

文章编号:1001-9081(2009)01-0143-03

## 面向短文本的命名实体识别

王丹,樊兴华

(重庆邮电大学 计算机科学与技术研究所,重庆 400065)

(wdan02@gmail.com)

**摘要:**针对短文本命名实体识别这项紧缺任务,提出了一种面向短文本的快速有效的命名实体识别方法。该方法主要分成三步:第一步,针对短文本表达不规范特性对命名实体识别的干扰,采取去干扰字符,化繁为简等规范化操作。第二步,针对短文本语意不完整特性,提出用HMM(隐马尔可夫模型)以词性做观察值进行初步命名实体识别。第三步,据初步识别结果,构建拼音同指关系库来识别潜在实体。在由8464篇短文本构成的测试集上运行的实验表明,该方法能较好地短文本命名实体识别。

**关键词:**短文本;隐马尔可夫模型;命名实体识别;拼音同指关系库;词性

**中图分类号:** TP18 **文献标志码:** A

## Named entity recognition for short text

WANG Dan, FAN Xing-hua

(Institute of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

**Abstract:** Aiming at the urgent task of named entity recognition for short text, a fast and effective method was proposed. The method comprised three steps: Firstly, according to the disturbance of non-standard expression in short text, the elimination of interferential characters and text simplification were adopted. Secondly, according to the non-integrity of short text, Hidden Markov Model (HMM) was employed to preliminarily name entity recognition, in which the part of speech was used as observed value. In the end, by means of the preliminary recognition result, a pinyin co-referential relation library was established to identify the potential entity. The experiment on the test-set including 8464 short texts shows that this method has better performance to named entity recognition for short text.

**Key words:** short text; Hidden Markov Model (HMM); named entity recognition; pinyin co-referential relation library; part of speech

### 0 引言

随着互联网和通信产业的快速发展,各种形式的信息扑面而来。BBS、聊天室、即时通信软件、手机短信等的出现无时无刻不在影响着人们的日常生活。这些形式的信息都有一些共同特点:文本比较短,用语不规范,语意不完整等,我们统称其为短文本(Short Text)。由于短文本在Internet信息流中日益突出的地位,人们迫切需要一些自动化工具帮助进行短文本海量信息处理,短文本中信息抽取、信息检索、机器翻译、文摘生成等技术正是在这种背景下产生的。在这些技术中,一个共同而基础的问题就是短文本命名实体识别(Named Entity Recognition)。短文本命名实体识别作为这些研究中非常重要并且是必不可少关键技术,越来越得到人们的重视和关注。短文本命名实体识别的质量会直接影响到后续的一系列工作,例如在信息抽取中如果没有先识别实体,根本就不可能识别实体关系<sup>[1]</sup>;在文摘生成中,很多时候是对固定模式的填充,填充内容大都是“谁”,“干什么”,“什么时候”,“在哪里”,等等,这正是命名实体的内容。因此,短文本命名实体识别已经越来越成为短文本处理中的关键技术。

目前,一般长文本命名实体识别技术已较为成熟,而短文本命名实体识别这一领域仍属一片空白。短文本命名实体识

别较于一般长文本具有以下几个难点:

1)短文本<sup>[2]</sup>表达不规范:如1)特殊符号(“:一”)表示一个笑脸);2)繁简夹杂,中英文夹杂(“在日本並沒有什麼影響~~~やれやれ”)。这些表达的不规范,极大影响了命名实体识别。

2)语意不完整:如“我中奖了”会写成“我中了”,“好多了”写成“好大发了”。

3)同音字现象,如:“周恭伟”会写成“州工尾”、“舟公委”等,这样会导致大量同音错别字实体无法识别出。

因此本文针对短文本这些特殊特性<sup>[3-4]</sup>,提出一种面向短文本命名实体识别方法,其基本思路是:1)规范化短文本;2)利用隐马尔可夫模型(Hidden Markov Models, HMM)<sup>[5]</sup>以词性作观察值进行初步命名实体识别;3)根据初步命名实体识别结果,构建实体拼音同指关系库,用来识别潜在同音字实体。

### 1 面向短文本的命名实体识别方法

#### 1.1 短文本规范化

短文本规范化具体来说:

1)清除所有的干扰字符。因为这些字符对于短文本来说没有什么实际意义,而且还会造成对命名实体识别的干扰。

收稿日期:2008-07-14;修回日期:2008-09-16。 基金项目:国家自然科学基金资助项目(60703010);重庆市自然科学基金资助项目(2006BB2374);重庆市教委科学技术研究项目(KJ070519);教育部回国留学人员启动基金资助项目(教外司留[2007]1109号)。

作者简介:王丹(1982-),女,湖北鄂州人,硕士研究生,主要研究方向:中文信息处理、机器学习;樊兴华(1972-),男,重庆人,教授,博士,主要研究方向:人工智能、自然语言处理、信息检索。

这里干扰字符包括诸如标点、特殊字符等。例如：“OK `` 呵呵 `` 他是真的”，将所有的非法字符去掉，得到“OK 呵呵他是真的”。

2) 将所有繁简夹杂，中外文夹杂的短文本全转换为中文简体。由于短文本中大量出现繁简夹杂，中外文夹杂的现象，这些繁体，外文里包含的实体识别不出，而且还会影响识别其他实体。因此文中将它们统一转化为中文简体。比如“这倒可以从侧面看出，柏芝这个在中国的天后级人物，在日本并没有什么影响良好的悲痛”。

### 1.2 HMM 模型以词性做观察值训练参数

目前，针对一般长文本的命名实体识别已有大量方法，其主要方法有基于规则的方法与基于统计的方法。因 HMM 能较好地捕获所需要的状态转移信息，而且由于经典的 Viterbi 算法在求取最佳状态序列的高效性，使得 HMM 在该领域中的应用很广泛。

HMM<sup>[6]</sup> 可以表示为一个五元组  $\{S, O, P, A, B\}$ ，其中：

$S = \{S_1, \dots, S_n\}$  表示状态的集合；

$O = \{O_1, \dots, O_m\}$  表示观察值的集合；

$P = \{P_i\}$  表示状态的初始概率；

$A = \{a_{ij}\}$  表示从状态  $S_i$  到状态  $S_j$  的转移概率矩阵；

$B = \{b_{jk}\}$  表示从状态  $S_j$  观察到  $O_k$  的发射概率矩阵。

对于某个给定的词性序列  $W = w_1, w_2, \dots, w_m$ ，NE 识别的目的是要找到一个最优的 NE 标注序列  $T = t_1, t_2, \dots, t_m$  使得条件概率  $P(T|W)$  达到最大。由贝叶斯公式可得：

$$P(T|W) = \frac{P(W, T)}{P(W)} = \frac{P(T)P(W|T)}{P(W)} \quad (1)$$

假设转移概率只与前一个状态有关，输出观察值的概率只与当前状态有关，则有：

$$P(T) = \prod_{i=1}^m P(t_i | t_{i-1}) \quad (2)$$

$$P(W|T) = \prod_{i=1}^m P(w_i | t_i) \quad (3)$$

其中： $P(t_i | t_{i-1})$  表示状态  $t_{i-1}$  到状态  $t_i$  的转移概率。 $P(w_i | t_i)$  表示在状态  $t_i$  出现的条件下观察到  $w_i$  的发射概率。而对于一个给定的词性序列  $W$  来说， $P(W)$  的值是确定的，可以不考虑。于是最终输出  $T^*$  可以表示为：

$$T^* = \arg \max_T \left[ \sum_{i=1}^m (\log P(t_i | t_{i-1}) + \log P(w_i | t_i)) \right] \quad (4)$$

HMM 模型的命名实体识别过程就是利用训练参数为当前输入的词性序列标注最优的状态序列的过程。即在给定模型  $\lambda = \{A, B, P\}$  和观察值序列  $W$  的条件下找出概率最大的状态序列：

$$Q^* = \arg \max_Q P(Q|W, \lambda) \quad (5)$$

本文采用动态规划的韦特比算法求解最佳的状态序列，归纳后有：

$$\delta_{s+1}(j) = (\max_i \delta_s(i) a_{ij}) \times b_j(w_{s+1}) \quad (6)$$

本文针对短文本语意不完整的特性，提出以词性<sup>[7]</sup>做观察值进行命名实体识别。文中所用词性共 48 种（分别是 "Ag" "a" "ad" "an" "Bg" "b" "c" "Dg" "d" "e" "f" "g" "h" "i" "j" "k" "l" "Mg" "m" "Ng" "n" "nr" "ns" "nt" "nx" "nz" "o" "p" "Qg" "q" "Rg" "r" "s" "Tg" "t" "Ug" "u" "Vg" "v" "vd" "vn" "w" "x" "Yg" "y" "z" "c"

"zzz")。识别的 NE 类别共三类，包括人名 (nr)、地名 (ns) 和组织名 (nt)。每一种类别根据它的组成部分在 NE 中出现位置的不同又可以分为 NE 开头 (B-NE)，NE 内部 (I-NE)，NE 结尾 (E-NE) 以及独立 NE (S-NE)，再加上不属于任何 NE 类别的其他类型 (o)，共 13 种，这 13 种即为 13 种状态。

### 1.3 构建拼音同指关系库识别潜在同音字实体

经过初步识别得出的结果，由于短文本，主要由于拼音输入法的使用，导致大量出现同音错别字实体，这些同音错别字实体经初步识别，未能识别出来。因此本文提出据初步识别结果中识别出的实体，去重及人工筛选后，构建实体拼音同指关系库，用来识别潜在同音字实体。比如文中出现的“周恭伟”经初步识别能识别出来，但是“周恭伟”由于拼音输入法的使用，可能被误写成“州工尾”、“舟公委”，这些同音错别字实体就未能识别出。据本文方法，先将初步识别结果全转化为汉字拼音声调形式，如初步识别结果中的“州/n+o 工/n+o 尾/ng+o”转化为汉字拼音声调形式就为“州(zhou1)/n+o 工(gong1)/n+o 尾(wei3)/ng+o”。然后据初步识别结果中实体“周恭伟”构建的拼音同指关系库里规则“周(zhou1)恭(gong1)伟(wei3)”，根据拼音匹配，这两个词具有同指关系，因此就将“州工尾”是人名识别出来。

### 1.4 面向短文本的命名实体识别方法

流程如图 1 所示。

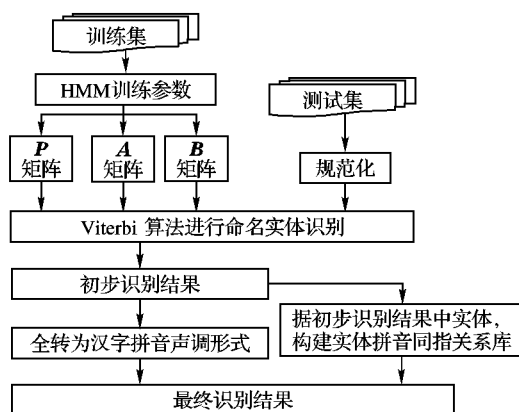


图 1 面向短文本的命名实体识别方法流程

该方法的基本过程为：

第 1 步 导入测试集，对短文本测试集规范化。

第 2 步 导入训练集，用 HMM 模型以词性做观察值训练参数，得到  $P$  矩阵（初始状态矩阵）， $A$  矩阵（状态转移矩阵）和  $B$  矩阵（观察值矩阵）。

第 3 步 对于规范后的短文本测试集与  $P$  矩阵， $A$  矩阵， $B$  矩阵，通过 Viterbi 算法进行命名实体识别，得到初步识别结果。

第 4 步 提取初步识别结果中实体，去重及人工筛选后构建实体拼音同指关系库，同时将初步识别结果全转化为汉字拼音声调形式。然后根据实体拼音同指关系库里规则进行匹配。若匹配则修正标注，从而得到最终识别结果。

## 2 实验结果及分析

### 2.1 实验设置

#### 2.1.1 实验语料

测试集原始语料：全部收集于新浪和网易的各大新闻标题或专题的网友评论，过滤掉文本长度大于 160 个字的长文

本,保留 4 类共计 8464 篇评论,其中科技 2028 篇、娱乐 2213 篇,书评 2013 篇,女性 2210 篇。

测试集命名实体正确标注语料的构建:采用中科院张华平博士的 ICTCLAS 1.0 系统先自动标注,标注后人名、地名的精确率达到 90% 以上,组织名稍低。然后再人工校对,得到一个较正确的命名实体标注集,作为评判标准。

训练集语料采用了 2 个:

1) 北京大学收集的《人民日报》语料库中 199801 月份语料。

2) 采集于网易的科技专题的网友评论 2000 篇原始语料,经过 ICTCLAS 1.0 系统自动标注后,再人工校对,得到一个较正确的 2000 篇短文本的训练集。

### 2.1.2 实验工具和参数设置

HMM 模型以词性做观察值,采用了共 48 种词性("Ag" "a" "ad" "an" "Bg" "b" "c" "Dg" "d" "e" "f" "g" "h" "i" "j" "k" "l" "Mg" "m" "Ng" "n" "nr" "ns" "nt" "nx" "nz" "o" "p" "Qg" "q" "Rg" "r" "s" "Tg" "t" "Ug" "u" "Vg" "v" "vd" "vn" "w" "x" "Yg" "y" "z" "c" "zzz"),13 种状态("Snr" "Bnr" "Inr" "Enr" "Sns" "Bns" "Ins" "Ens" "Snt" "Bnt" "Int" "Ent" "o")。

汉字拼音声调转换采用的是 KTestpinyin 4.7 版工具。

外文转换成中文采用的是 google 翻译工具。

### 2.1.3 评价标准

命名实体(NE)评测指标选用准确率(P)、召回率(R)和 F 值。具体定义如下:

$$P = \frac{\text{系统标注正确的 NE 总数}}{\text{系统标出的 NE 总数}} \quad (7)$$

$$R = \frac{\text{系统标注正确的 NE 总数}}{\text{测试集中出现的 NE 总数}} \quad (8)$$

$$F = \frac{2PR}{P + R} \quad (9)$$

## 2.2 试验结果

### 实验 1 验证该方法的有效性

短文本测试集规范化后,以人民日报 199801 月份语料做训练集,用 HMM 模型以词性做观察值得到训练参数  $P$  矩阵,  $A$  矩阵和  $B$  矩阵,然后以 Viterbi 算法进行命名实体识别,得到初步识别结果。在初步识别结果基础上,利用构建的拼音同指关系库识别出潜在实体,得到最终识别结果(见表 1)。

表 1 验证本文方法的有效性 %

NE 类型	模型	精确率(P)	召回率(R)	F 值
人名	初步识别结果	89.37	74.05	80.99
人名	最终识别结果	88.51	89.01	88.76
地名	初步识别结果	86.93	92.97	89.84
地名	最终识别结果	85.87	97.12	91.15
组织名	初步识别结果	45.28	77.41	57.14
组织名	最终识别结果	47.27	83.87	60.46

由表 1 实验结果可以看出,初步识别结果达到了 HMM 方法用于识别一般长文本中命名实体相当的水平。说明规范化并结合 HMM 以词性做观察值可以有效地识别短文本中命名实体。同时最终识别结果中提出的拼音同指法,从试验结果上看在保持精确率相当的情况下,识别出了短文本中特殊的大量潜在的命名实体,召回率,总的 F 值都有大幅提高。

### 实验 2 验证规范化短文本的影响。

以人民日报 199801 月份语料做训练集。同时将收集于新浪和网易的各大新闻标题或专题的网友评论的原始语料共 8464 篇评论,其中科技 2028 篇、娱乐 2213 篇,书评 2013 篇,女性 2210 篇。一份对其规范化处理后,作为测试语料;一份不做任何处理,直接作为测试语料。然后分别用 HMM 模型进行命名实体识别。

实验结果如表 2 所示。从表 2 可以看出,规范化后精确率,召回率, F 值都有较大提高,说明短文本的规范化能有效提高命名实体识别性能。

表 2 验证规范化短文本的影响 %

NE 类型	是否规范化	精确率(P)	召回率(R)	F 值
人名	未规范化	66.08	64.01	65.03
人名	规范化	89.37	74.05	80.99
地名	未规范化	84.23	90.01	87.02
地名	规范化	86.93	92.97	89.84
组织名	未规范化	42.30	70.96	53.01
组织名	规范化	45.28	77.41	57.14

实验 3 验证以现有一般长文本语料标注集做训练语料的可行性。

目前由于没有现有一般长文本命名实体标注集,经过分析,短文本在规范化后与一般长文本在语法形态上基本相似。因此文中分别用人民日报标注集 199801 月份前 2000 篇文本作训练集和以人工精确标注短文本科技类语料 2000 篇做训练集,然后用 HMM 模型进行命名实体识别,实验结果如表 3 所示。从表 3 可以看出:人名、地名的精确率、召回率, F 值均相差在 5 个百分点左右,组织名的识别结果差异稍大,这是由于短文本中组织名出现较少,导致偏差较大的缘故。由此说明在无现有一般长文本命名实体标注集情况下,可以采用已有的一般长文本语料标注集作训练语料来初步标注短文本。

表 3 实验结果 %

NE 类型	训练语料	精确率(P)	召回率(R)	F 值
人名	人民日报	89.03	73.01	80.22
人名	人工标注短文本	83.72	68.18	75.15
地名	人民日报	89.90	87.46	88.66
地名	人工标注短文本	85.72	93.53	89.45
组织名	人民日报	39.39	41.93	40.62
组织名	人工标注短文本	25.80	25.80	25.80

## 3 结语

有关短文本命名实体识别技术是现今大量出现短文本处理的重要基础技术,本文针对当前这项急需解决的现实任务,提出了一种面向短文本的命名实体识别方法,该方法采用 3 步:1)规范化短文本。2)利用 HMM 以词性作观察值进行初步命名实体识别。3)根据初步命名实体识别结果中实体,构建拼音同指关系库,用来识别出潜在的命名实体。在由 8464 篇短文本构成的测试集上运行的实验表明:本文方法能有效进行短文本命名实体识别。同时由本文的实验还能得到如下结论:1)短文本的规范化能有效提高命名实体识别性能。2)在无现有一般长文本命名实体标注集情况下,可以采用已有的一般长文本语料标注集做训练语料来初步标注短文本。

(下转第 171 页)

功能是加密 Irp 指定缓冲区中内容,并传向下层。处理过程如下:

```
//取得文件名
fn = GetFileName( Irp );
//取得文件的加密密钥,解密后得到文件密钥
EncryptedFEK = GetEncryptedFEK( fn, UserID );
UserPriKey = GetUserPriKey( UserID );
FEK = GetFEK( EncryptedFEK, UserPriKey );
//加密文件内容,并传向下层
EncryptFileIrp( Irp, FEK );
```

#### 2.7.4 设置共享文件

设置共享文件主要函数是 SetShareFile,参数为文件名 FileName 和共享用户名 ShareUserID,其功能是把共享用户信息添加到密钥文件中。处理过程如下:

```
//取得文件的加密密钥,解密后得到文件密钥
EncryptedFEK = GetEncryptedFEK( FileName, UserID );
UserPriKey = GetUserPriKey( UserID );
FEK = GetFEK( EncryptedFEK, UserPriKey );
//取得共享用户的公钥
ShareUserPubKey = GetShareUserPubKey ( ShareUserID );
//加密文件密钥
EncryptedFEK = CreateEncryptedFEK( FEK, ShareUserPubKey );
//取得密钥文件
FEKFile = GetFEKFile( FileName );
//把加密密钥保存于密钥文件中
AddFEKFile( FEKFile, EncryptedFEK, UserID, ShareUserID );
```

#### 2.7.5 读取共享文件

读取共享文件主要函数是 ReadShareFile,参数为文件名 FileName 和共享用户 ShareUserID,其功能是把指定共享密文文件和加密密钥传给共享者。处理过程如下:

```
//检查指定用户是否有权限读取指定文件
if ( CheckShareUser( FileName, ShareUserID ) )
{ //有读取权限
//把该用户指定文件加密密钥取出,放入一临时密钥文件中
TempFEKFile = CreateTempFEKFile ( FileName, ShareUserID );
//把密文文件和临时密钥文件传送给共享者
SendShareFile( FileName, TempFEKFile );}
```

### 3 文件保护方案分析

1) 可操作性。基于文件系统过滤驱动的文件保护方案与 Windows 文件系统无缝集成,应用程序访问方式和用户操作习惯无需作任何改变,复杂的密钥管理由过滤驱动程序自动完成,用户唯一要做的是保护好用户私钥 U 盘。文件内容加解密采用快速的对称加密算法,效率可以接受。

2) 安全性。文件解密依赖于用户私钥 U 盘,用户取走私钥 U 盘后,入侵者不能得到安全目录中文件的明文。即使攻

破一个文件,其他文件仍是安全的。用户私钥已经加密,即使 U 盘丢失也不易泄露私钥;而且可以用备份私钥解密密钥文件中加密密钥,用另外新的公钥重新加密,并生成新的用户私钥 U 盘。

3) 共享安全。共享者从共享目录中得到的是密文文件和加密密钥,传输过程安全;通过密钥文件的控制,共享者只能访问指定的共享文件,不能访问非共享文件。共享者 UID 保存于共享目录所有者的计算机中,不能伪造共享者。

4) 可扩展性。加密算法和密钥长度都可以由用户指定,私钥 U 盘可方便地用智能卡替换。

5) 成本低廉。在具有较高安全性的同时,硬件开销是目前非常廉价的 U 盘,具有很强的实用性。

### 4 结语

一些研究者已证明基于文件系统过滤驱动的加解密处理能有效保证 Windows 文件的安全<sup>[1-3]</sup>,本文在此基础上采用一文件一密钥方式及双密钥体系进一步增强了文件系统的安全性,利用密钥文件保存文件加密密钥,通过有效的密钥管理,在提高安全性的同时不增加安全成本,具有较好的实用性。作者以此方案开发的加密目录系统经试用系统稳定、安全性好,获得用户良好的评价。本方案将在以下方面进一步改进:1) 把私钥 U 盘替换为专门的智能卡;2) 密钥文件中加密密钥信息更丰富一些,并研究如何高效地访问密钥文件;3) 把文件加密与病毒防治结合起来。

#### 参考文献:

- [1] 王雷,庄毅,潘龙平. 基于强制访问控制的文件安全监控系统的设计与实现[J]. 计算机应用, 2006, 26(12): 2941 - 2944.
- [2] 翟进,李清宝,白燕,等. 文件过滤驱动在网络安全终端中的应用[J]. 计算机应用, 2007, 27(3): 624 - 626.
- [3] 郑磊,马兆丰,顾明. 基于文件系统过滤驱动的安全增强型加密系统技术研究[J]. 小型微型计算机系统, 2007, 28(7): 1181 - 1184.
- [4] 魏不会,卿斯汉,刘海峰. 基于安全操作系统的透明加密文件系统的设计[J]. 计算机科学, 2003, 30(7): 132 - 135.
- [5] 林昊翔,董渊,张为,等. 基于 Linux 的通用加密文件系统 Waycryptic 的设计与实现[J]. 小型微型计算机系统, 2007, 28(1): 122 - 126.
- [6] 李凡,刘学照,卢安,等. Windows NT 内核下文件系统过滤驱动程序开发[J]. 华中科技大学学报:自然科学版, 2003, 31(1): 19 - 21.
- [7] NAGAR R. Windows NT file system internals [M]. New York: O'Reilly & Associates, 1997.

(上接第 145 页)

#### 参考文献:

- [1] FAN X F, SUN M S. Knowledge representation and reasoning based on entity and relation propagation diagram / tree [J]. Intelligent Data Analysis, 2006, 10(1): 81 - 102.
- [2] 黄永光,刘挺,车万翔,等. 面向变异短文本的快速聚类算法[J]. 中文信息学报, 2007, 3(21): 63 - 68.
- [3] 樊兴华,王鹏. 基于两步策略的中文短文本分类研究[J]. 大连海事大学学报, 2008, 34(3): 121 - 124.
- [4] 宋东风,张志浩. 短文本数据的自动分类[J]. 电脑与信息技术, 2007, 2(15): 36 - 39.
- [5] ZHOU G D, SU J. Named entity recognition using an HMM-based chunk tagger [C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: [s. n.], 2002: 473 - 480.
- [6] 廖先桃,于海滨,秦兵. HMM 与自动规则提取相结合的中文命名实体识别[D]. 哈尔滨: 哈尔滨工业大学信息检索研究室, 2004.
- [7] 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, 28(6): 135 - 139.