doi: 10.3969/j.issn.1006-1576.2011.10.028

基于多模型结合的军事命名实体识别

姜文志[1]，顾佼佼[1]，胡文萱[2]，王彦[1]
（1. 海军航空工程学院兵器科学与技术系，山东 烟台 264001；2. 海军航空工程学院外训系，山东 烟台 264001）

摘要：针对军用命名实体的识别难点，采用多模型结合的方法进行军事命名实体识别。在对军用文本信息进行深入的研究和分析的基础上，根据其特点设计了一个基于双层模型的军事命名实体识别系统，并将军事命名实体分为简单实体名和复杂实体名 2 类：在底层采用 CRFs 模型进行简单实体名识别，再采用 SVM 与 CRFs 结合判断命名实体的左右边界，进行复杂命名实体的辅助识别，最后进行两层识别结果的合并。实验结果表明：该识别模型是一种比较理想的方法。

关键词：军事命名实体；命名实体识别；条件随机场；支持向量机
中图分类号：TJ03　　**文献标志码**：A

Military Named Entity Recognition Based on Multi-Models

Jiang Wenzhi[1], Gu Jiaojiao[1], Hu Wenxuan[2], Wang Yan[1]
(1. Dept. of Ordnance Science & Technology, Naval Aeronautical & Astronautical University, Yantai 264001, China;
2. Dept. of foreign training, Naval Aeronautical & Astronautical University, Yantai 264001, China)

Abstract: In order to identify military named entities a multi-models method was proposed. After detailed research and analysis on military text efficient military named entity recognition system, which is based on a dual-layered model, is presented. The military named entity group is divided into two categories: simple name and complicated ones. First the simple named entity is recognized in the first layer with CRFs, the results are transferred to the second layer to support the recognition of complicated named entity. In the second layer, SVM was used to decide the right boundary of a complicated named entity and then CRFs model was used to tag the former boundary. Last the simple and complicated named entity recognition results are combined together. Experimental results show that the model is effective.

Keywords: military named entity; named entity recognition; conditional random fields; support vector machine

0　引言

计算机以及相关信息处理技术的日渐成熟和发展，为进一步提高军事指挥效能提供了有效手段[1]。现代战争中，信息化[2]体现得越来越明显。军用信息处理中，进行军用命名实体的识别是一项非常重要的基础性工作，它是分析、理解和处理军用信息的前提，能否正确识别出军事命名实体是军事领域信息检索和信息抽取的关键。

目前命名实体识别研究的方法[3]主要有 3 种：基于规则的方法、基于统计的方法和基于统计与规则相结合的方法。特定领域机构名识别的研究不多，但意义重大。早期的研究中，主要是在分析词法和语义结构的基础上依赖人工制定规则，但其识别效果不高，而且制定规则库需要大量的时间和精力，移植能力低，针对大批量真实语料中的军事命名实体的识别效果并不理想。基于统计的识别模型相对基于规则的方式有较好的表现，但每种模型都有自己的优点和缺点。有的模型在处理信息方面能力不足但效率较高，而有些模型则在可有效的利用上下文信息的同时计算效率不高，存在优势互补。

因此，笔者在前人工作和实验的基础上，采用支持向量机[4]（support vector machine，SVM）模型和条件随机场[5]（conditional random fields，CRFs）模型相结合的方法进行军事命名实体识别，在计算效率和识别精度之间寻求最佳状态。

1　概述

1.1　命名实体识别

命名实体[6]（named entity, NE）是中文文本的基本信息单位，是信息的主要载体，正确理解中文文本的基础。从狭义上讲命名实体可以有人名、地名和机构名 3 种最主要的命名实体；广义的讲可以分为好多特定领域特定类型的命名实体。因此命名实体识别是语言信息处理技术中的关键技术，是理解和处理文本信息的前提和基础。

1.1.1　命名实体识别难点

分析和总结了大量的中文命名实体结构特点后

收稿日期：2011-06-10；修回日期：2011-07-28
作者简介：姜文志（1964—），男，山东人，博士，从事计算机应用、武器装备与作战指挥一体化研究。

发现，中文命名实体的构词结构不受限制而且数量庞大，将之悉数列举或全部收录都是不可能的。中文文本中词语边界不是用空格来标识的，也不区分大小写，这增加了识别的难度，而且复杂命名实体经常是嵌套的。即使同一命名实体在同一文本中其上下文的表达形式也有可能是不同的，例如机构名的简写形式。

1.1.2　军事命名实体特点

经过对大量的中文军事命名实体结构特点的分析研究和总结经验，将中文军事命名实体分为简单实体和复杂实体 2 类。简单实体如：“解放军”、“军委”等，即由一个词组成；复杂实体如：“中国人民解放军”等，即由 2 个或 2 个以上词语组成。一个有规律的复杂军事命名实体的组成形式是：一个或多个前部词与一个实体特征词组合而成；无规律的复杂军事命名实体由于存在省略或简写等多种情况，构成了结构更加复杂的机构名。

1.2　支持向量机模型

SVM 是一种基于统计学习理论的模式识别方法，其在解决小样本、非线性及高维模式识别问题中具有非常大的优势，目前已在许多领域成功应用。其主要思想可概括为 2 点：一是针对线性可分情况进行分析，对于线性不可分情况通过转化为高维特征空间使其线性可分；再就是 SVM 基于结构风险最小化理论之上在特征空间中建构最优分割超平面，算法其实就是一个二次寻优过程，可使得学习器得到全局最优化。

1.3　条件随机场模型

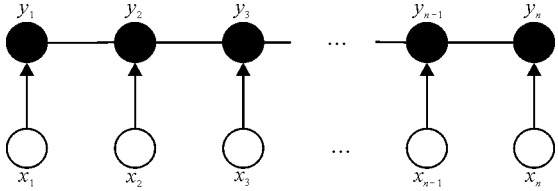
CRFs 是一个基于最大熵模型发展而来的无向图模型，用来在给定输入节点条件下计算输出节点的条件概率。CRFs 不仅可抽取丰富的上下文信息而且还克服了标记偏置问题，在各领域广泛应用。

CRFs 的结构可以是任意的图结构，当图形中各输出节点被连接成一条线性链的情况下，CRFs 假设在各输出节点之间存在一阶马尔科夫独立性。链式 CRFs 是现今最常用的结构之一。在给定观察序列 x，线性链的 CRFs 定义状态序列 y 的条件概率为：

$$p_\theta(y|x) \propto \exp\left(\sum_{e\in E,k}\lambda_k f_k(e,y|_e,x)+\sum_{v\in V,k}\mu_k g_k(v,y|_v,x)\right) \tag{1}$$

其中：$f_k(e,y|_e,x)$ 为状态转移特征函数；$g_k(v,y|_v,x)$ 为状态特征函数；λ_k、μ_k 分别是 $f_k(e,y|_e,x)$ 和 $g_k(v,y|_v,x)$ 的特征权重。

条件随机场模型训练主要有 2 个基本任务：特征选择和参数评估。特征选择就是选一个能表达这个随机过程的统计特征的集合，参数评估是指为入选的每个特征估计权重。一般采用极大似然估计方法来计算特征权重，从训练数据 $D=(x^i,y^i)_{i=1}^N$ 中计算出 $\theta=(\lambda_1,\lambda_2,\cdots;\mu_1,\mu_2,\cdots)$。CRFs 指数模型为凸函数，可以采用迭代方法找到全局最优解。目前常用的是 L-BFGS 迭代方法。一阶链式 CRFs 图形结构如图 1。



注：黑色代表隐藏节点，白色代表观察节点

图 1　一阶链式 CRFs 图形结构

2　基于 SVM 和 CRFs 结合的复杂军事命名实体识别

军事命名实体在识别的过程中可以分为简单军事命名实体和复杂命名实体，简单命名实体结构简单，复杂命名实体结构层次比较复杂，经常嵌套简单命名实体，常由 2 个或 2 个以上的词语组成，完整的构造形式可用“前部词+特征词”表示，经分析发现，完整的命名实体构成还是有一定规律的。笔者构建了一个双层模型，首先在低层采用 CRFs 模型识别简单命名实体，然后将识别结果传至高层，为复杂命名实体的识别提供更多可利用的信息并辅助进行复杂命名实体的识别。最后结合两层的识别结果，识别流程如图 2。

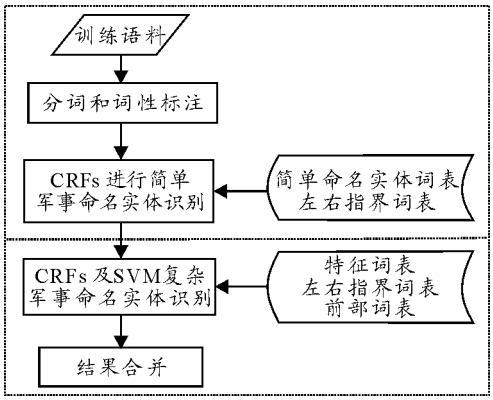


图 2　命名实体识别流程图

具体步骤如下：

1) 对语料进行分词和词性标注，构建所需词表；

2) 选取合适的特征训练 CRFs 模型，进行简单军事命名实体的识别；

3) 将 SVM 和 CRF 相结合，进行复杂军事命名实体识别：根据选定的特征构建向量，SVM 训练得到复杂军事命名实体的右边界识别模型，用该模型识别右边界词；对被确定为右边界的词向前利用 CRFs 进行前部标注。

简单军事命名实体的结构简单，一般只由一个词组成。一个词的上下文是判断该词是否为命名实体的主要条件之一，CRFs 可以做到识别过程中尽可能充分地利用上下文信息。因此，可利用 CRFs 这一优点，在低层基于 CRFs 模型对简单军事命名实体进行识别，在多次试验后，选出 CRFs 的几个综合效能较好的特征。理论上讲，特征多则可利用的上下文信息也多，识别效果会更好，但特征多了会产生冗余信息，影响效率。经过分析及大量实验，笔者人工选择并实验调整选取出特征模板，原子模板如表 1，组合模板如表 2。

表 1 原子特征

原子特征	释义
Word(n)	当前词的词形
Pos(n)	当前词的词性
L_spe(n)	若当前词前面第一个词在左指界词表中标记为 Y，否则标记为 N
R_spe(n)	若当前词前面第一个词在右指界词表中标记为 Y，否则标记为 N
Smp_org(n)	若当前词存在于简单机构名表中则标为 Y，否则标为 N

表 2 组合特征

组合特征	释义
word(n−1)word(n)	[n=−1,0,1]
Pos(n−1)pos(n)	[n=−1,0,1]
L_spe(n−1)Smp_org(n)	[n=−1,0,1]
Smp_org(n−1)R_spe(n)	[n=−1,0,1]

在特征模板选取之后，利用 CRFs 进行训练，获得特征函数和相应的权重。特征函数的权重表示该特征函数对标注结果影响力的大小。

复杂军事命名实体的识别分为 2 步：右边界识别和前部标注。首先寻找实体的右边界，以特征词为牵引，对出现在特征词表中的词使用 SVM 进行右边界判断，SVM 在处理这种二值分类问题上的识别效果和速度有明显优势。若被确定为实体右边界词，接下来以右边界词为起点向前标注，寻找实体的左边界。若其左右边界均被确定，则完成识别。右边界识别流程图如图 3。

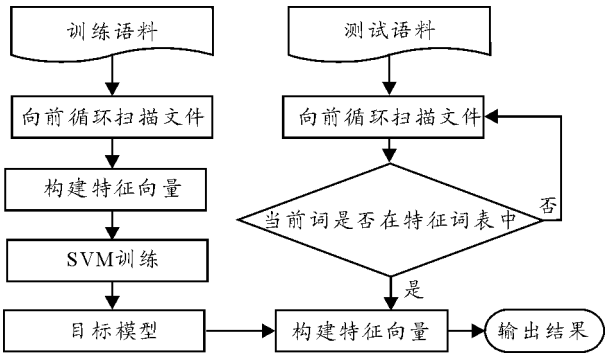


图 3 复杂实体识别中右边界识别流程

确定了右边界后，左边界的确定由此依次向前寻找，直至找到为止。整个过程就是前部标注的过程。此处 CRFs 模型尤其着重上下文信息的提取。经过大量的实验，选定前部标注的原子特征模板及组合模板如表 3、表 4。

表 3 原子特征

原子特征	释义
Word(n)	当前词的词形
Pos(n)	当前词的词性
Former_word(n)	若当前词在前部词表中则标为 Y，否则标为 N
L_spe(n)	若当前词前面第一个词在左指界词表中则标为 Y，否则标为 N
Is_smp(n)	若当前词识别为简单机构名则标为 Y，否则标为 N
Is_candidate(n)	若当前词确定为右边界则标为 L；若为候选词则标为 U，否则标为 0

表 4 组合特征

组合特征	释义
word(n−1)word(n)	[n=−1,0,1]
Pos(n−1)pos(n)	[n=−1,0,1]
L_spe(n−1)Former_word(n)	[n=−1,0,1]
Is_candidate(n−1)R_spe(n)	[n=−1,0,1]

3 实验

笔者使用的语料是部分新华军事网上近期军事类新闻和部分军用文书文集合并构成，大小为 2.2 MB。对其进行分词、标注形成标准语料。在进行命名实体识别时需要很多信息，如词表、特征模板等资源，命名实体识别时用到的资源有特征词表、常用前部词表、左右指界词表、常用简单命名实体名表。笔者用基于词的粒度来进行简单军事命名实体的识别。实验采用 CRF++[7]作为 CRF 训练工具。

为减少实验规模，只针对军事机构名的识别进行实验验证。简单机构名识别效果如表 5。

表 5 简单机构名识别 %

准确率	召回率	F−值
97.77	95.89	96.82

因为简单军事机构名结构简单，而 CRFs 可充

分利用上下文信息，从表 5 可见，识别效果是理想的。若用单层 CRF 来同时识别简单命名实体和复杂命名实体，其效率就有所降低，结果如表 6。

表 6　CRFs 识别机构名效果　　%

准确率	召回率	F-值
91.98	88.29	90.10

效率降低的原因一方面是因为模型以词为识别粒度，因此识别效果受到前期分词精度的影响。另外不规则或不完整的复杂机构名本身就有较大难度，再加上不完备的词表资源，都会影响识别精度。

然后进行 SVM 和 CRFs 相结合的实体识别实验，效果如表 7。

表 7　SVM 及 CRFs 结合识别军事机构名　%

准确率	召回率	F-值
94.18	93.29	92.90

从表 7 可见，识别结果明显优于单纯使用 CRFs 的方法。

4　结束语

实验结果证明：SVM 具有较好的推广和高维处理能力，识别实体右边界时效果较好；就确定为右边界的词向前采用 CRFs 进行前部标注，提高了识别精度和效率。采用 SVM 和 CRFs 相结合的方法进行军事命名实体识别的总体效果是比较理想的。

参考文献：

[1] 刘博. 基于作战文书的标图系统设计与实现[D]. 郑州: 解放军信息工程大学, 2009.

[2] 姜文志, 蒋伟俊, 张金乙, 等. 军用词典库的设计[J]. 兵工自动化, 2007, 26(8): 50-52.

[3] 曾冠明. 基于条件随机场的中文命名实体识别研究[D]. 北京: 北京邮电大学, 2009.

[4] 王国胜. 支持向量机的理论与算法研究[D]. 北京: 北京邮电大学, 2008.

[5] Charles Sutton, Andrew McCallum. An Introduction to Conditional Random Fields[M]. Foundations and Trends in Machine Learning, 2010.

[6] 余军, 陈晓鸥. 命名实体识别: One-at-a-time or All-at-once? Word-based or Character-based?[C]//萧国政, 何炎祥, 孙茂松. 中国计算技术与语言问题研究: 第七届中文信息处理国际会议论文集. 北京: 电子工业出版社, 2007: 81-89.

[7] CRF++: Yet another crf toolkit, http://crfpp.sourceforge.net/.

****

（上接第 89 页）

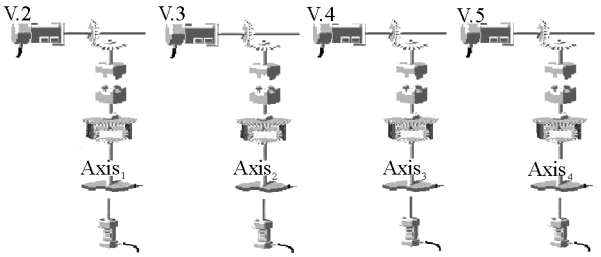


图 6　整定单轴机械程序配置图

其中 $\Delta X_{\min}$ 为主动轴和从动轴位置差的极限值。$\Delta X_{\min}$ 为两轴误差的允许值，在此值范围内不需要补偿，$\Delta X_{\max}$ 为人为设定的阈值，用来区分采用 PID 控制还是 PD 控制[4]。根据减速比、最高使用速度和同步精度要求等确定采样周期 T 为，公式如下：

$$T = \frac{s_j \rho}{V S_L} \tag{2}$$

其中 S_j 为同步精度，ρ 为减速比，V 为最高使用转速，S_L 为螺距。在采样周期内，各轴位置偏差超过安全值时进行整定，位置偏差超过极限值时紧急停止。把采样周期内通过 PID 计算得到的 $u(i)$ 值输出给超差轴。整定时单轴的机械系统程序配置如图 6。

2.3　其他功能应用

通过运动控制器还实现了原点回归、单轴升降调试、虚拟位置设定、伺服故障报警及保护等功能。

3　结束语

该系统已经在某风洞中进行了应用。结果证明：该控制系统能满足风洞地面效应试验装置同步和定位精度要的要求，并提升了试验效率，可推广应用于其他类似的多轴同步系统中。

参考文献：

[1] 王勋年. 低速风洞试验[M]. 北京: 国防工业出版社, 2002.

[2] 李耿轶, 王宇融. 数控机床多轴同步控制方法[J]. 制造技术与机床, 2000(5): 23-25.

[3] 吴其华, 徐邦荃. 多电机同步传动控制系统分析[J]. 自动控制技术, 2003, 22(1): 20-24.

[4] 戴永红. 以 FANUC-15i 为例浅析数控系统同步控制在双驱中的应用[J]. 机械技术与机床, 2005(5): 121-124.