在篇章中面向产品类的命名实体识别研究

李治国[1]，周俏丽[1]

（沈阳航空工业学院自然语言处理实验室，沈阳， 110034）

摘 要： 命名实体识别是中文信息处理的一个重要问题。本文根据篇章中利用互信息识别命名实体的方法，引入词性互信息和有距离的匹配策略，通过词表层信息和有距离匹配策略的融合方法识别出面向产品类的命名实体。同时融入一些知识和规则方法，取得了很好的效果。

关键词： 产品类 命名实体 互信息 匹配模式

Study on Named Entity Recognition for Product Class in the Text Information

LI Zhiguo[1], ZHOU Qiaoli[1]

(Shenyang Institute of Aeronautical Engineering,Shenyang 110034 ,China)

Abstract: It is a import problem to identify Named Entities in the Chinese Information Processing, By mean of Using Mutual Information Recognizing Named entity this paper presents a method of that importing POS information recognizes Class-based Product through POS information sequence and searching strategy which have some distance in the text information. By surface Information and having-distance pattern, Named Entity Recognition which includes knowledge and some rules get well effect.

key words: Class-based Product Named Entity Mutual Information Matching Pattern

1 引言

随着新产品新技术日新月异的发展，新品种新型号的产品层出不穷，越来越多地融入到这个信息时代当中，这就为中文信息处理提出了一个新的命题----面向产品类信息抽取的命名实体识别。对于在篇章信息中以产品系列（同类产品不同型号）为主要特点的命名实体，仅仅靠常规命名实体识别无法有效的地抽取文本中的关键信息，尤其是篇章中产品的相关信息，所以面向产品类的命名实体识别研究势在必行。

本文对面向产品类信息抽取的命名实体进行了定义描述，对面向产品类的命名实体识别的特点和难点进行了分析，提出了一种基于词表层信息和篇章中的关联信息相融合的方法。该方法通过篇章中的文字信息的词形、词性和统计词频的融合，综合了一些语言的知识，取得了较为满意的效果。

作者简介：伍建军（1982-），男，湖南祁阳人，硕士研究生，研究方向为 Internet 信息检索、数据挖掘

E-mail： happier5281@yahoo.com.cn ；

康耀红（1963-），男，陕西韩城人，教授，博士生导师，研究方向为 Internet 信息检索等

2 相关工作

命名实体（Named Entity， 简称 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[10]。

NE 识别是目前公认的中文自动切词的难点。对中文来说，因为一方面他没有空格标志词语边界，另一方面也没有词语的明确定义，所以中文 NE 识别比英文 NE 的识别要困难。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；另一类是用统计或学习的方法对各种命名实体做一揽子处理。它的健壮性和灵活性都比规则的方法好，且代价小，但需要大规模的语料进行训练。具有代表的机器学习算法有 HMM，Maximum Entropy，Memory-based learning，Support Vector Machine，Mutual Information 等。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[10]。

3 面向产品类的命名实体识别任务分析

3.1 问题提出

随着互联网越来越多的融入到我们的生活中，我们接触着越来越多的信息，尤其一些产品不断更新替换，而且不同地方的人们联系也越来越多，同时从不同的环境中对于同一类的产品会给我们带来了不同的信息，所以，我们在浏览一些网页和文章时总是发现在这些文本信息中总有一些命名实体存在着不同的称谓、不规范的书写、以及具有系列型的产品命名实体，例如，在有关“歼 8 传”和“神舟笔记本电脑推荐”的文章中，我们可以看到对于两种产品命名实体的描述：

表 1 关于“歼 8 传”的一篇文章部分描述

歼 8 飞机的试飞与攻关

“文化大革命”的动乱使 *歼 8 飞机*的研制深受其害。...*歼 8* 研制过程一再受阻...

...保护了 *歼 8 飞机*免遭夭折。...

...

*歼 8I 飞机*的研制

*歼 8 型飞机*是白天型飞机 ...从一九七六年起开始研制 *歼 8I 型飞机* ...

...

...最快的速度又装出一架 *歼 8I 型飞机* ...第二架 *歼 8I* 上天。...

...

...航空产品定型委员会正式批准 *歼 8I 型*设计定型 ...

在表 1 里我们发现：“歼 8” 、“歼 8 飞机” 、“歼 8 型飞机” 、“歼 8I 型飞机” 、“歼 8I” 、“歼 8I 型”等命名实体由于不规则的书写、不规范的称谓、产品的系列型号、以及三者中同时的作用，造成了同一命名实体的不同词形，为命名实体的识别带来了困难。

表 2 关于“神舟笔记本电脑推荐”的一篇文章中的部分描述

…*神舟优雅 Q300T*：64 位 512M 内存 13.3 寸本仅 5388 元 ，...

…、*同方 V30*：4999 元超值之选 还有可能看世界杯， ...

…*长城 E530*：14 寸宽屏镁铝合金本仅 5199 元，…

TCL K40:512M 内存 60G 硬盘宽屏本仅 5998 元，…

… *华硕 A6517CR-D* 采用了 Intel Celeron 390 处理器 …

…这款*海尔 H30* 在市场上很受欢迎…

…*方正 R350* 采用了 Intel 赛扬 M 370 处理器…

…*宏碁 TM 2424NWXM* 不仅采用 60GB 大容量硬盘…

…*戴尔 Inspiron 1300* 具有 3 个 USB2.0 ...

…

… *优雅 Q310Y* …. *承运 L420E* ， … *承运 B380R* …

在表 2 中我们同样可以看到：“神舟优雅 Q300T” 、“同方 V30” 、“长城 E530”等命名实体都是属于笔记本一类或者系列的产品，在其他的文章中我们找到了许多类似“神舟天运 Q230N” 和“神舟天运 420S” 、诺基亚手机中的“诺基亚 N-GAGE” 、“诺基亚 N-GAGE II”和“诺基亚 N-GAGE QD”等都属于一系列的产品或一类的产品命名实体，这种类型的命名实体在大量的文献篇章网页信息中出现过，它们有一个共同的特点就是经过分词后属于同一种句法结构。所以对于这类命名实体的识别不能用常规的命名实体识别方法去处理，而是需要一种融合的策略。

3.2 面向产品类的命名实体的界定

在文献〈1〉提出：一个产品的命名实体由品牌名称、产品型号和产品类别词以及其他一些产品属性信息组成。在真实文本中，一个名词性结构需要含有两个确定性产品信息，才可构成产品命名实体：（1）含有产品品牌或者型号实体任何一个或两个；（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。由于面向产品类的产品命名实体和常规命名实体相比存在许多形式上和结构上的差异，所以，本文针对这一特征，采用多策略融合识别的方法来实现对于面向产品类的命名实体识别。在这里我们提出了面向产品类的命名实体的定义描述：

（1）对于同一产品命名实体的不规则的书写（“歼 8 飞机” 写成“歼 8 飞机”、“歼八飞机”）；

（2）对于同一产品命名实体的不规范的称谓（“歼 8”、“歼 8 型飞机”都是对“歼 8 飞机”的不规范称谓）；

（3）由同一类产品命名实体衍生（“歼 8I 飞机”、“歼 8II 飞机”）；

（4）由（1）（2）（3）共同作用而产生的产品命名实体（“歼 8I 型飞机”、“歼 8I 型”）；

文献〈12〉提出根据字符匹配的方法计算候选命名实体的在篇章中的互信息，我们在此基础上根据产品类命名实体的特点加入词性匹配的信息，通过计算产品类命名实体的词性互信息，计算出产品类的命名实体。为此，本文对产品命名实体的词性进行了具体的分析，见表 3

表 3 面向产品类的命名实体结构词性标注分析

	产品类实体分析	词性分析	举例说明
1	产品实体（PRO）	{n}/名词	飞机/n ， 手机/n ， 笔记本/n
2	品牌实体（BRA）	{n}/名词，{nz}/其他专名，{nr}/人名，{ns}/地名，{nt}/团体机构，{j}/简称	诺基亚/n ， 联想/n，清华/n，摩托罗拉/n，西门子/n， …
3	型号实体（TYP）	{m}/数词，{q}/量词，{b}/区别词，{k}/候接成分，{ng}/名素语，{nz}/其他专有名词 ，{n}/名词	8I/m ， 型/b ， 同方 *V30*/n …
4	结构实体（STR）	{v}/动词，{a}/形容词，{vg}/动语素 ，{n}/名词	*歼*/vg 8 飞机 *超音速*/n 飞机
5	实体边界（BBB）	{u}/其他助词，{v}/动词，{c}/连词， {d}/副词，{b}/区别词，{p}/介词， {m}/数词，{dg}/副语素词，{w}/标点符号等	的/u 以谓语动词结尾等

4 面向产品类的命名实体识别

4.1 预处理和构建产品类词表

（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）构建面向产品类的词表：对于要进行命名实体识别的产品我们进行了对这类产品词表的建设，首先我们将

产品分成三个目录，第一个目录是总目录用于列举产品类的总称，例如：飞机、电脑、手机、音响等，第二个目录用于每个产品类的各自小类别，例如飞机类中：战斗机、轰炸机、运输机、民航飞机等，手机类中有诺基亚、摩托罗拉、西门子等，第三个目录用于列举第三个目录中具体产品名，例如战斗机中有歼 8 飞机、F111、苏 34 等，诺基亚手机中有诺基亚 N-GAGE 等。我们做的工作就是将完善第二个目录的中产品类命名实体抽取，对于符合要求的我们都将识别出来的命名实体放在相应的目录下，通过一小部分命名实体“种子”，相应识别出更多的同类命名实体，同时完成了产品类词表的自动更新。

4.2 面向产品类的命名实体识别模型

$$I(x,y)=\log\frac{p(xy)}{p(x)p(y)} \tag{1}$$

信息论中的互信息是衡量两个信号的关联尺度，后来引中为对两个随机变量间的关联程度进行统计描述，如上公式中，x ，y 表示两种模式，p（x），p（y）表示两种模式在篇章中出现的频次概率，p（xy）表示模式 xy 在篇章中出现的概率。I（x，y）>>0，表示 x，y 关联的程度强，I（x，y）≈0，表示 x 和 y 的关联程度弱，I（x，y）<<0，表示他们不存在关联关系。具体的点互信息描述参见文献〔3〕。由于本文计算的互信息是针对命名实体内部的互信息，计算互信息的值相对来说不是很大，在做实验时对互信息的阀值作了一个假设，只要 I(x,y)>0.1 我们就认为 x 和 y 模式具有关联性。

a） 利用词形之间的互信息

本文通过计算词形之间的互信息来进行产品类的命名实体识别。以表 1 中的文章为例，首先我们从产品类词表中找到“歼 8 飞机”的词条，根据分词工具将“歼 8 飞机”分成“歼”、“8”、“飞机”三个部分，统计全文词数有 6412 个词，其中“歼”在全文中出现 37 次，“8”在全文中出现 37 次，“歼 8”在全文中出现 34 次，计算“歼”与“8”的互信息：

$$I(x,y)=\log\frac{\frac{34}{6412}}{\frac{37}{6412}*\frac{37}{6412}}=7.3129 \tag{2}$$

这说明“歼”与“8”的关联程度很强，两者可以合为一个模式，所以“歼 8”是一个候选命名实体模式。依次递推再计算“歼 8”和“飞机”的互信息为 6.2854。同样，互信息也很大，所以“歼 8 飞机”也成为一个候选模式。在这里，本文提出这种纯字符的词形匹配不能够识别出“歼 8 型飞机”、“歼 8I 飞机”、“歼 8 飞机”等类的实体。所以本文提出进行词性互信息与策略相融合的计算方法可以解决以上出现的问题。

b） 利用词性之间的互信息

$$I^{*}(t_1{}^{*}t_2)=\log\frac{p(t_1{}^{*}t_2)}{p(t_1)p(t_2)} \tag{3}$$

本文在通过对篇章中的词性序列观测计算互信息来对产品类命名实体进行识别。t_1，t_2 分别代表公式(1)中 x，y 对应的词性，$t_1{}^{*}t_2$ 表示改进互信息采用词性匹配的模式，计算 $t_1{}^{*}t_2$ 模式在篇章中的频次概率时，采用 3 种词性模式匹配的策略，具体的策略介绍见 4.3 中的面向产品类的命名实体识别策略。在这里我们再次利用改进的互信息，通过词性计算互信息，在表 1 中“歼/vg 8/m 飞机/n”中的“vg”词频数是 47 次，“m”词频数是“98”次，“vg m”词频出现的次数“29”次，它的互信息为：

$$I^{*}(t_1{}^{*}t_2)=\log\frac{\frac{29}{1678}}{\frac{98}{1678}*\frac{47}{1678}}=3.3219 \tag{4}$$

说明互信息强，“vg m”模式可以认为是一种候选实体模式，同理我们也可以计算出“vg m”和“n”的互信息为1，互信息也较强，“vg m n”模式也可以认为是一种实体后选模式。但由于$I^*(vg,m) >> I^*(<vg\ m>,n)$，所以两种模式可以认为是独立存在的。我们在采用词性匹配的同时，融合了一些策略，这样也增强了产品类的命名实体识别的深度和宽度，以下介绍融合面向产品类的命名实体识别的策略。

4.3 面向产品类的命名实体识别策略

文献〈3〉和〈12〉都提出了利用篇章的互信息进行命名实体识别的方法，基本采用的策略是利用字符的匹配统计模式在篇章中出现的次数，并且取得了很好的效果。本文在此基础上加入了词性的信息，通过对词性在篇章中的序列观测，计算在篇章中共现情况，得出互信息，最后观测的词性序列确定影射到字符的匹配，完成面向产品类的命名实体的识别。

我们对篇章中的文本通过已有的工具进行分词，形式化为 w1/t1 ， w2/t2 ， w3/t3，… wi/ti ，… wn/tn，其中 wi 表示文本中第 i 个词，ti 表示第 i 个词的词性(标记集使用的是北大语言所的汉语文本词性标注标记集)，n 表示词的个数。通过对大量网页文本分析可知，产品实体一般出现在最后，品牌实体、类型实体、结构实体三者没有先后顺序，品牌可以不用出现，但是产品型号和属性一定要出现。同时文献〈1〉为我们提供了产品命名实体难点分析。我们结合难点分析提出了几种匹配策略。

首先，根据表〈3〉建立了 5 个状态集合{PRO，BRA，TYP，STR，BBB}，其中每个状态对应了各自的词性标注集合。面向产品类的命名实体识别的过程是：给出一个观测序列 W=w1 ，w2，…wi…wn，我们对这个观测序列进行分词和标注，建立分词序列{w1 ，w2 ，…wi… ，wn}和词性序列{t1，t2 ，…ti…，tn}，通过在篇章信息中计算词性序列的互信息，搜索对应的词序列，然后找出命名实体。整个过程通过以下识别策略得出命名实体结果：

4.3.1 进行在篇章中词性完全的匹配

利用产品类词表，在篇章中查找与词表中对应的词性序列。例如“歼/vg 8/m 飞机/n”是词表中的一条记录，我们确定词性序列“vg m n ”是面向产品类的命名实体的一个词性序列，我们从文章中找出词性符合“vg m n”的分词序列。

通过 4.2.2 改进互信息模式公式，我们得出“vg m ”与“n”模式的互信息是 1，说明“vg m n”模式是可行的，通过查找“vg m n”对应的词序列，我们就找到了产品类的命名实体，根据表 1 的实验结果：“歼/vg 8/m 飞机/n” 、“歼/vg 8/m□飞机/n”、“歼/vg 8I/m 飞机/n”、“歼/vg 8/m 原型机/n”等都能够被识别出来。

4.3.2 进行在篇章中词性部分匹配

由表 4 的分析我们知道产品类命名实体是由产品实体（PRO）、品牌实体（VRA）、类型实体（TYP）结构实体（STR）组合而成，有的时候并不完全同时出现，可能出现含有产品品牌或者型号实体任何一个或者两个，可能含有某种品牌所特有的产品系列或者版本信息等，所以，从产品实体，品牌实体，类型实体，结构实体中取出部分实体也能构成产品类的命名实体。

我们在“vg m n”模式中取出“vg m”、“m n”、“vg n”组合模式，进行互信息计算，在互信息较大的情况下，取出模式对应的词序列。我们从 4.2.2 中计算“vg m”模式的互信息较大情况下，可以看出进行词性的部分匹配是可取的，这样我们找出的词序列有：“歼/vg 8/m”、“歼/vg 8I/m”等。

4.3.3 进行有距离的词性匹配

当我们进行词性的完全匹配或部分词性匹配时，很有可能漏掉了一些信息，漏掉的信息是状态集合{{PRO}，{BRA}，{TYP}，{STR}}中一个或多个，所以我们采用比较宽松的策略，在对（1）（2）策略进行获取互信息的时候，在匹配模式的前、中、后位置有距离获取词性，重新组合成新的词性匹配模式，再从篇章中获取对应的词序列，这样的策略我们称之为是有距离的词性模式匹配策略。三种有距离的模式可表示为：

$$T(t_i^* t_{i+1}) = T(<t_{i-n}...t_{i-2}t_{i-1}>, t_i, t_{i+1}) \qquad (5)$$

$$T(t_i^* t_{i+n+1}) = T(t_i, < t_{i+1} t_{i+2} ... t_{i+n} >, t_{i+n+1}) \quad (6)$$

$$T(t_i^* t_{i+1}) = T(t_i, t_{i+1}, < t_{i+2} t_{i+3} ... t_{i+n+1} >) \quad (7)$$

其中 T 表示模式，t 表示词性，下标表示词性在词性序列中的位置，试验中 n=1、2，在计算互信息时，由于要计算的实体是稀疏数据，本文采用了平滑的方法，在计算频率时，采用了以下方法：

$$r = r_1 + r_2 \quad (8)$$

其中 r_1 表示在（5）、（6）、（7）下计算模式的频次，r_2 表示在没有距离条件下产生的频次，r 表示前两者的频次之和。以下是具体介绍有距离匹配的方法：

1）在（t_1，t_2）模式位置前有距离获取词性的词性模式匹配方法。当通过策略[1]或[2]进行互信息计算时，如果互信息较高，就把词性序列前几个邻接词性一同与（t_1，t_2）模式取出，构成公式(5)中的模式，计算互信息并把对应的词序列从文本信息中抽取出来。在实验中，我们对表 2 中的信息“同方 V30”、“长城 E530”、“TCL K40”模式词性匹配出“n+n”词性模式，当我们增加距离时获得新的词性模式“n+n+n”，从文本中找到对用的词序列“神舟优雅 Q300T”。

2）在（t_1，t_2）模式位置中间有距离获取词性的词性模式匹配方法。当通过策略[1]或[2]进行互信息计算时，如果互信息较高，就把 t_1 和 t_2 之间含有词性的模式取出，构成公式(6)中的模式，计算互信息并把对应的词序列从文本信息中抽取出来。在实验中我们对表 1 中的信息进行测试，针对“歼/vg 8/m 飞机/n”的模式进行有距离的匹配，其中词性模式为“vg m n”。

首先我们要计算“vg”和“m”的互信息，在上文 4.2.2 中我们计算出 I(vg,m)=3.3219，可以认为“vg m”是一种关联模式，并在“vg m”模式基础上继续计算 I(<vg,m>,n),其中“<vg，m>”在文中出现 30 次，“n”出现 435 次，“vg m n”出现了 22 次，文章中词性的词频出现次数为 1678 次，以此计算出互信息为：

$$I^*(t_1^* t_2) = \log \frac{\frac{22}{1678}}{\frac{30}{1678} * \frac{435}{1678}} = 1.0000 \quad (9)$$

说明“vg m n”模式也是具有较强的互信息。然后，我们在“vg m”和“n”之间加入距离（实验中加入距离为 1），使词性模式变成了词性(6)中的模式，在词性序列中搜索出了“vg m k n”、“vg m nz n”、“vg m n n”、“vg m b n”等词性模式，同时计算模式出现的频次为 r_1=8，r=8+22=30。计算

$$I(T(t_i^* t_{i+n+1})) = \log \frac{\frac{30}{1678}}{\frac{30}{1678} * \frac{435}{1678}} = 1.5850 \quad (10)$$

说明模式（6）互信息增强了，把模式（6）的词性信息从篇章信息中找出词序列：“歼 8 型飞机”、“歼 8I 型飞机”、“歼 8 飞机液压”等。

3）在（t_1，t_2）模式位置后有距离获取词性的词性模式匹配方法。当通过策略[1]或[2]进行互信息计算时，如果互信息较高，就把（t_1，t_2）词性模式与后几个有距离的邻接词性取出，构成公式(7)中的模式，计算

互信息并把对应的词序列从文本信息中抽取出来。

本文在表 2 中的文本信息做实验，计算方法与 I 基本一致，找出的词序列有“歼/vg 8/m 型/n”等。

从试验结果来看，随着距离取值的增长，一些与实体不相关的信息也都加进来了，例如：“歼 8 飞机液压”、“歼 8 飞机尾部”等给实验的结果带来了噪音，如果不给模式加以长距离就不能把“歼教 6 型超音速教练机”等长实体名识别出来。所以存在着相互制约的矛盾。所以实验中采用了一些知识规则和语言知识，从而减少了噪音重现。例如在通过以上三种策略进行匹配时，我们还要考虑实体边界{BBB}的条件，通过实体边界的判断，把一些不符合要求的模式排除出去等。

4.3.4 去重和消伪的工作

1) 本文利用知网的相关性的资源，对命名实体的产品实体（“飞机”、“计算机”等产品）进行相关性查找，抽取出其中的名词词条，做成相关性词表。例如当我们把“飞机”放到知网中查找时，可以查到 198 个相关词，我们按照名词词性进行过滤查到与“飞机”对应的词性词有 177 个。当我们对候选产品命名实体类进行查找时，发现“书生/n 型/k”候选实体与相关性词表没有关系，所以我们就可以去掉这一个候选词。通过试验我们发现简单的利用相关性词表不能完全的识别出命名实体类中的“伪词”，例如“歼 8/n 飞机/n 尾部/n”、“白天/t 型/k 飞机/n”等，通过命名实体相关词表的约束，对面向产品命名实体类的识别只能起到一定的辅助作用，我们还应采用其他的规则共同完成面向产品类命名实体识别的“消伪”任务。

2) 命名实体边界的确定

文献〈1〉中提出：产品命名实体很难给出确切的定义。由于产品命名实体内部多样化，所以不能通过常规的命名实体的边界确定方法去确定产品的命名实体，同样也无法去确定面向产品类的命名实体。我们进行命名实体边界的是通过对篇章中词性序列的观测得出的结论，即满足表 4 中实体边界状态集合中的任意一个状态就可以确认为命名实体的边界，从而识别出面向产品类的命名实体。此外，我们通过一些语言知识规则强化了面向产品类命名实体边界的确定。

5 实验结果及分析

本文进行测试采用的语料来自Internet网页中，包括笔记本、飞机两个领域共100个网页文本，其中笔记本、飞机各占50个文本。测试结果对于“笔记本”类召回率达到91%，准确率达到100%，原因在于命名比较规范基本形式都可归结为“产品品牌+产品类型”，例如在神舟笔记本中：天运Q370S 承运L230E 优雅Q310Y 承运B380R 承运L240R等等，对“飞机”类进行测试时，召回率是87.5%，准确率是100%，召回率有些降低，原因在于在飞机领域中，命名实体表现形式很到多，在同一个文本中存在对同一命名实体的不同称谓也很多，甚至还存在词串较长的命名实体。例如在进行“歼8飞机”的测试中，针对“歼8飞机”的称谓就有：歼8、歼8机、歼8原型机、歼8型机等，存在较长搭配的命名实体有：歼教6型超音速教练机。

考虑的问题比较少，所以随着大语料和多领域的产品命名实体的测试，一定还会出现的更多问题。针对这些现象我们通过分析添加规则可以完善面向产品类的命名实体的识别。从而提高召回率和准确率。同时实验也证明了对于特殊的产品类的命名实体识别通过一定的方法，可以实现在网页信息中的自动抽取。为命名实体识别提供了新的手段和方法。

6 结论和下一步工作

本文在产品命名实体识别的基础上，提出了面向产品类的命名实体识别，给出了一个通过篇章中利用互信息进行产品类的命名实体识别的方法，目前来说，命名实体的方法很多，单纯的依靠一种方法和策略不能满足目前的命名实体的识别任务，并且命名实体的研究方法，也都不是很深入，所以只有深入研究命名实体的方法，采用多种方法结合的策略才是命名实体识别的正确道路。

由于命名实体识别的起步较晚，识别技术还很不完善，面向产品类的识别也是在初步的探索之中，所以仍需要更加深入细致的研究和大量的后续工作：

. 在大语料中进一步进行测试，完善系统；

. 在统计篇章中的信息中，还需考虑上下文的信息；

. 还需要更多的词义支持；
. 对于允许范围内的匹配技术改进。

参考文献：

[1] LIU Fei-fan，ZHAO Jun，LV Bi-bo,Xu Bo,Yu Hao,XIA Ying-ju: Study on Product Named Entity Recognition for Business Information Extraction [J].In: Journal of Chinese Information Processing, 2006 Vol.20 No.1

[2] Jian Sun , Jianfeng Gao , Lei Zhang ,Ming Zhou ,Changning Huang. Chinese Named Entity Identifucation 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.

[3] HUANG De-gen, Ma Yu-xia, YANG Yuan-sheng: Chinese names identification based on mutual information [J]. In: Journal of Dalian University of Technology,2004 Vol.44 No.5.

[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] Butte AJ.Kohane IS.MuTual Information Relevance NetWorks: Punctional Genomic Clustering Using Pairewise Entropy Measurements.Pacific Symposium on Biocomputing (PSB 2000).

[6] Chen Xiaohe. A package scheme for identifying unlisted words in Chinese segmen –tation [J] : Application of Language Letter , 1999,vol31.

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

[8] 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.

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

[10] Xiantao Liao Haibin Yu Bing Qin Ting Liu. HMM combined with automatic rules-extracting for Chinese Named Entity recognition [A]. In：Proceedings of the 2nd SWCL [C]，Beijing，China，2004，232 – 237.

[11] 汉语文本词性标注标记集 ，北京大学计算语言所。

[12] 郭志立 ：使用互信息辅助在篇章范围内识别命名实体。语言计算与基于内容的文本处理。北京：清华大学出版社.2003.