 吉久明, 等. 基于启发式规则的中文化学物质命名识别研究[J]. 现代图书情报技术, 2010(5): 13-17.
LI Nan, ZHENG Rongting, JI Jiuming, et al. Research on Chinese chemical name recognition based on heuristic rules[J]. New technology of library and information service, 2010(5): 13-17.

[11] ELSEBAI A. A rules based system for named entity

recognition in modern standard Arabic[D]. Manchester: University of Salford, 2009.

[12] MORWAL S, JAHAN N, CHOPRA D. Named entity recognition using Hidden Markov Model (HMM)[J]. International journal on natural language computing, 2012, 1(4): 15-23.

[13] AHMED I, SATHYARAJ R. Named entity recognition by using maximum entropy[J]. International journal of database theory and application, 2015, 8(2): 43-50.

[14] 张玥杰, 徐智婷, 薛向阳. 融合多特征的最大熵汉语命名实体识别模型[J]. 计算机研究与发展, 2008, 45(6): 1004-1010.
ZHANG Yuejie, XU Zhiting, XUE Xiangyang. Fusion of multiple features for Chinese named entity recognition based on maximum entropy model[J]. Journal of computer research and development, 2008, 45(6): 1004-1010.

[15] BENAJIBA Y, DIAB M, ROSSO P. Arabic named entity recognition: an SVM-based approach[J]. IEEE transactions on audio, speech and language processing. special issue on processing morphologically rich languages, 2009, 15(5): 926-934.

[16] 潘正高. 基于规则和统计相结合的中文命名实体识别研究[J]. 情报科学, 2012, 30(5): 708-712, 786.
PAN Zhenggao. Research on the recognition of Chinese named entity based on rules and statistics[J]. Information science, 2012, 30(5): 708-712, 786.

[17] 蔡月红, 朱倩, 程显毅. 基于 *Tri-training* 半监督学习的中文组织机构名识别[J]. 计算机应用研究, 2010, 27(1): 193-195.
CAI Yuehong, ZHU Qian, CHENG Xianyi. Chinese organization names recognition with Tri-training learning[J]. Application research of computers, 2010, 27(1): 193-195.

[18] BISWAS S, MOHANTY S, MISHRA S P. A Hybrid Oriya named entity recognition system: integrating HMM with MaxEnt[C]//Proceedings of the Second International Conference on Emerging Trends in Engineering & Technology. Nagpur: IEEE, 2009: 639-643.

[19] MESELHI M A, BAKR H M A, ZIEDAN I, et al. A novel hybrid approach to Arabic named entity recognition[M]//SHI Xiaodong, CHEN Yidong. Machine Translation. Communications in Computer and Information Science. Berlin Heidelberg: Springer, 2014, 493(1): 93-103.

[20] 尹继豪, 樊孝忠, 赵攀超, 等. 基于组块分析技术的中文机构名称识别[J]. 哈尔滨工程大学学报, 2006, 27(S1): 466-470.
YIN Jihao, FAN Xiaozhong, ZHAO Panchao, et al. Identification of Chinese organization name based on Chinese chunking[J]. Journal of Harbin engineering university, 2006, 27(S1): 466-470.

[21] NGUYEN V H, NGUYEN H T, SNASEL V. Named entity recognition in Vietnamese tweets[M]//THAI M T, NGUYEN N P, SHEN Huawei. Computational Social Networks. Switzerland: Springer International Publishing, 2015: 205-215.

[22] SAM R C, LE H T, NGUYEN T T, et al. Combining proper name-coreference with conditional random fields for semi-supervised named entity recognition in Vietnamese text[M]//HUANG J Z, CAO Longbing, SRIVASTAVA J. Advances in Knowledge Discovery and Data Mining. Berlin Heidelberg: Springer, 2011: 512-524.

[23] 闫丹辉, 毕玉德. 基于规则的越南语命名实体识别研究[J]. 中文信息学报, 2014, 28(5): 198-205, 214.
YAN Danhui, BI Yude. Rule-based recognition of Vietnamese named entities[J]. Journal of Chinese information processing, 2014, 28(5): 198-205, 214.

[24] 潘清清, 周枫, 余正涛, 等. 基于条件随机场的越南语命名实体识别方法[J]. 山东大学学报: 理学版, 2014(1): 76-79.
PAN Qingqing, ZHOU Feng, YU Zhengtao, et al. Recognition method of Vietnamese named entity based on conditional random fields[J]. Journal of Shandong university: natural science, 2014(1): 76-79.

作者简介：

刘艳超，男，1990 年生，硕士研究生，主要研究方向为自然语言处理与信息抽取。

郭剑毅，女，1964 年生，教授，主要研究方向为自然语言处理、信息抽取、机器学习。主持并参与了多项国家自然科学基金、云南省信息技术重大专项基金、云南省自然科学基金，获得云南省科技进步一等奖 1 项、云南省自然科学二等奖各 1 项。发表学术论文 60 余篇，主编教材 2 部。

余正涛，男，1970 年生，教授，博士生导师，博士，主要研究方向为自然语言处理、信息检索、机器学习。以排名第一获得云南省科技进步一等奖、云南省自然科学二等奖、云南省科技进步三等奖各 1 项。发表学术论文 150 余篇，被 SCI、EI 检索 80 余篇。