EntNER

企业信息的命名实体识别工具

|  |  |
| --- | --- |
| 版本 | 0.1 |
| 作者 | 张知临 |
| 联系方式 | zhzhl202@163.com |
| 最后更新 | 2012/03/22 |
| 系统主页 | [http://code.google.com/p/entner/](http://code.google.com/p/tmsvm/) |

目录

[命名实体识别 1](#_Toc320218752)

[基础 1](#_Toc320218753)

[基本概念 1](#_Toc320218754)

[难点 1](#_Toc320218755)

[现有方法 2](#_Toc320218756)

[系统中的识别算法 3](#_Toc320218757)

[企业注册信息识别 3](#_Toc320218758)

[企业简称识别 5](#_Toc320218759)

[日期、货币和地点识别 6](#_Toc320218760)

[商标信息识别 6](#_Toc320218761)

[使用说明： 7](#_Toc320218762)

[参考文献 8](#_Toc320218763)

命名实体识别

命名实体识别（Named Entity Recognition ,NER）是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等，该技术也是自然语言处理中的重要基础技术。命名实体识别的过程通常包括两部分：（1）实体边界识别；（2）确定实体类别（人名、地名、机构名或其他）。

# 基础

## 基本概念

命名实体是文本中承载信息的重要语言单位。按照 Automatic Content Extraction (ACE) 评测计划的定义，实体概念在文本中的引用（也可称为指称项）可以有三种形式：命名性指称、名词性指称和代词性指称。例如在句子“ [中国总理][温家宝]29日开始[他]在韩国第二天的访问行程。”中，实体概念“温家宝”的指称项有三个，其中“中国总理”是名词性指称，“温家宝”是命名性指称，“他”是代词性指称。一般实体识别技术研究的对象都是命名性指称。

围绕命名实体有一系列的研究任务，例如：命名实体的识别、排歧、属性抽取、关系抽取等。其中，命名实体识别任务是识别出文本中实体概念的命名性指称项，并标明其类别（例如人名、地名、机构名、产品名等）；命名实体排歧解决的是一个命名性指称项指称多个实体概念的问题以及多个命名性指称项指称同一个实体概念的问题。命名实体的属性抽取，指的是从网页中抽取出特定实体概念的属性类别和值，例如，企业就包含名称、简称、成立日其、注册资金、经营范围、住所等属性；实体关系检测指的是通过分析网页信息判断两个实体是否存在关系，存在什么类型的关系，例如同一段文字中出现的两个企业，他们的关系可以是合作者、厂家与客户、投资者与子公司、竞争者等。命名实体的识别、排歧、属性抽取以及关系抽取等技术在网络信息抽取、网络内容管理和知识工程等领域中占有非常重要的地位。

## 难点

命名实体识别中人名、地名、机构名的用字灵活，识别的难度很大，因此命名实体识别通常指的是人名、地名和机构名的识别。

本课题关注的是中文领域的命名实体识别，相对英语来说，中文命名实体识别更加复杂：1）词的界限不明显，中文所有的字都是没有空隙的，因此分词是一切中文处理的前提；2）实体没有明显的标记，英语中实体的首字母往往是大写的，而汉语却没有类似的标记，实体边界的识别就显得尤为重要[1]。

命名实体识别的主要难点有[1]：

a) 命名实体形式多变：命名实体的内部结构很复杂，有时还存在不同名称表示同一个实体，或者一个实体中嵌套另一个实体等。

b) 命名实体的语言环境复杂：命名实体可能出现在各种语言环境中。同样的汉字序列在不同语境下，可能具有不同的实体类型，或者在某些条件下是实体，在另外的条件下就不是实体。

## 现有方法

有关命名实体识别的研究已有大量成果, 比较常用的是机器学习方法，包括：语言模型方法[3][6]、隐马尔可夫模型[4][9]、最大熵模型[5]等。

文献[3]提出了基于分类的语言模型方法，该方法将中文分词和命名实体识别纳入统一的框架，并且建立一个层次结构的语言模型以便于识别嵌套在某个组织名称中的实体。该模型由两个子模型组成：上下文模型表示在给定的上下文中存在实体的概率，实体模型则用于估算给定字符串属于某个实体类别的概率。该模型对于某些缩写的实体识别有显著的提高。

文献[5]则是采用的最大熵模型。作者不仅利用关键词所在句子的上下文知识，而且从同一文档其他出现该词的地方抽取特征。其最大熵分类器将词根据其在实体内的位置分为：实体的首词、中间词、末位词，以及单个词组成的实体。根据词在实体内的位置进行分类同样可以适用于中文命名实体识别领域。

文献[4]文针对三种重要的命名实体，即人名、地名、组织名，提出了一种隐马尔可夫模型（HMM）和最大熵模型（ME）相结合的汉语命名实体识别的方法。该方法的特点在于：使命名实体识别和词性标注两个任务一体化；融合两种统计模型进行命名实体识别，其中HMM从整体上（句子范围内）对命名实体识别进行约束，ME则在局部范围内（当前词的上下文范围）估计一个词串被标记为某种命名实体的概率。

文献[9]则对HMM进行了深入研究，并进行了一定的改进。HMM是一个双重随机过程，包含马尔可夫链和一般随机过程两个部分，分别描述状态的转移和状态与观察值间的关系。HMM一般使用五元组λ=(N, M, π, A, B)来描述（有时也简写为三元组(π, A, B)），其中N为状态数目，M为观察值数目，πi表示初始时选择状态Si的概率，Aij为状态转移概率矩阵，表示从状态Si转移到Sj的概率，Bik表示状态Si下观察到vk的概率。HMM包含三个基本问题：

1. 评估问题

给定观察序列O=O1O2…OT和模型λ=(π, A, B)，计算P(O | λ)。即给定模型和观察序列，计算从模型生成观察序列的概率。常用算法有向前法、向后法。

1. 解码问题

给定观察序列O=O1O2…OT和模型λ=(π, A, B)，择一个对应的状态序列 S = q1,q2,…qT，使得S能够最为合理的解释观察序列O。常见的有Viterbi算法。

1. 学习问题

如何调整模型参数λ=(π, A, B)，对于一个给定的观察序列O=O1O2…OT，使得P(O | λ)最大。参数训练算法有Baum\_Welch算法、K均值算法等。

文献[8]介绍了一种基于规则的识别方法用于识别组织机构。该方法首先通过训练建立一定规模的规则集，并且构造组织机构后缀词库。在文本中发现后缀词时进行规则匹配，得到若干候选规则后通过一定的策略进行筛选，得到最终的识别结果。

# 系统中的识别算法

该模块主要对企业和商标相关的实体信息进行识别，如企业、日期、地点、商标、产品名等。本系统针对不同实体采用了两种不同的识别方法，分别是基于规则匹配和HMM。其中企业、地点、日期和货币都有明显的特征词和名称规则，因此采用基于规则的识别方法。商品实体中型号和修饰词有着明显的词性特征，因此采用HMM利用其统计信息进行识别，然后再识别出其中的商标和产品名。

## 企业注册信息识别

企业注册信息包括企业名称、日期、地址、货币，采用基于规则的识别方法。首先以指定后缀词激活规则匹配器，然后通过实体名称规则进行匹配和识别。其中以企业名称识别最为复杂，其流程如图 1‑1所示，该方法文献[8]的基础上，将企业名称长度纳入规则决策时的计算因子，并且在最后增加了一个左边界修正的步骤。



图 1‑1 企业名称识别流程

**企业名称识别流程**

1. 机构称谓词识别

输入文本为一个句子，在文本中发现指定企业后缀词时触发规则识别器。

1. 规则匹配[8]

定义规则集R（词性序列），候选规则集CR，指针\*pr和\*pw。匹配过程如下：

1. 对于 R中下一条规则r，pr指向规则r中的最后一项
2. pw指向文档中的当前词
3. 如果\*pw与\*pr词性相同，则 w--，p--；转到 ③
4. 如果pr 为null，则匹配成功，将r加入CR，转到 ①
5. 如果pr不为null，则匹配失败，转1）
6. 规则决策

在若干候选规则中选择最优解，即求出 P(ri|O)最大时的ri，O为当前观察值。

首先，对于规则集R中的每一个规则 ri, 假设它的出现次数为ni，总的出现次数为ntotal，可以计算该事件出现的频率：



其次，还要对每一事ri 计算P(O|ri)，即当事件ri发生时，组织机构名称被正确识别的概率,特征函数:





根据贝叶斯定理得



因此



上述方法是按照统计结果进行匹配，然而根据实际观察，我们发现包括企业在内的组织机构名称更倾向于匹配最长的规则，因此我们增加了一个参数param（0～1），该参数意义如下：

param=1：选择最长的候选规则作为识别结果

param=0：选择最短的候选规则作为识别结果

param=0.5：没有人工干预，纯粹按统计结果计算

0.5～1：规则越长，越大

0～0.5：规则越短，越大



1. 左边界修正

由于企业字号用字非常自由，其词性多变，因此会经常与现有规则无法匹配，导致左边界识别错误，如“上海宇青文体用品器材厂”，其词性序列为“/ns /ng /a /n /n /n /n” ，而规则集中只有“/n /n /n /n”，因此该企业可能被识别为“文体用品器材厂”。

考虑到企业多半以地名开头，因此如果匹配的规则不是以地名开头，则继续往前回溯指定距离。若找到地名，则更新识别结果，若遇到单独分词的“把”、“是”、语气词等(Stop Words)或者已达到最大长度限制，则返回原来的结果。

## 企业简称识别

本文采用根据全称推测简称的方法。企业名称最多包括行政区划（地名）、企业字号、行业、组织形式四个字段。将企业名称按下述格式进行分解，得到若干候选简称，然后扫描当前页面，就能发现是否存在匹配的简称。

1. 企业字号
2. 行政区划+企业字号
3. 企业字号+行业
4. 企业字号+组织形式

## 日期、货币和地点识别

分别将时间单位、货币单位、地点后缀词作为以上三种实体的触发关键词。

识别过程的关键步骤如下：

1. 标注单独构成实体的词；
2. 遇到关键词，则根据关键词类型（前缀或者后缀）分别往后或往前搜索，若找到符合条件的词则识别为一个实体，其中日期实体搜索的是数字词和时间词，货币实体搜索的是数字词，地点实体则搜索直到遇到地点词或Stop Words。
3. 将相邻的同一种实体合并为一个实体。

## 商标信息识别

本文采用HMM进行商品名称的识别。模型参数设置如表 1‑1所示。

表 1‑1 HMM模型参数表

|  |  |
| --- | --- |
| 参数 | 含义 |
| N | 状态数目：品牌、型号、修饰词、产品名、其他 |
| M | 词性数目，本文采用中科院计算所二级词性标注，总共96种 |
| Aij | 与时间无关的状态转移概率矩阵 |
| Bij | 每个状态下词性的分布概率 |
| πi | 句子第一个词属于某个状态的概率 |

任何词都应属于上述五种状态中的一种，其中修饰词包括颜色、大小以及其他特殊描述。常见的商品名称格式有：

1. 品牌+型号+修饰词+产品名：海尔 XQB50-918A 全自动洗衣机
2. 品牌+型号+产品名：三洋 XQB50-M805Z 洗衣机
3. 品牌+修饰词+产品名：数源25寸纯平CRT电视机
4. 品牌+修饰词+产品名+型号：索尼32英寸高清液晶电视KLV-32BX205
5. 品牌+产品名：飞利浦剃须刀
6. 品牌+产品名+型号：TCL冰箱BC-92B
7. 品牌+型号：天逸F41(GT5250-P)

可以建立有限状态机表示其状态转移过程，如图 3‑4所示：



图 1‑2 HMM状态转移拓扑

HMM在进行命名实体识别前需要进行参数训练，本文采用Baum-Welch算法进行迭代计算。该算法主要步骤如下：

1. 初始模型(待训练模型) λ0；
2. 基于λ0以及观察值序列O，训练新模型λ；
3. 如果 log P(O|λ)-log(P(O|λ0) < Delta，说明训练已经达到预期效果，算法结束。
4. 否则，令λ0＝λ，继续第2步工作；

商品识别的计算过程包括状态解码、状态识别、识别结果修正三个步骤。

1. 状态解码：根据模型和观察值序列，采用Viterbi算法计算出可能性最大的状态序列。
2. 状态识别：根据图 1‑2所示，将符合商品名称状态序列的若干个词合并成一个商品实体。
3. 识别结果修正：对于被识别为商品的实体，在本地产品名列表中查找其产品名称。这一步骤可以有效的剔除错误识别，提高识别准确率，同时还可以获取产品分类用于验证商标的国际分类号。

使用说明：

规则的建立：

如果想通过训练样本为某种实体建立规则，则需要以下步骤：

假如想建立以TAOBAO命名的规则：

1. 首先必须在配置文件XMLConfig.java中写入：

**public** **static** **final** String *RULE\_TAOBAO\_LIST*="TAOBAO";//增加淘宝的一类实体

**然后在rule.**xml中加入以*RULE\_TAOBAO\_LIST命名*的目录，并新建一条新规则。即

<TAOBAO>

<total>8276</total>

<rule>

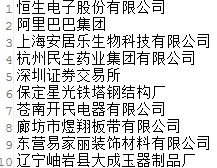
<POS></POS>

<count>0</count>

</rule>

</TAOBAO>

1. 同样在keyword.xml中需要配置上要识别该实体时，需要触发的关键词，比如识别企业名称时，如果句子中遇到”公司”时，就会触发实体识别程序，对前面的句子识别企业名称。
2. 接下来需要准备训练样本，训练样本的格式是每个实体一行，如企业名称



，然后运行训练程序，对实体的构词规律进行训练。并将训练的规则放入到rule.xml中。

1. 训练好模型之后，就可以对样本进行测试了。可以找一些实例检验模型的效果。如果一些实例没有识别，还可以把这些实例加入到样本中重新训练。

可以参考QuickStart.java

参考文献

1. 赵军. 命名实体识别、排歧和跨语言关联[J]. 中文信息学报. 2009, 23 (2): 3-17
2. Casey Whitelaw, Alex Kehlenbeck, Nemanja Petrovic, Lyle Ungar. Web Scale Named Entity Recognition [J]. Proceeding of the 17th ACM conference on information and knowledge management, 2008: 123-132
3. Jian Sun, Jianfeng Gao, Lei Zhang, Ming Zhou, Changning Huang. Chinese Named Entity Identification Using Class-based Language Model[C]. Proceeding of the 19th International Conference on Computational Linguistics. 2002: 967-973
4. 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]. 计算机工程与科学. 2006, Vol.28 (6): 135-139
5. Hai Leong Chieu, Hwee Tou Ng. Named Entity Recognition with a Maximum Entropy Approach [J]. Proceedings of the seventh conference on Natural language learning at HLT-NAACL. 2003, Vol. 4: 160 - 163
6. Youzheng Wu, Jun Zhao, Bo Xu, et a1．Chinese Named Entity Recognition Model Based on Multiple Features [C]. Human Language Technology Conference & Conference on Empirical Methods in NLP (HLT/EMNLP), Association for Computational linguistics, 2005: 427-434
7. 刘非凡，赵军，吕碧波，徐波，于浩，夏迎炬. 面向商务信息抽取的产品命名实体识别研究[J]. 中文信息学报, 2006, 20(1): 7-13
8. 沈嘉懿，李芳，徐飞玉，Hans Uszkoreit. 中文组织机构名称与简称的识别[J]. 中文信息学报. 2007, 21(6): 17-21
9. 赵琳瑛. 基于隐马尔科夫模型的中文命名实体识别研究[D]. 西安电子科技大学. 2008:1-50
10. 李荣，郑家恒，郭梅英. 基于遗传算法的隐马尔可夫模型在名词短语识别中的应用研究. 计算机科学. 2009,36(10): 244-246
11. 郑家恒，王兴义，李飞. 信息抽取模式自动生成方法的研究. 中文信息学报，2004, 18 (1): 48-54