基于规则的越南语命名实体识别研究*

闫丹辉[1]，毕玉德[2]

（1.洛阳外国语学院，河南省 洛阳市 471003；2. 洛阳外国语学院，河南省 洛阳市 471003）

摘要：命名实体识别是信息抽取的重要研究内容，主要包括对组织机构名、地名和人名的自动识别。针对英语和汉语的命名实体识别研究开始较早，主要采用基于规则和基于统计的方法进行识别，但目前国内还少有针对越南语命名实体识别的研究。本文分析了越南语命名实体的语言学特点，对其分类并进行了形式化表达，提出了一种基于规则的越南语命名实体识别方法，实验结果显示，该方法能够达到较高的识别准确率。

关键字：命名实体识别；越南语；规则

中图分类号：TP391　　**文献标识码**：A

Rule-based Recognition of Vietnamese Named Entities

Yan Danhui[1], Yude Bi[2]

(1. University of Foreign Languages, Luoyang, Henan 471003, China; 2. University of Foreign Languages, Luoyang, Henan 471003, China)

Abstract: Named Entity Recognition (NER) is an important task for Information Extraction. NER mainly includes the recognition of person names, location names and organization names. Studies on English and Chinese NER began relatively earlier. Under the impetus of some international conferences such as MUC, ACL, SIGHAN, etc., some practical systems have been developed, mainly using rule-based methods or statistical methods. There are fewer studies carried out on Vietnamese NER, and there are even no domestic studies. This paper presents a rule based method to recognize Vietnamese Named Entities.

Keywords: Named Entity Recognition, Vietnamese, rule.

1 引言

命名实体识别是信息抽取的重要内容，同时，其在信息检索、机器翻译和问答系统等自然语言处理领域都有着广泛的应用。命名实体是指文本中的固有名称、缩写及其他唯一标识。主要包括文本中出现的组织机构名、地名、人名、时间表达及数值表达等。

从目前所掌握的资料来看，在现代越南语（以下简称越南语）命名实体识别方面的研究仍相对较少，其具有重大的研究意义及应用价值。

越南学者 Tri Tran Q., Thao Pham T. X., Hung Ngo Q., Dien DINH, Nigel COLLIER（2007）提出了一个基于支持向量机（SVM）的越南语命名实体识别模型[1]。这也是首次将该机器学习方法应用于越南语命名实体识别。实验表明，该模型的识别效果超过了基于条件随机域（CRF）的模型，总体 F 值达到了 87.75。作者分析了越南语的构词特征及词形特征，在此基础上提出了一个基于 SVM 的越南语命名实体识别模型。实验结果表明，该系统对越

*收稿日期：2014-06-15　　定稿日期：2014-07-28

作者简介：闫丹辉（1987—），男，硕士，助教，主要研究方向为自然语言处理；毕玉德（1967—），博士，教授，主要研究方向为自然语言处理。

南语人名、机构名、地名识别的准确率分别达到了 92.91%，85.16%，89.13%，召回率分别达到了 87.09%，77.11%，88.75%。

本文首先从语言学角度分析了每类命名实体的构成规律，并将这些规律在计算机中进行了实现。该方法具有较好的扩展性，并在实验中取得了较好的识别效果。本文语料来源为 2009 年洛阳市社科规划项目重点项目“现代越南语语料库建设”成果，2009A028。语料涵盖人文社科、自然科学及综合类等各方面。

2 越南语组织机构名识别

越南语是一种孤立语，没有词汇的形态变化，声调多，“越语音节界限分明，一个音节就是一个字”，书写时每个音节之间用空格分开，如：

“Đẹp vô cùng Tổ quốc ta ơi!（我们无比美丽的祖国啊！）

这句话有 7 个音节，也就是 7 个字。有三个单音节词 đẹp、ta、ơi，有两个双音节词 vô cùng、Tổ quốc，这两个双音节词分别可拆分为 vô 和 cùng、tổ和 quốc 四个语素。”

越南语采用拉丁字母书写系统，是一种记音文字，因而存在字母的大小写问题。越南官方出台有相关的书写规范，如，规定越南人名、地名在书写时每个音节首字母必须为大写形式，这在越南人名、地名识别中是个非常有用的特征。对于机构名的书写来说，并没有明确的书写规定，但是通过对语料进行分析，我们发现机构名存在着特殊的书写习惯，可用于对越南机构名识别。

我们将越南语文本中可能出现的机构名称分为以下三类，以进一步对越南机构名的自动识别进行探讨。

2.1 越南国家权力、行政及司法类机构

行政机构是最有代表性的社会组织形式，越南行政机构分为四级，如下表所示：

表 1 越南行政机构设置

1	Trung ương	中央
2	Tỉng/Thành phố trực thuộc Trung ương	省、直辖市
3	Huyện/Thành phố thuộc tỉnh/Thị xã/Quận	县、县级市、郡
4	Xã/Phường/Thị trấn	乡、坊、镇

“每一级都设有四大领导班子：Đảng ủy（党委）、Hội đồng nhân dân（人民议会）、Ủy ban nhân dân（人民委员会）、Mặt trận tổ quốc（祖国阵线）。”[2]

经过对语料的分析，可归纳出越南国家行政系统机构类别如下表 3 所示：

表 3 越南国家行政系统机构类别

机构类型	翻译	机构类型	翻译
văn phòng	办公厅	sở	厅
hội đồng	议会	ty	厅
ủy ban	委员会	chi cục	支局
ban	办公室	phòng	局/处/科
bộ	部	toà án	法庭
vụ	司	toà	庭

tổng cục	总局	cục	局

这些类别名称通常与机构名同时出现，可作为用以识别机构名的特征词。如：Văn phòng Tỉnh ủy Yên Bái（安沛省省委办公厅）、Ủy ban Trung ương Mặt trận Tổ quốc Việt Nam（越南祖国阵线中央委员会）等等。

对以上表中列出的越南机构名进行分析可以发现其具有一个显著的特点，即特征词位于机构名短语开头。

通过对语料进行分析，我们发现此类机构名的用词可分为特征词、工作类用词、地名三类。特征词位于机构名短语的开头，如上文列出的 văn phòng（办公厅），hội đồng（议会），ủy ban（委员会）等；工作类用词指某机构所分管的工作，如 Thuế（税务）、Thống kê（统计）、Công thương（工商）等；而地名为机构所在地的地名。

形式化表达式中，双引号中的字（“a”）代表字符本身，尖括号内（< >）包含的为出现 1 次的必选项，方括号（[ ]）内包含的为可重复 1 至有限次的项，大括号（{ }）内包含的为可重复 0 至有限次的项，竖线（ | ）表示在其左右两边任选一项。

从分析结果来看，当划分为较细的类别时可分别对每类进行形式化表达，举例如下：

（1）<特征词><工作用词>

如 Cục Thống kê（统计局），Ban Pháp chế（法制办公室）等等。

（2）<特征词><工作用词>[<”và”>|<”-”>]<工作用词>

当出现多个工作用词时往往使用“và”或“-”相连，如 Sở Giáo dục và Đào tạo（教育与培训厅），Ban Kinh tế - Ngân sách（经济-预算办公室）等等。

（3）<特征词>{<工作用词>}[<”tỉnh”>|<” huyện”>|<”quận”>]{<地名>}

如 Công an tỉnh（省公安），Cục Thuế tỉnh（省税务局）等等。

2.2 越南政党及社会团体类机构

当前，越南是一个一党执政的国家，越南共产党是越南的执政党。越南共产党各级组织机构的设置与我国类似，主要设置有以下几类机构：

表 4 越南共产党组织机构类别

机构类型	译文
Bộ chính trị	政治部
Ban Bí thư	书记处
Văn phòng	办公厅
Ủy ban kiểm tra	纪检委
Đảng ủy	党委
Ban	部
Thành/Tỉnh/Huyện/ Quận/Phường/Xã ủy	市委、省委、县委、 郡党委、坊党委、乡党委

这些机构类别同样可以作为越南共产党机构名识别的特征词。

越南社会同样存在大量的社会团体，如越南祖国阵线、越南劳动者联合会总会、越南农民协会等，其中越南祖国阵线是越南具有协商性质的民族统一战线组织，是越南最重要

的社会团体，其机构设置具有典型的代表性。下面，我们以该团体为例，介绍越南社会团体中的机构设置。

在中央一级，越南祖国阵线主要设置有如下机构：

表 5 越南祖国阵线机构示例

Đại hội đại biểu toàn quốc Mặt trận Tổ quốc Việt Nam	越南祖国阵线全国代表大会
Ủy ban Trung ương Mặt trận Tổ quốc Việt Nam	越南祖国阵线中央委员会
Đoàn Chủ tịch Ủy ban Trung ương Mặt trận Tổ quốc Việt Nam	越南祖国阵线中央委员会主席团
Ban Thường trực Ủy ban Trung ương Mặt trận Tổ quốc Việt Nam	越南祖国阵线常务委员会

在省一级，越南祖国阵线机构设置如下：

表 6 越南祖国阵线机构示例

Đại hội đại biểu Mặt trận Tổ quốc tỉnh	省级祖国阵线代表大会
Ủy ban Mặt trận Tổ quốc Việt Nam tỉnh	省级越南祖国阵线委员会
Ban Thường trực Ủy ban Mặt trận Tổ quốc tỉnh	省级祖国阵线常务委员会
Văn phòng	办公厅

对于越南政党及社会团体类机构，我们采取如下策略进行识别。首先，建立社会团体数据库，收录越南社会各主要社会团体，以静态匹配的方式对越语文本中的社会团体进行识别。其次，对越南政党类机构进行详细分类，并对每类机构进行形式化表达，以形式化表达作为其识别规则。举例如下：

（1）<特征词>{<工作用词>}[<”Trung ương”>|<”Trung ương Đảng”>]

如 Ban Bí thư Trung ương（中央书记处），Ban Bíthư Trung ương Đảng（党中央书记处）等。其中，“工作用词”指该机构所分管的工作，如 Quân sự（军事），Công an（公安），Tổ chức（组织）等。

（2）<特征词><工作用词>{[<”tỉnh uỷ”>|<”huyện uỷ”>|<” Thành ủy”>]}

如 Ban Tuyên huấn（宣训处），Ban Thường vụ Tỉnh ủy（省委常务委员会）等。

（3）<特征词><工作用词>[<”tỉnh uỷ”>|<”huyện uỷ”>|<” Thành ủy”>]<地名>

如 Ban Thường vụ Tỉnh ủy Thái Bình（安沛省常务委员会），Ban Dân vận Tỉnh ủy Yên Bái（安沛省委民运处）等等。

（4）<特征词>{<”tỉnh”>|<”huyện”>|<” Thành phố”>}{<地名>}

如 Đảng bộ tỉnh Cà Mau（金瓯省党委会），Đảng bộ Thành phố（城市党委会）等等。

（5）<特征词><社会团体名>[<” tỉnh”>|<” huyện”>]{<地名>}

如 Ủy ban Mặt trận Tổ quốc Việt Nam tỉnh（省级越南祖国阵线委员会）等。

2.3 文化教育、公司企业类机构

对于此类机构名，我们将主要从越南教育系统内的各学校、越南的公司企业等两方面进行介绍和分析。

我们语料中收集了 820 所越南高校的名称，如下表所示：

表 7 越南高校示例

1	Đại học Đà Nẵng	岘港大学
2	Đại học Quốc gia Hà Nội	河内国家大学
3	Học viện An ninh Nhân dân	人民安全学院
4	Học viện Báo chí và Tuyên truyền	宣传与报纸学院
5	Học viện Công nghệ Bưu chính Viễn thông cơ sở 2	邮政电信工艺学院
6	Trường CĐ Giao thông vận tải II	第二交通运输高等学校
7	Trường CĐ Thực hành FPT	
8	Trường Đại học Chu Văn An	周文安大学

我们发现，总体上越南高校的命名方式并没有统一的规律可循，且在高校名称中使用缩略语是很常见的现象，如将 cao đẳng（高等）缩略为 CĐ，将 trung học（中学）缩略为 TH，将 kinh tế（经济）缩略为 KT 等等，这更增加了对其识别的难度。但对越南高校名称进行详细分类后，我们发现每类高校名的命名方式仍有一定的规律可循，可据此制定相关的识别规则。

在用词上，越南高校名均以 Đại học（大学），Học viện（学院），Trường（学校），Viện（院）开头，这可作为越南高校名识别的特征词。通过对越南 820 所越南高校名称进行分析发现，高校名称用词的最后一个词也有一定的规律可循，可分为地名（表 7 中 1、2）、普通名词（表 7 中 3、4）、人名（表 7 中 8）、数词（表 7 中 5、6）、缩略语（表 7 中 7）等 5 类。我们对 820 所越南高校名称最后一个词进行了统计，如下表所示：

表 8 越南高校名称用词统计

地名	Đà Nẵng/ Hà Nội/ Thanh Hoá	70%
普通名词	Nhân dân/ đô thị/ phát triển	16%
人名	Chu Văn An/ Nguyễn Huệ	3%
数词	1/2/3/ I/ II/ III	6%
缩略语	FPT/ VN/ PTNT/ HP	5%

从上表可以看出，以地名结尾的高校名称最多，占到了约 70%；其次是以普通名词结尾的高校名，占约 16%；以人名、数词和缩略语结尾的高校名所占比例较小，依次为 3%，6%和 5%。

在命名方式上，不同类别的高校名称具有不同的构成特点。从分析结果来看，当划分为较细的分类时可分别将每类高校名进行形式化表达。在此，我们同样借用巴科斯-诺尔范式对高校名的构成进行形式化表达。举例如下：

（1）<特征词>{性质}{<学科>|<行业>}<地名>

如 Đại học Đà Nẵng，Đại học Huế，Trường CĐ Bến Tre，Trường Đại học Đồng Nai 等等。Trường CĐ 与 Trường Đại học 均是学校的类型，在此也将其作为高校名识别的特征词。此处 的“性质”指国立、民立、国家等高校名称用词；“地名”是广义上的地名，包括如 Đông Á（东亚）、Miền Trung（中部地区）等地理名词和方位名词；“学科、行业”指学校涉及的如工业、科技、音乐等学科和行业。又如 Đại học Quốc gia Hà Nội，Trường CĐ Dân lập Công nghệ thông tin Thành phố Hồ Chí Minh 等。

（2）<特征词>[<学科>|<行业>] {地名}[<数量词>|<缩略语>]

“数量词”指“第一、第二、第三”等。如 Trường CĐ Xây dựng số 1， Trường TH Công Nghiệp III 等。“缩略语”指由大写字母、数字及连接符“-”等组成的缩略语。又如 Trường Đại học FPT，Trường TH Đường Sắt HN 等。越语里缩略语的使用情况比较广泛，诸如“学科”、“行业”及“地名”等均可能为缩略语，而且，当名称中出现两个“学科”、“行业”用词时，采用连接符“-”或连词“và”进行连接。

（3）<特征词>[<学科>|<行业>] {普通名词}

高校名称用词中的“普通名词”指除“学科、行业”用词外的普通名词，如 nhân dân（人民）、đô thị（城市）等。如 Học viện An ninh Nhân dân，Học viện Cán bộ quản lý xây dựng và đô thị 等。

（4）<特征词>{<学科>|<行业>}<人名>

此类高校名以越南历史名人的姓名结尾，如 Trường Đại học Chu Văn An，Trường Đại học Phạm Văn Đồng 等。

在公司企业类机构方面，我们对在河内证券交易所、胡志明证券交易所上市的各类公司企业名称进行了整理，收集了其中的 779 家各类公司企业的名称，如下表所示：

表 9 越南公司企业名示例

1	Công ty Cổ phần Alphanam	Alphanam 股份公司
2	Công ty Cổ phần Bao bìDầu thực vật	植物油包装股份公司
3	Công ty Cổ phần Cao su Đà Nẵng	岘港橡胶股份公司
4	Công ty cổ phần Lilama 3	第 3Lilama 股份公司
5	CTCP Bao bìBiên Hòa	边和包装公司
6	CTCP Chứng khoán Hòa Bình	和平证券股份公司
7	CTCP Sông Đà 1	第一沱河股份公司

越语文本中公司企业名最显著的特点是缩略语的使用，“Công ty Cổ phần（股份公司）”通常都会取各音节的首字母而缩写为 CTCP，如表 9 中的 5、6、7；其次是外来语的使用，如 Alphanam，Lilama，Petrolimex 等等。

根据统计结果，我们将越语文本中公司企业名用词分为如下 5 类：第一类是公司企业的特征词，如 Công ty，Công ty Cổ phần，CTCP，Công ty CP，Cty CP 等等；第二类是公司企业生产经营所涉及的产品类名词，如 cao su（橡胶）、cáp（电缆）、gỗ（木材加工）等等；第三类是公司企业生产经营所涉及的行业类名词，如 chế biến（加工）、Vận tải（运输）等等；第四类是地名用词，这些地名一般为公司企业所在地的地名，如 Việt Nam（越南）、Biên Hòa（越南地名，边和）等等；第五类是数词，如 1、2、3 等等；第六类是外来词及缩略语，如 BECAMEX，LIX，CADOVIMEX 等等。

从分析结果来看，当划分为较细的分类时可分别将每类企业名称进行形式化表达。举例如下：

（1）<特征词>[<地名>|<产品>]

如 Công ty Cổ phần cơ điện lạnh（冷机电股份公司），CTCP công nghệ mạng và truyền thông（通讯与网络工艺股份公司）等等。

（2）<特征词><外来词>{<数词>|<地名>}

如 Công ty Cổ phần Alphanam，Công ty Cổ phần Bourbon Tây Ninh（西宁 Bourbon 股份公司）等等。

（3）<特征词>{行业}<产品>[<地名>|<外来词>|<缩略语>|<数词>]

如 Công ty Cổ phần Cao su Đà Nẵng，Công ty Cổ phần Bột giặt LIX（LIX 洗衣粉股份公司），Công ty Cổ phần Thủy sản số 4（第四水产股份公司）等等。

3 越南人名、地名识别

3.1 越南人名识别

一般来讲，越南人名均由 2-4 个越南语音节组成，下面，我们以一条语料为例来进行分析：

“Đoàn kiểm tra của Tổng cục Chính trị do (Trung tướng) **Nguyễn Tuấn Dũng** dẫn đầu đã làm việc tại Cục Cảnh sát Biển.”（由阮俊勇中将率领的政治局检查团到海警局指导工作。）

这条语料中出现了一个越南人名 Nguyễn Tuấn Dũng（阮俊勇），这个人名由三个音节组成，其中第一个音节 Nguyễn（阮）为越南的一个姓氏，第二、三个音节 Tuấn Dũng（俊勇）为此人的名字。

比较明显的特征是，人名中三个音节的首字母 N，T，D 均为大写。上文中括号内的部分 Trung tướng 意为“中将”，是一种称谓用语，表明 Nguyễn Tuấn Dũng 在此的身份是一名中将，Trung tướng 与 Nguyễn Tuấn Dũng 总是同时出现，可作为识别人名 Nguyễn Tuấn Dũng 的特征词。

在越南人名识别方面，特征词通常表现为称谓用语。现代越南社会的称谓系统比较复杂，《越南语人际称谓研究》一书提出，以人际称谓的内涵作为分类标准，可以“将人际称谓分为亲属称谓、社会称谓、姓名称谓和指代称谓 4 部分”。其中，“亲属称谓是指互相有直接和间接血缘、婚姻、法律等关系的亲戚和亲属的名称”。如汉语中：父亲、儿子、丈夫……；越语中的 ông（先生）、bà（夫人、女士）、bác（老伯）……等。“社会称谓指作为社会群体的人在互相交际时根据对方的社会角色所使用的称谓”。如汉语中：部长、局长、处长……；越语中的 chủ nhiệm（主任）、thủ tướng（总理）、bộ trưởng（部长）……等。“姓名称谓是指人类社会中每一个具体成员的正式的代指符号”，即人名，如李白、曹操……；越语中的 Hồ Chí Minh（胡志明）、Phan Bội Châu（潘佩洲）……等。“指代称谓是指人们在交际时对自身、对方和他方所使用的代称” [2]。如汉语中：我、你们……；越语中的 ta（咱们）、tao（我）、mày（你）……等。

经过对语料的分析，我们发现，出现在语料中的越南人名称谓用语主要涉及上述分类中的亲属、社会称谓，极少量涉及其它分类。鉴于此，我们收集整理语料中出现的称谓用词，构建用于越南人名识别的特征词库。如：

表 10 越南语人名特征词示例

类别	称谓	汉语
官衔	Chủ tịch	主席
	Bí thư	书记
	Bộ trưởng	部长
军衔	Đại tá	大校

	Đại úy	大尉
	Thượng sĩ	上士
学衔	Tiến sĩ	博士
	Thạc sĩ	硕士
	Cử nhân	学士

越南人名与中国人名类似，由“姓”+“名”两部分组成。越南人名的姓有单字姓和双字姓之分，名也有单字名和双字名之分。除此之外，越南人名还有一个独特的特点，即大量使用垫名，越南人名的垫名位于姓和名中间，也分为单字垫名和双字垫名。以单字垫名“Thị”（氏）为例，“Thị”是常用于越南女性姓名的垫名，“据越南学者 1992 年统计，女性姓名的垫名用“Thị”的在农民中为 84%，工人中为 62%，知识分子中为 42%”。[2]

《越南语人际称谓研究》一书对越南人名的构成形式进行了分析，提出了 10 种常见构成形式和 4 种罕见构成形式，认为越南人名的常见构成形式如下表所示：

表 11 越南人名常见构成形式[2]

Họ 姓		Tên 名			
Họ đơn 单字姓	Họ kép 双字姓	Tên đệm 垫名		Tên chính 主名	
		Tên đệm đơn 单字垫名	Tên đệm kép 双字垫名	Tên chính đơn 单字主名	Tên chính kép 双字主名

通过对语料进行分析，我们发现，出现在语料中的人名绝大多数为 2-4 个音节，极少量由 4 个以上音节组成。如：

表 12 越南语常见人名示例

越南人名	构成	对应汉语
Hồ Huy	2 音节，姓+主名	胡辉
Đinh Văn Ân	3 音节，姓+单字垫名+主名	丁文恩
Nguyễn Thị Thanh Tâm	4 音节，姓+双字垫名+主名	阮氏青心

姓氏是人名中必不可少的组成部分，由此，我们构建越南常用姓氏库，用于对越南人名的识别。越南常用姓氏库按照《越南语人际称谓研究》一书指出的“以较为保守的态度确定的，应该是目前最无争议的越人的姓”[2]进行构建，列举如下：

表 13 越南常用姓氏库[2]

姓氏							数量
An	Âu	Bạch	Bùi	Ca	Cái	Cam	88
Cao	Cấn	Cù	Chu/Châu	Chử	Diệp	Doãn	
Dư	Dương	Đái	Đàm	Đào	Đặng	Đinh	
Đoàn	Đỗ	Đồng	Giang	Giáp	Hà	Hạ	

Hàn	Hoàng	Hồ	Hồng	Hứa	Kiều	Khiếu	
Khu	Khúc	Khưu	La	Lã/Lữ	Lại	Lâm	
Lê	Lý	Lộ	Lương	Lưu	Ma	Mã	
Mạc	Mai	Mạnh	Ninh	Ngạc	Ngọ	Ngô	
Ngụy	Nguyễn	Nhữ	Ông	Phạm	Phan	Phù	
Phùng	Quách	Quan	Quản	Sử	Tạ	Tăng	
Tiêu	Tô	Tống	Thái	Thành	Thẩm	Thân	
Trang	Trần	Triệu	Trình	Trương	Ưng	Uông	
Văn	Vì	Vũ/Võ	Vương				

另外，需要注意的是，不同题材的文本中出现的人名往往具有不同的特点，如小说中的人名通常以“称谓+名”的方式出现，如：anh Huy（小辉），cô Nga（小娥），em Hoa（小花）等等，在制定规则时需要加以考虑。

根据以上分析，可采取基于规则的方法对人名进行识别，举例如下：

（1）采取静态匹配的方式识别越南社会各领域内重要人物。

（2）特征词+（2 至 4 个）首字母大写音节。

（3）以姓氏库位基础，姓氏+（2 至 3 个）首字母大写音节

需要指出的是，这些规则并不是唯一的，在实际的使用过程中，需要不断地分析实际情况，对规则进行修改，以更好地进行识别。

3.2 越南行政地名识别

行政地名在一个社会所含的地名系统中是最具代表性的，同时，囿于篇幅所限，在此我们将越南地名类命名实体识别的研究范围限于越南行政地名。越南行政地名共有以下 8 个类别，分别为：thành phố（直辖市/省辖市），tỉnh（省），thị xã（县级市），quận（郡），huyện（县），thị trấn（镇），phường（坊），xã（乡）。截止到 2004 年，越南全国共划分出了 5 个直辖市、59 个省。

行政地名是一个较为稳定的地名系统，因而可以构建越南行政地名词库，采取静态匹配的方法对其进行识别。经过对语料的分析，我们发现，越南行政地名中的通名如：tỉnh，thàng phố，quận，xã等，以及 đến（到），tại（在），đi（去）等通常与地名同时出现，可作为地名识别的特征词，收录到特征词库中。

陆利军在《越南行政地名研究》中对越南行政地名的音节结构形式进行了分析，提出越南行政地名用词可分为单音节、双音节、三音节和四音节四类，绝大多数为双音节和三音节。作者将越南全国有代表性的省份所属市县名进行了统计，如下：

表 15 越南部分省所属市县名音节结构表：（%）[3,4]

音节结构 / 行政区划	单音节	双音节	三音节	四音节
河江省		90	10	
义安省	5.26	94.74		
海防省		100		
奠边省		62.5	37.5	
广南省	14.29	85.71		
多乐省	7.14	28.57	50	14.28
嘉莱省		50	50	
平顺省		77.78	22.22	
西宁省		88.89	11.11	
安江省		100		

从上表可以看出，双音节和三音节地名占了绝大多数，其中，海防省和安江省全部为双音节地名，而河江省和义安省超过 90%以上的地名为双音节。根据分析，我们可以采取基于规则的方法对越南行政地名进行识别，举例如下：

（1）采取静态匹配识别行政地名。

（2）特征词+（2 至 4 个）首字母大写音节。

4 实验结果分析

依据以上分析，我们共制定出越南语命名实体识别规则 152 条，构建了越南语命名实体通用词典，收录越南的常见命名实体，包含越南通用人名库共计 20361 条实例、地名库共计 11911 条实例、机构名库共计 3180 条实例。开发了基于规则的越南语命名实体识别系统，采用 Acess 2003 作为数据库。数据库中包含 ORG（机构名表）、LOC（地名表）和 PER（人名表）三个表，每个表中设置 ID（编号）、NAME（名称）、INTRO（简介）三个字段。系统的主要功能为调用规则识别待处理文本中的人名、地名和机构名，同时将识别结果存入数据库中。

系统包括以下五个模块：

（1）文本预处理模块。该模块的功能是将待处理的文本进行格式化处理，删除多余的部分，只保留文本的正文，以更适合其它模块的处理。

（2）机器词典查询模块。该模块利用通用词典查询待处理文本，识别出常见命名实体。

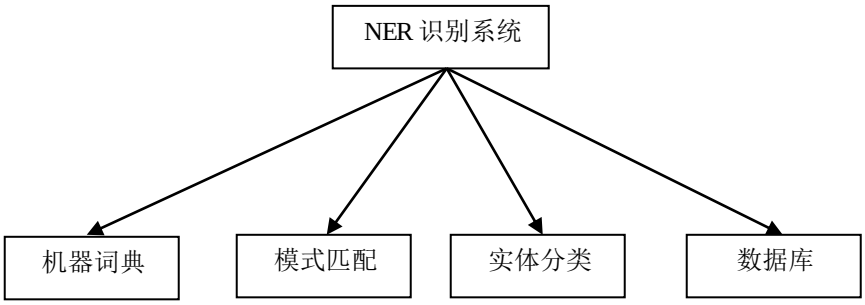
（3）模式匹配模块。该模块集成了用以识别越南语命名实体的各种规则，将实现系统的主要功能，识别待处理文本中的命名实体。

（4）命名实体分类模块。该模块的主要功能为对识别出的越南语命名实体进行自动分类。

（5）命名实体数据库模块。对识别出的命名实体识别进行分类，并存入三个数据库中。

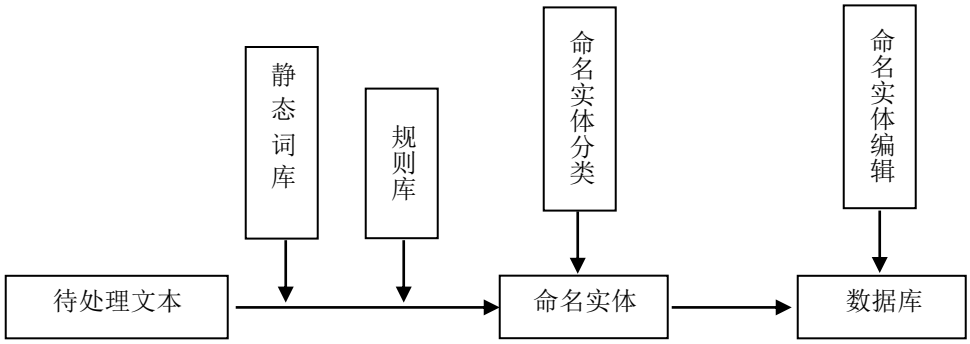
系统结构图如下：

图 1 系统结构图



系统流程图如下：

图 2 NER 系统流程图



我们用 500 篇越南政治、经济类语料，对系统进行了封闭测试，以下为测试结果：

性能指标	机构名	人名	地名
准确率	90.5%	92.7%	95.8%
召回率	74.3%	79.2%	77.5%

该方法的最大优势在于原理简单，容易实现，识别准确率高。在实际应用过程中可以随时添加规则，提高系统召回率。此外，该方法大量避免了识别结果中错误实例。

该方法的缺点难于手工总结出所有可能的规则，制约了系统召回率的提高。同时，要正确识别出某些嵌套结构的、由较多音节组成的组织机构名还比较困难，需要进一步提升规则的覆盖范围并优化算法中规则的执行顺序。

初次测试中系统规则的执行顺序为：人名，地名，组织机构名。这导致了对机构名的识别不完全的现象，如对这条语料：

“…nhà đầu tư tin rằng Cục Dự trữ Liên bang Mỹ sẽ sớm tung ra các gói kích thích mới…”

（投资者认为美国联邦储备局会尽早出台各项新的刺激措施）

系统识别出了”Cục Dự trữ”（储备局）为组织名，”Liên bang Mỹ”（美国联邦）为地名，

而预期的识别结果应为”Cục Dự trữ Liên bang Mỹ”（美国联邦储备局）。鉴于此，我们调整了规则执行顺序，先识别组织名，这次系统虽然能正确识别出预期结果，但是也出现了一些其他问题。

由于未能获取到越南学者开发的越南语 SVM 命名实体识别工具，在此并未进行对比试验。但是从系统的测试结果来看，需要在算法中加入适当的统计因素，这将作为我们的下一步工作进行。未来，我们计划进一步扩大语料分析量，进一步完善规则库并优化规则执行顺序，提高系统召回率。同时考虑结合适当的统计方法[5,6]，深入借鉴中、英文及越南语相关领域的研究成果[7-19]，在保持该方法优势的同时提高召回率。

参考文献

[1] Tri Tran Q., Thao Pham T. X., Hung Ngo Q., Dien DINH, Nigel COLLIER. Named Entity Recognition in Vietnamese documents [J]. Progress in Informatics, No.4, pp.5-13,(2007).

[2] 孙衍峰. 越南语人际称谓研究[M]. 北京:外文出版社，2009.

[3] 陆利军：《越南行政地名研究》,（硕士论文）广西民族大学，2007 年，第 26 页。

[4] 丛国胜. 越南行政地名译名手册[M]. 北京：军事谊文出版社，2004.

[5] 宗成庆. 统计自然语言处理[M]. 北京：清华大学出版社，2008.

[6] Daniel Jurafsky, James H. Martin, 冯志伟 孙乐译. 自然语言处理综论[M]. 北京：电子工业出版社，2005.

[7] 俞鸿魁，张华平，刘群等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报，2006，27（2）.

[8] 张晓艳，王挺，陈火旺. 基于混合统计模型的汉语命名实体识别方法[J]. 中文信息学报，2009，（2）.

[9] Chen, Hsin-His, Yang Changhua & Ying Lin. Learning Formulation and Transformation Rules for Multilingual Named Entities [C] // Proceedings of ACL-2003.

[10] Chieu, Hai Leong & Hwee Tou Ng. Named Entity Recognition with a Maximum Entropy Approach [C] // Proceedings of CoNLL-2003.

[11] Curran, James R. & Stephen Clark. Language Independent NER using a Maximum Entropy Tagger [C] // Proceedings of CoNLL-2003.

[12] Dat Bat Nguyen, Son Huu Hoang, Son Bao Pham & Thai Phuong Nguyen. Named Entity Recognition for Vietnamese [J]. ACIIDS 2010. Part II, LNAI 5991, pp. 205-214.

[13] Klein, Dan, Joseph Smarr, Huy Nguyen & Christopher D. Manning. Named Entity Recognition with Character-Level Models [C] // Proceedings of CoNLL-2003.

[14] Mayfield, James, Paul McNamee & Christine Piatko. Named Entity Recognition using Hundreds of Thousands of Features [C] // Proceedings of CoNLL-2003.

[15] Thao Pham T. X, Tri T. Q., Ai Kawazoe, Dien Dinh & Nigel Collier. Construction of Vietnamese Corpora for Named Entity Recognition [C] // Conference RIAO2007. Pittsburgh PA, U.S.A. May 30-June 1, 2007.

[16] Tri Tran Q., Thao Pham T. X., Hung Ngo Q., Dien DINH, Nigel COLLIER. Named Entity Recognition in Vietnamese documents [J]. Progress in Informatics, No.4, pp.5-13,(2007).

[17] Whitelaw, Casey & Jon Patrick. Named Entity Recognition Using a Character-based Probabilistic Approach [C] // Proceedings of CoNLL-2003.

[18] WU, Youzheng, ZHAO Jun & XU Bo. Chinese Named Entity Recognition Combining a Statistical Model

with Human Knowledge [C] // Proceedings of ACL-2003.

[19] Zhou, Junsheng & He Liang. Chinese Named Entity Recognition with a Multi-Phase Model [C] // Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006: 213-216.

联系方式：闫丹辉，河南省洛阳市 036 信箱 150 号，邮编：471003；电话：13523619594；电子邮箱：diem1987@163.com