中 文 信 息 学 报
第 20 卷 第 1 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 20 No. 1

文章编号:1003 - 0077(2006)01 - 0007 - 07

面向商务信息抽取的产品命名实体识别研究*

刘非凡[1],赵 军[1],吕碧波[1],徐 波[1],于 浩[2],夏迎炬[2]

(1. 中国科学院自动化研究所 模式识别国家重点实验室 北京 100080; 2. 富士通研究开发中心有限公司 北京 100016)

摘要:市场信息化使得商务信息抽取、市场内容管理日益成为信息科学领域的一个研究热点。产品命名实体识别作为其中非常重要的关键技术之一也逐渐受到人们的关注。本文面向商务信息抽取对产品命名实体进行了定义并系统分析了其识别任务的特点和难点,提出了一种基于层级隐马尔可夫模型(hierarchical hidden Markov model)的产品命名实体识别方法,实现了汉语自由文本中产品命名实体识别和标注的原型系统。实验表明,该系统在电子数码和手机领域均取得了令人满意的实验结果,对产品名实体、产品型号实体、产品品牌实体整体识别性能的 F 值分别为 79.7 %,86.9 %,75.8 %。通过和最大熵模型相比较,验证了 HHMM 对于处理多尺度嵌套序列有更强的表征能力。

关键词:计算机应用;中文信息处理;产品命名实体识别;商务信息抽取;层级隐马尔可夫模型

中图分类号:TP391 **文献标识码**:A

Study on Product Named Entity Recognition for Business Information Extraction

LIU Fei-fan[1], ZHAO Jun[1], LV Bi-bo[1], XU Bo[1], YU Hao[2], XIA Ying-ju[2]

(1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China;
2. FUJITSU R&D, Beijing 100016, China)

Abstract: Electronic business has fueled increasing research interest recently in business information extraction and market intelligence management. As one of the key techniques, product named entity recognition (product NER) has also begun to draw more attention in the field of natural language processing. In the paper, characteristics and challenges in product NER are explored and analyzed deliberately, and a hierarchical hidden Markov model (HHMM) based approach to product NER from Chinese free text is presented. Experimental results in both digital and mobile phone domains show that our approach performs quite well in these two different domains and achieves F-measures of 79.7 %, 86.9 %, 75.8 % on the whole for three types of product named entities respectively. In comparison with maximum entropy model, HHMM is experimentally proved to be more powerful for dealing with multi-scale embedded sequence problem.

Key words: computer application; Chinese information processing; product named entity recognition; business information extraction; hierarchical hidden Markov model (HHMM)

1 引言

市场信息化、商务电子化的飞速发展给商务信息智能处理提供了非常广阔的应用空间,其中商务领域命名实体识别(Named Entity Recognition)是最为基础的关键技术之一。对于商务领域信息抽取任务来说,仅仅靠常规命名实体识别不能有效地抽取文本中的关键信息,尤其是产

* **收稿日期**:2005 - 05 - 03 **定稿日期**:2005 - 11 - 03
基金项目:国家自然科学基金资助项目(60372016);北京市自然科学基金资助项目(4052027)
作者简介:刘非凡(1979 —),男,博士生,主要研究方向是信息抽取、文本挖掘.

品相关信息,产品命名实体识别研究势在必行。

本文对产品命名实体进行了定义,对产品命名实体识别的特点和难点进行了细致的分析,提出了基于层级隐马尔可夫模型(HHMM)[8]的产品命名实体识别方法。该方法通过统计模型的融合,可以综合利用不同层次的语言学特征,并将其与知识库、启发式规则有机地融合到一个统一框架中,在电子数码和手机领域均取得了令人满意的效果。

2 相关工作

目前国内外涉及产品命名实体识别的工作很少。文献[1]利用简单布尔分类器进行英文产品名识别,类似字符匹配模式,性能受到限制。文献[2],该方法过分依赖于句法解析器性能,移植性较差。文献[7]采用自举(bootstrapping)学习方法进行英文命名实体识别,在产品名命名实体(相当于本文的 PRO)识别中获得 69.8 %的 F 值。

尽管产品命名实体识别刚刚起步,近年来在常规命名实体识别方面已经开展了大量的研究工作。汉语常规命名实体识别研究相对英文起步较晚,目前主流的汉语命名实体识别方法以统计学习为主,趋于各种方法的融合。表 1 从三个方面对其进行了比较。

表 1 几个汉语命名实体识别系统的比较

汉语 NER 系统	统计模型	特征选择	融合策略
文献[4]	HMM	语义角色,词形	模式规则
文献[3]文献[9]	基于类的语言模型(LM)	词形,命名识别类别	线索词表,人工知识
文献[5]	基于类的 LM	词形,部分词性	
文献[6]	最大熵模型	词形	知识描述框架(深加工)

3 产品命名实体识别任务分析

3.1 产品命名实体界定

对于产品命名实体很难给出很确切的定义。我们认为,一个产品命名实体应由品牌名称、产品型号和产品类别词以及其他一些产品属性信息组成。例:“摩托罗拉 V8088 折叠手机”这个产品命名实体中,“摩托罗拉”是品牌名,“V8088”是产品型号,“手机”是产品类别词,“折叠”为属性信息。

在真实文本中,一个名词性结构需要含有以下确定性产品信息,才可以构成产品命名实体。

(1) 含有产品品牌或者型号实体任何一个或者两个,如:“爱国者闪存”是一个产品名实体,而“数码相机产品”则不是一个产品名实体;

(2) 尽管没有含有品牌或者型号信息,但是含有某种品牌所特有的产品系列或者版本信息,如:“EasyShare 系列数码相机”是一个产品名实体,因为 EasyShare 是柯达品牌所独有的系列,而“智能型手机”则不是,因为“智能型”是所有品牌可共有的属性信息。

本文研究的是产品实体(PRO)以及品牌实体(BRA)和产品型号实体(TYP)的识别,其中品牌实体和产品型号实体可能嵌套在产品实体中,品牌实体也可单独出现。

例如:①明基/BRA 品牌市场占有率稳步上升。

②佳能/ORG 即将推出[Canon/BRA 334 万像素数码相机 Pro90IS/TYP]/PRO。

3.2 产品命名实体识别难点分析

产品命名实体和常规命名实体相比存在许多形式和结构上的差异,主要体现在:

(1)同人名、地名、机构名等常规命名实体相比,产品名实体上下文环境中不存在一些特定的特征词或者线索词。例如:“区”、“先生”、“街”、“公司”等特征词对于常规命名实体的识别有很好的表征作用。但是,产品实体的上下文环境中很少有这样的特征词语,因此产品名实体识别中候选实体触发难以控制,边界歧义复杂度增加。

(2)产品命名实体中更加多样化的内部类别歧义、以及同常规命名实体之间的类别歧义给产品命名实体识别带来了更严峻的技术挑战。

●品牌词可能充当品牌实体、机构名实体、地名实体、人名实体、普通词;

例如:①“苹果”可以是品牌,也可以是表示水果的普通名词或者公司名称;

②“青岛”可以是品牌,也可以是地名。

●英文单词可能是一个普通词,可能是英文品牌,也可能是产品型号实体的构成部分;

例如:①数字专业编码(Digital Professional Encoding);

②[佳能/BRA DIGITAL IXUS 40/TYP]/PRO。

●数字串可以是产品型号实体的组成部分,也可以是时间或者数量表达式。

例如:①[诺基亚/BRA 8850/TYP]/PRO;

②2003 - 06 - 23/TIM。

(3)同常规命名实体相比,产品命名实体形式更加灵活。产品命名实体灵活表现形式不仅包括位置结构上的多变,还包括语言表达上的各种变异和省略现象。

例如:①柯达 600 万象素数码相机 DX7630;

②600 万象素数码相机柯达 DX7630;

③柯达 DX7630。

3.3 本文思路

(1)控制产品命名实体候选触发:我们依靠知识库结合一些启发式规则进行触发。考虑到系统的可移植性,我们选择容易获得的知识库(自动下载品牌列表)和规则(自动抽取)。

(2)利用统计模型融合各种知识消歧:产生候选之后,选择一个适合的统计模型综合运用实体内部和外部的词汇、语法、语义等各个层次的上下文特征进行统计消歧。由于产品命名实体具有很明显的嵌套特性,内部组成、长度灵活多变,我们选用层级隐马尔可夫模型(HHMM[8]),因为 HHMM 对不同尺度的、多层次的嵌套序列具有更强的描述和表征能力。

4 基于层级隐马尔可夫模型的产品命名实体识别

4.1 系统流程

(1)预处理:对输入文本运用已有工具进行分词、词性标注、常规命名实体识别。

(2)候选实体产生:首先根据品牌列表产生品牌实体、机构名实体的候选,这样可以处理品牌词作为公司名的情况;其次为了触发产生型号实体候选我们定义了六种型号特征类(TCC):纯英文字母(YZ)、字母数字(ZS)、纯数字(SH)、全角数字(QS)、全角字母(QZ)以及其他。根据 TCC 产生型号实体候选;最后根据型号实体、品牌实体、以及诸如“版”、“型”、“系列”等特征词触发产生候选产品命名实体。

(3)统计消歧:根据第二步生成的 HHMM 动态拓扑结构,运用 viterbi 算法进行寻优从而得到最后的识别结果。

4.2 基于 HHMM 的产品命名实体识别模型

我们的输入序列是经过预处理以后的文本,形式化为 $w_1/t_1, w_2/t_2, \cdots, w_i/t_i, \cdots, w_n/t_n$,其

中 w_i 表示输入序列中的第 i 个词，t_i 表示第 i 个词的词性(标记集包括北京大学常规词性标记集{ *POS*}和人名、地名、机构名、时间、数量词五类常规实体类{ *GEN*})，n 表示词的个数。我们建立状态集合 $S = \{\{GEN\}, PRO, BRA, TYP, \{V\}\}$，其中 *PRO*、*BRA*、*TYP* 三种状态对应三类产品命名实体，*V* 为词表，同时建立观察序列集合 $O = \{V\}$，即可构成层级隐马尔可夫模型。这里，只有 *PRO* 为中间状态，其余状态均为产生状态[8]，不属于实体内部的词均单独作为一个状态。为了同 *S. Fine* 对 *HHMM* 的描述一致[8]，用 $q_i^d (1 \leq d \leq D)$ 表示第 d 层子模型的第 i 个状态。

产品命名实体识别的过程是：给定输入序列 $W = w_1, w_2, \cdots, w_i, \cdots, w_n$，触发产生各类实体候选形成 HHMM 的拓扑结构，然后运用 viterbi 算法寻找概率最大的状态激活序列(state activation squence) Q^*，从而得到识别结果。根据贝叶斯公式我们不难得到($P(W) = 1$)：

$$Q^* = \arg\max_Q P(Q \mid W) = \arg\max_Q P(Q)\ P(W \mid Q) \tag{1}$$

该式描述了从 HHMM 根节点开始根据水平转移和垂直转移概率逐步激活各层中间、产生状态节点直至产生观察序列的过程。以第 k 层为例(设由 $k-1$ 层第 m 个状态触发)，右边可写成：

$$P(Q) \cong \underbrace{p(q_1^k \mid q_m^{k-1})}_{\text{vertical transition}} \underbrace{P(q_2^k \mid q_1^k) \prod_{j=3}^{|q^k|} P(q_j^k \mid q_{j-1}^k, q_{j-2}^k)}_{\text{horizonal transition}} \tag{2}$$

$$P(W \mid Q) = \begin{cases} \cong \prod_{j=1}^{|q^k PS|} P([w_{q_j^k\text{-}begin} \cdots w_{q_j^k\text{-}end}] \mid q_j^k) & \text{如果 } q_j^k \text{ 为产生状态} \\ \text{递归激活下一层状态} & \text{如果 } q_j^k \text{ 为中间状态} \end{cases} \tag{3}$$

其中 $|q^k|$ 为第 k 层模型的状态个数；$|q_{PS}^k|$ 为第 k 层模型的产生状态数；$w_{q_j^k\text{-}begin} \cdots w_{q_j^k\text{-}end}$ 表示和状态 q_j^k 对应的词序列。

①若 $q_j^k \in \{\{GEN\}, \{V\}\}$，相信文本预处理结果，则 $P([w_{q_j^k\text{-}begin} \cdots w_{q_j^k\text{-}end}] \mid q_j^k) = 1$。

②若 $q_j^k = BRA$，由于品牌词不仅可以产生品牌候选，还可以产生公司名候选，而且品牌词中用字比较灵活，缺乏规律性，这里我们令 $P([w_{q_j^k\text{-}begin} \cdots w_{q_j^k\text{-}end}] \mid q_j^k = BRA) = 0.5$。

③若 $q_j^k = TYP$，由于产品型号实例可重复性较差，我们采用 4.1 中定义的型号特征类(*TCC*)来进行计算。即将型号候选的组成部分映射为 *TCC* 组合，作为观察序列计算发射概率，即

$$P([w_{q_j^k\text{-}begin} \cdots w_{q_j^k\text{-}end}] \mid q_j^k = TYP) \cong p(tc_1 \mid begin)\ P(end \mid tc_{|q_j^k|}) \prod_{m=2}^{|q_j^k|} p(tc_m \mid tc_{m-1}) \tag{4}$$

④若 $q_j^k = PRO$，由于 *PRO* 是中间状态按照(3)式将激活下一层状态，实现 HMM 的嵌套。

HHMM 中参数可用最大似然方法从训练语料库获得，其中采用 Jelinek and Mercer 提出的删除插值法进行了参数平滑。

4.3 HHMM 模型的融合

上节介绍的基于词形信息的层级隐马尔可夫模型(HHMM - 1)中，没有用到词性信息，为此我们基于词性信息构建层级隐马尔可夫模型 HHMM - 2，并期望通过两个模型的融合实现实体内外多层次特征的有效利用。

同样是上面的输入序列，我们将词性序列看作观察序列，构建另一个层级隐马尔可夫模

型。即将 $T = t_1, t_2, \cdots, t_i, \cdots, t_n$ 看作是观察序列,相应观察序列集合和状态集合为 $Q_{II} = \{POS\}, S_{II} = \{\{POS\}, \{GEN\}, BRA, TYP, PRO\}$。同样对于 HHMM - 2 有

$$Q_{II}^{*} = \arg\max_{Q_{II}} P(Q_{II} \mid T) = \arg\max_{Q_{II}} P(Q_{II})\ P(T \mid Q_{II}) \tag{5}$$

HHMM - 1 侧重使用词形信息,从而模型具有很好的区别性;HHMM - 2 侧重词性更大颗粒度的特征,使得模型具有很好的鲁棒性。很自然,我们期望通过融合两个模型来提高产品命名实体识别的整体性能。融合模型的对数形式如式(6),其中 β 为可调参数。

$$(Q^{*}, Q_{II}^{*}) = \arg\max_{Q, Q_{II}}\{\log(P(Q)) + \log(P(W \mid Q)) + \beta[\log(P(Q_{II})) + \log(P(T \mid Q_{II}))]\} \tag{6}$$

5 实验结果及分析

5.1 实验数据集

实验采用 Casia_Pro1.2 语料库,规模 1,000,000 字左右,包括通讯、电子数码两个领域共 1,500 个网页文本。从 Casia_Pro1.2 中任意抽出 140 个文本作为本文开放测试数据集 OpenTestSet,其余作为训练数据集 TrainSet。从训练集中任意抽取 160 个文本作为封闭测试集 CloseTestSet。各语料库实体分布情况见表 2。

表 2 语料中各实体分布情况简表

语料库	PRO	BRA	TYP	PER	LOC	ORG
Casia_Prol.2	12,432	5,047	10,606	424	1,733	4,798
Open TestSet	1800	803	1364	39	207	614
Close TestSet	1553	513	1296	55	248	619

5.2 实验结果

产品实体表达形式非常灵活,有些实体边界尽管和答案不一致,但是也比较合理或可以接受,因此我们采取了软评测方法,对位置检测、类型识别均正确而边界错误的情况在原有基础上也给予一定的折扣分。本文实验如果不加特殊说明,识别结果均为开放测试结果。

5.2.1 模型融合参数 β 对系统性能的影响

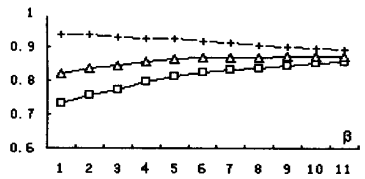


图 1 产品实体识别性能随β的变化曲线

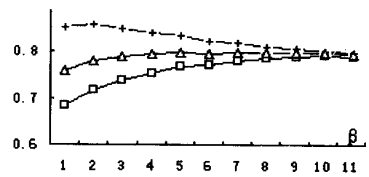
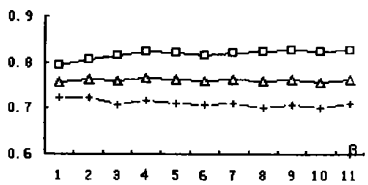


图 2 产品实体识别性能随β的变化曲线



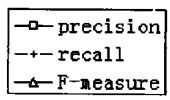


图 3 产品品牌实体识别性能随β的变化曲线

融合模型中 β 值越大 HHMM - 2 在系统识别中贡献就越大。图 1 - 3 分别为系统对产品

三类实体识别性能随着 β 值的变化曲线。图表中 $\beta=1$ 表示 HHMM - 1 和 HHMM - 2 的均等融合,可以看出,随着 β 值的提高,产品名实体和产品型号实体的 F 值呈上升趋势,在一段平稳后开始小幅度下降。数据表明 HHMM - 2 对 HHMM - 1 起到了互补作用,系统整体性能得到提高,这是因为,HHMM - 2 采用了词性信息,而 HHMM - 1 采用了词形信息,因此从模型描述问题的精确度、区分度来说,HHMM - 1 要好一些,而 HHMM - 2 则更能有效的缓解数据稀疏问题对系统带来的不利影响。而且由于目前产品标注语料有限,数据稀疏问题比较突出,从而使得 HHMM - 2 的作用得到了凸现。目前系统中 β 取值为 8。因此,HHMM - 2 和 HHMM - 1 融合可以起到很好的互补作用,在目前语料有限的情况下,HHMM - 2 占有更重要的地位。

5.2.2 识别系统的领域移植性测试(开放测试、封闭测试)

从表 3 和表 4 中可以看出,识别系统在手机领域和电子数码领域均取得了比较满意的结果,表现出一定的领域可移植性。

表 3 电子数码领域识别结果 $\beta=8$

产品实体	封闭测试			开放测试		
	Precision	Recall	F - measure	Precision	Recall	F-measure
产品名	0.864	0.799	0.830	0.762	0.744	0.753
产品型号	0.903	0.906	0.905	0.828	0.944	0.882
产品品牌	0.824	0.702	0.758	0.723	0.705	0.714

表 4 手机领域识别结果 $\beta=8$

产品实体	封闭测试			开放测试		
	Precision	Recall	F-measure	Precision	Recall	F-measure
产品名	0.917	0.935	0.926	0.799	0.856	0.827
产品型号	0.959	0.976	0.967	0.842	0.886	0.864
产品品牌	0.911	0.741	0.818	0.893	0.701	0.785

5.2.3 三种模型以及与最大熵模型的比较

该实验中,我们不仅比较了本文三种模型的性能,还和传统的最大熵模型做了比较。我们使用 Maxent 最大熵工具包(http://homepages.inf.ed.ac.uk/s0450736/maxent_tool-kit.html)进行了实验。因为产品的三类实体的嵌套标注特点,我们训练了两个最大熵模型,顺序进行内层的品牌、型号标注和外层的产品名标注。表 5 中,“1”表示 HHMM - 1 模型,“2”表示 HHMM - 2 模型,“1 + 2”表示二者的融合模型,“ME”表示最大熵模型顺序识别结果。

表 5 模型识别性能比较

HHMM	产品名实体			产品型号实体			产品品牌实体		
	precision	recall	F - 值	precision	recall	F - 值	precision	recall	F - 值
1	0.63	0.84	0.718	0.70	0.94	0.800	0.74	0.73	0.737
2	0.83	0.70	0.760	0.93	0.78	0.851	0.83	0.68	0.743
1 + 2	0.78	0.81	0.797	0.84	0.90	0.869	0.82	0.70	0.758
ME	0.81	0.59	0.683	0.82	0.43	0.564	0.58	0.62	0.60

(1) 三个 HHMM 模型中,HHMM - 1 和 HHMM - 2 在召回率和精确率上是互补的,通过模型的融合,三类实体的 F 值均有较明显的提高。另外,从整体上看,HHMM - 2 从整体性能上要好于 HHMM - 1,这也说明目前语料不充足的情况下,HHMM - 2 在融合模型中应该占据更重要的位置,和实验 5.2.1 是一致的。

(2)对于最大熵模型,我们在特征选择上保持同 HHMM 的融合模型一致,利用了词、词性的窗口上下文特征(窗口宽度 5)、前一个位置标记特征、以及词表层特征(数字字母组合特征),同样也利用了外部知识库(品牌列表),但实验结果表明,HHMM 融合模型在整体性能上要优于顺序识别的最大熵模型。一方面来源于顺序识别造成的累积错误蔓延,层与层之间无法形成互补;另一方面也说明 HHMM 模型由于可以很好的综合各层内部以及层与层之间的约束,两层之间在某种程度上形成互补,对嵌套的、内部组成和长度均灵活多变的产品命名实体来说,具有更好的表征能力。

6 结束语与展望

本文系统分析了产品命名实体识别任务的特点和难点,并提出了一种基于层级隐马尔可夫模型的产品命名实体识别方法,该方法通过融合两个统计模型以及同知识库、启发式规则的有机结合,综合利用了不同层次的上下文特征进行产品命名实体识别,在电子数码和手机领域均取得了令人满意的效果。

目前系统识别性能仅限于小规模测试语料,仍需要更深入细致的研究和大量的后续工作:

- 研究语言模型的长距离依存信息在产品实体识别中的应用;
- 实现分词、词性标注、常规命名实体识别一体化的框架,避免错误的蔓延;
- 进一步修改标注规范,增强训练语料的标注一致性。

参 考 文 献:

[1] John M. Pierre. Mining Knowledge from Text Collections Using Automatically Generated Metadata [A]. In: Proceedings of Fourth International Conference on Practical Aspects of Knowledge Management [C]. London, UK: Springer-Verlag, 2002, 537 - 548.

[2] Bick, Eckhard. A Named Entity Recognizer for Danish[A]. In: Lino et al. (eds.), Proc. of 4th International Conf. on Language Resources and Evaluation (LREC2004) [C], Lisbon, 2004, 305 - 308.

[3] Jian Sun, Jianfeng Gao, Lei Zhang, Ming Zhou, Changning Huang. Chinese Named Entity Identification Using Class-based Language Model [A]. In: Proceedings of the 19th international conference on Computational Linguistics [C]. Morristown, NJ, USA, Association for Computational Linguistics, 2002, 1 - 7.

[4] Huaping Zhang, et al. Chinese NER Using Role Model [J]. Special Issue of the International Journal of Computational Linguistics and Chinese Language Processing, 2003, 8(2): 29 - 60.

[5] Guohong Fu and Kang-Kwong Luke. Chinese Unknown Word Identification Using Class-based LM[A]. In: Proceedings of the First International JointConference on Natural Language Processing (IJCNLP - 04) [C]. Hainan, China, 2004, 262 - 269.

[6] Tzong-Han Tsai, et al. Mencius: A Chinese Named Entity Recognizer Using the Maximum Entropy-based Hybrid Model [J]. International Journal of Computational Linguistics & Chinese Language Processing, 2004, 9(1): 62 - 82.

[7] Cheng Niu, Wei Li, Jihong Ding and Rohini K. Srihari. A Bootstrapping Approach to Named Entity Classification Using Successive Learners [A]. In: Proceedings of the 41st ACL [C], Sapporo, Japan, 2003, 335 - 342.

[8] Shai Fine, Yoram Singer, Naftali Tishby. (1998) The Hierarchical Hidden Markov Model: Analysis and Applications [J]. Machine Learning. 1998, 32(1): 41 - 62.

[9] Y. Z. Wu, J. Zhao, B. Xu. Chinese Named Entity Recognition Combining Statistical Model with Human Knowledge [A]. Workshop of 41st ACL: Multilingual and Mix-language NER [C], Sapporo, Japan, 2003, 65 - 72.