基于小规模尾字特征的中文命名实体识别研究

冯元勇[1,2,3],孙 乐[1],张大鲲[1,3],李文波[1,3]
(1. 中国科学院软件研究所基础软件工程研究中心,北京 100080;
2. 广州大学计算机学院,广东广州 510006; 3. 中国科学院研究生院,北京 100049)

摘 要: 本文针对难度最大的两类命名实体(地名和机构名)在条件随机场框架下首次引入了小规模的常用尾字特征. 实验表明,该特征与词类特征具有一定的互补性,联合使用可以以较小的训练代价显著提高专有名词的识别性能,特别是机构名的识别精度. 该系统在我国 863 简体命名实体识别评测语料上专名(人名、地名和机构名)总体 F1 值达 88.76 %,超过当年最佳系统 8.63 个百分点. 在 SIGHAN 2006 命名实体识别语料上的结果也居于前列.

关键词: 中文命名实体识别; 小规模尾字特征; 条件随机场; 自然语言处理; 机器学习
中图分类号: TP391.1 **文献标识码**: A **文章编号**: 0372-2112 (2008) 09-1833-06

Study on the Chinese Named Entity Recognition Using Small Scale Character Tail Hints

FENG Yuan-yong[1,2,3], SUN Le[1], ZHANG Da-kun[1,3], LI Wen-bo[1,3]
(1. *Institute of Software, Chinese Academy of Sciences, Beijing* 100080, *China*;
2. *Guangzhou University, Guangzhou, Guangdong*510006, *China*;
3. *Graduate University of Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract: We propose small-scale-hint-character-list (SSHCL) features for location and organization names under the conditional random fields framework. As experiments show, SSHCL features provide significant gains in precision, especially for organization names, showing complementary property to part-of-speech. It also lowers construction and training cost greatly that a common large scale feature set demands. The overall proper nouns F1 measurement of integrated system on simple Chinese 863 program 2004 NER corpora reaches 88.76 %, gaining 8.63 % improvement over the best system in the evaluation. The performance on SIGHAN 2006 is also remarkable.

Key words: Chinese named entity recognition; small-scale-tail-hint-character-list feature; conditional random fields; natural language processing; machine learning

1 引言

命名实体识别是计算机理解文本信息的基础. 命名实体(Named Entity)指那些能够明确指称外部世界某一对象的名词或名词短语. 命名实体识别(Named Entity Recognition, NER)就是确定文档中的人名、地名和机构名等文本片段并识别其类型的过程. 它是信息抽取、问答系统、机器翻译、文档摘要、跨语言检索等自然语言处理研究的关键技术之一.

在常见的五类命名实体中,数量和时间表达式相对比较容易,几乎完全可以依靠几种模式匹配完成[1]. 而人名、地名和机构名则相对复杂. 特别是中文机构名和地名识别,由于跨度大,内部结构复杂,成为命名实体识别的难点. 因此本文主要研究这三类命名实体(称为专有名词, Proper Nouns)的识别.

条件随机场(Conditional Random Fields, CRF)[2]模型为中文命名实体识别提供了一个多特征融合框架. 在这个框架下,多层次的、长距离的特征可以方便地引入,并且不要求特征间的相互独立. 在 NER 这样的序列文本标注任务上,大量实验证实 CRF 模型综合了最大熵(MaxEnt)和隐马尔科夫模型(HMM)[1,3]的优点[4,5],在多方面优于其它模型,因而近年来在词类标注[2]、中文分词[6]、命名实体识别[4,7,8]和蛋白质名称识别[5]等研究中广为应用.

收稿日期:2007-03-12;修回日期:2007-12-06
基金项目:国家自然科学基金(No. 60773027, 60736044);863 重点项目(No. 2006AA010108);国家 242 项目计划(No. 2006A40)

但是 CRF 模型同时也存在着训练时间长、收敛慢的缺点. 由于其标准线性链(linear-chain)模型[2]单轮训练迭代时间复杂度与特征规模成正比,我们设计了一个小规模的用字列表特征来直接降低特征集的规模,达到提高算法运行速度的目的.

本文布局如下:第二节介绍国内外在基于字词特征选取方面的相关工作,以及我们提出小规模提示用字,特别是地名和机构名尾字特征的动机. 第三节简要介绍融合多特征的条件随机场模型框架下的命名实体识别原理,分析其训练复杂度,指出特征规模与复杂度之间的关系. 然后在第四节结合中文命名实体的特点,给出我们的小规模常用提示字特征融合到 CRF 下进行中文命名实体识别的基本原理. 第五节给出了实验数据,并进行了结果分析. 最后为全文总结,提出了将来工作的方向.

2 相关工作

目前,中文命名实体识别大多建立在分词的基础上. 这些系统的特征大多集中在词文本、词类、常用词和名称列表、字号列表[3,9,10],其中常用词的研究最多. 据我们掌握的文献来看,除人名外*,还没有直接以提示字特别是尾字为特征的中文 NER 系统. 我们认为,这样的特征系统存在着两个方面的局限. 首先,由于命名实体多为新词,分词预处理极易在边界处错误切分. 当系统模型特征以词为基本粒度时,这些错误难以在模型内纠正,形成错误积累. 其次,尽管词和名称列表具有一定的稳定性,但过长的词串容易造成数据稀疏,影响系统的召回率.

通过对地名和机构名常见尾字的分析,我们认为这两类命名实体的尾字同样具有强烈的提示作用. 在大多的情况下,同一个尾字可以覆盖多个尾词(地名和机构名提示词一般集中在后部),这是由于汉语组词时,尾字富含区别性意义的缘故. 例如"区"可覆盖"市区"、"郊区"、"自治区"、"地区"等提示词,均表示地理区域范围;"局"可覆盖"市局"、"公安局"、"邮局"、"卫生局"等,均表示特定级别的行政部门. 因此尾部提示字比提示词具有更好的覆盖能力,可缓解数据稀疏的问题,与字号、名称列表相比较,这种覆盖更加具有通用性.

据我们所知,本文是第一次将尾字特征用于中文地名和机构名命名实体识别. 由于单字特征的规模要小于词特征的规模,更小于大规模特征集[4,9],因此可以提高 CRF 模型训练速度. 此外,单字特征引入可以减少分词错误的影响,在目前分词效果不理想的情况下是有意义的.

3 条件随机场框架下的命名实体识别

3.1 特征表示

命名实体识别一般不处理嵌套,因此相应的 CRF 模型可简化为线性链(linear-chain)模型. 这种情况下,每个字为一个符号(token). 符号用它各个特征的特征值表示. 特征表达了对字或其环境的属性的判断. 这里,我们用到了三类特征,分别为字文本本身、分词后包含该字的词的词类和用字列表特征.

当符号的特征值确定以后,符号就用它各个特征的特征值集合(特征向量)表示,称为观察. 句中所有由特征值表示的符号形成一个序列,称为观察序列. 例如句子"张掖市民陈述军认为……"可表示为:

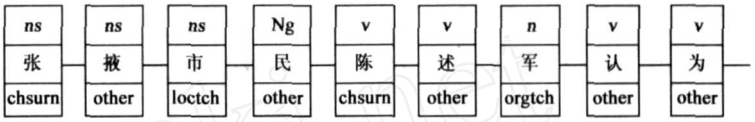


图 1 观察序列

其中,上框为符号的词类特征,中框为文本特征,下框为用字列表特征.

到目前为止,特征实际为观察特征,即输入序列在某个点上的属性或其组合. 后面我们将会看到,CRF 模型建立在观察特征与状态联合之上,我们用特征函数来刻画这种联合. 特征函数是对特定情况的判断,大多情况下取值为 0,只有判断成立时为 1. 例如"姓氏特征函数"定义为:

$$\mathrm{surn}(s,o)=\begin{cases}1, (s==\mathrm{B\text{-}PER}) \wedge \\ \quad (\mathrm{FirstChar}(o)\in \mathrm{List}(\mathrm{chsurn}))\\ 0, \mathrm{otherwise}\end{cases} \tag{1}$$

表示对状态值 s 与观察 o 的判断,其中"B-PER"为表示人名起始的标签(B 表示起始,PER 表示人名),FirstChar 为首字,List(chsurn)为中文姓氏列表中的所有条目.

3.2 状态表示

命名实体识别是对命名实体的边界确定和类型确定. 在 CRF 模型下,这两个问题可以综合考虑. 本文中,每个字符号对应着一个状态有待猜测的节点,该状态取值为这个符号的标签. 因此状态变量包含了两个方面的内容,即字符号所在命名实体的类型和表征符号在命名实体中位置的边界信息,对应的标签格式为"边界标签-类型标签". 我国 2004 年度 863 命名实体识别[11]任务定义有六类命名实体,分别是人名(PER)、地名(LOC)、机构名(ORG)、日期表达式(DAT)、时间表达式(TIM)、数量表达式(NUM),各对应一种类型标签,加上非命名实体标签(OTH),共七种类型标签. SIGHAN MSRA 任务只识别前三类命名实体(专有名词),故有四种类型标签. 边界标签则有命名实体起始(B)、命名实体接续(I)和非命名实体(O)三种. 外部边界标签 O 与 OTH 类对应,故 863 任务共有 $6\times2+1=13$ 种状态标签,MSRA 任务共有

* 尽管这些系统运用了单字姓氏和人名用字列表,但它们本质上是直接运用分词预处理的结果,为人名单字词.

7 种状态标签.

3.3 状态序列的概率估计

对命名实体识别而言,如果直接基于整个观察序列进行特征表示与状态序列估计,将会因为稀疏问题而不可计算. 因而,我们假设状态序列为满足一阶马尔科夫过程约束的线性链(linear-chain)模型,从而将基于观察序列的全局特征用待猜测节点的局部特征来表示. 对于由可观察符号特征组成的观察序列 $o = < o_1, o_2, \ldots, o_T >$,其对应状态序列 $s = < s_1, s_2, \ldots, s_T >$(在该模型中,待猜测节点与可观察字符号一一对应)概率的估计方法如下[2]:

$$p_\Lambda(s \mid o) = \frac{1}{Z_\Lambda(o)} e^{\sum_{t=1}^{T} \phi(s, o, t)} \tag{2}$$

其中,Λ 为模型的参数集,$\phi(s, o, t)$ 为序列第 t 个基团(clique)的势(potential);在标准的线性链模型中,基团由单个节点组成,势函数的具体定义将在下节给出,T 为序列的长度,$Z_\Lambda(o)$ 为由所有可能状态序列构成的配分函数(partition function):

$$Z_\Lambda(o) = \sum_{s} e^{\sum_{t=1}^{T} \phi(s, o, t)} \tag{3}$$

与 HMM 类似,可以定义 CRF 下的 Baum-Welch 算法. 例如前向变量 $\alpha_t(s_i)$ 为:

$$\alpha_{t+1}(s_i \mid o) = \sum_{s'} \alpha_t(s' \mid o)\, e^{\phi(s, o, t)} \tag{4}$$

标准线性链条件随机场模型的解码求得概率最高的状态序列,其过程与 HMM 中的 Viterbi 类似. 基于最大似然函数的学习过程也建立在序列估计基础之上. 它们的计算复杂度均与特征空间的规模和类别标签数的平方成正比,为 $O(L^2 FTN)$,L 为类别标签数,F 为特征规模,T 为样本序列平均长度,N 为训练样本数. 由于每轮训练主要为一个似然函数值计算过程,因此,单轮训练的计算复杂度与特征规模成正比,降低特征规模可以减少单轮训练的时间,也有利于加快收敛的速度,从而降低整个训练过程的代价.

4 基于小规模尾字的中文命名实体识别

在我们的模型中,基团的局部特征由状态转移特征和观察-状态特征两部分组成:

$$\phi(s, o, t) = \sum_k \lambda_k f_k(s_{t-1}, s_t) + \sum_k \mu_k g_k(s_{t-1}, o, t) \tag{5}$$

其中,s_t 为 t 时刻的状态值,f、g 分别为状态转移特征和局部观察特征,k 为索引,λ、μ 为权重. 局部观察特征一般取自邻近窗口,我们取左右各两个符号. 通过对观察特征的细分,减小特征空间的规模,减少势函数的计算量,达到缩短模型单轮学习时间的目的.

4.1 小规模尾字特征

为了减少词语切分错误会积累到命名实体识别中的影响,本文以字为模型的基本符号单元,并收集了 100 个地名尾字和 40 个机构名尾字,作为对常用类型提示词的替代,使得系统既能捕捉类型字对命名实体的提示信息,又能避开在词表查找匹配中引入的切分错误. 这些尾字与人名单字姓氏以及各种中文数字、字母等构成了一个用字列表特征(LIST,共计条目近 600 条. 为行文方便,有时不区分尾字特征与常用字特征). 该特征部分取值见表 1.

表 1　LIST 常用字特征

值	说明	样例
digit	数字	1,2,3
letter	字母	A, B, C, ..., a, b, c
chseq	中文序数	〇, ⑴, ①, Ⅰ
chdigit	中文基数	1,壹,一
tianseq	天干地支	甲,乙,丙,丁
chsurn	姓氏	李,吴,郑,王
notname	非人名用字	将,对,那,的,是,说
loctch	地名尾字	区,国,岛,堡,冲,庄
orgtch	机构名尾字	府,团,校,协,局,办

4.2 基于尾字特征的中文地名和机构名识别

局部特征主要为尾字和邻近符号的词类等. 在基于单字符号的模型中,为消解普通词内字与命名实体内部用字的歧义,我们在预处理中进行了分词,在命名实体识别中引入了字在切分词后所得的信息,包括所在词的词类、在词中的位置等信息. 这样使得在突出尾字特征的作用基础上,既充分利用分词结果,又不易受到分词错误的影响.

譬如,"局"字是一个常见的机构名尾字,但也经常出现在"局部"、"局域"、"局限"、"局促"等普通词中. 下句中

例 1　帮助 电力局 部分 职工 ……

"局"为词尾,同时也是小规模用字中的机构名常用尾字,因此命名实体识别倾向于将其与"电力局"作为机构名的类型词,而非与已成词的"部"字结合成普通词. 所用势函数形式为:

$$\begin{aligned}\phi(\text{I-ORG} \mid o, t) = & \sum_k \lambda_k f_k(s_{t-1}, \text{I-ORG}) \\ & + \sum_k \sum_{i=-2}^{2} \mu_{k,i} g_{k,i}(\text{I-ORG}, \text{pos}(o, t-i)) \\ & + \sum_k \theta_{k,i} h_{k,i}(\text{I-ORG}, \text{OrgtChh}(o, t)) \\ & + \sum_k \nu_k b_k(\text{I-ORG}, \text{wb}(o, t))\end{aligned} \tag{6}$$

其中,s_{t-1} 为前一状态取值,$\text{pos}(o, t-i)$ 为邻近符号所在词的词类,$\text{wb}(o, t)$ 表示该符号在所属词的位置(起始/结束/中间/单字词). $\text{OrgtCh}(o, t)$ 表示 o 中第 t 个符号是否为常用见机构名尾字,同样,f、g、h、b 为特征函数,λ、μ、θ、ν 为特征的权.

5 实验数据

我们所用的训练语料由 SIGHAN Bakeoff 3 (2006)

MSRA NER**任务提供,测试语料则还包括了 863 计划 2004 年命名实体识别评测简体语料***,见表 2.

表 2 语料概况

出现次数		人名	地名	机构名	专名总计
SIGHAN MSRA	训练集	17615	36860	20581	75056
	测试集	1973	2886	1331	6190
863 简体		6850	4017	8421	19288

由于识别难度大,我们将主要关注地名(LOC)和机构名(ORG)以及另外包含人名的专名总体(Proper Nouns)识别效果****.

5.1 基准系统

我们以字为唯一特征,建立了第一个基准系统(TXT,baseline1).此外,我们还将字与词类特征组合(含词边界),建立第二个基准系统(TXT + POS,baseline2).这里的词类指在分词结果中该字所在词的词类.它们在 SIGHAN MSRA 命名实体识别语料上的开放测试结果见表 3.

表 3 各基准实验开放测试主要指标

	LOC			ORG			Proper Nouns		
	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall
TXT	83.98	84.70	83.26	68.68	63.85	74.31	77.89	78.53	77.27
TXT + POS	86.72	88.81	84.72	76.00	73.32	78.89	83.49	84.68	82.33

5.2 小规模尾字特征

下面给出在 MSRA 语料上引入小规模用字提示特征后对各类型命名实体的识别效果.

表 4 引入小规模常用提示字特征后的主要测试指标

	LOC			ORG			Proper Nouns		
	F1	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall
TXT + LIST	85.19	85.25	85.14	74.99	73.35	76.71	80.65	81.80	79.53
TXT + POS + LIST	90.54	93.21	88.01	83.38	84.44	82.34	86.11	89.38	83.07

它们与基准系统的对比示于图 2.

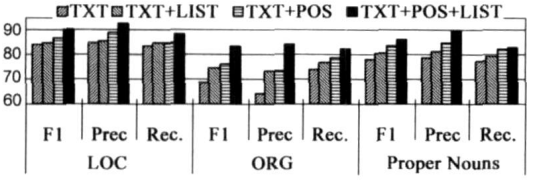


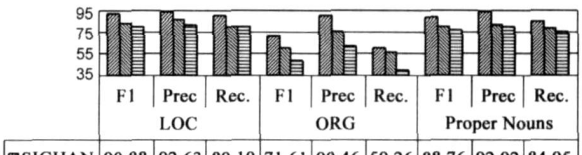
图 2 小规模常用提示字特征引入前后对比

可以看出,常用提示字特征的引入,使得字特征集(baseline1)的召回率和精确率均有提高,总体 F1 值提高约 3 个百分点.特别地,它对机构名的精度提高非常明显,幅度达 6 个百分点,接近于词类特征.这表明,小规模尾字特征在局部范围内突出了某些字文本对专有名词命名实体的提示,它对企业字号与名称列表的替代没有带来性能上的下降.

特别地,当我们将小规模常用提示字特征与词类相结合时,词类特征集(baseline2)的召回率和精确率均有较大幅度的提高,充分表现出小规模常用提示字特征对词类特征具有一定的互补性.从而使得词类特征集总体 F1 值提高约 3 个百分点,同样地主要表现在精确度的提高,特别地,将机构名精度提升了 7.38 个百分点.

5.3 基于 CRF 的多特征中文命名实体识别

基于通过将词类与小规模常用提示字特征、字文本特征三者结合,利用它们在不同粒度上的互补性,达到一个最佳的特征组合(*TXT + POS + LIST*),在各语料集上均表现出一致的良好的效果.我们用 *MSRA* 语料训练的系统(*SIGHAN*)在 863 简体语料上进行了测试,各项指标均超出当年评测的单项最佳结果,文献[11]没有列出专名总体指标,文中为根据单项结果以及标准答案命名实体个数推测所得)以及六类命名实体总体最佳系统的结果.特别地,我们的系统在机构名识别上有着明显的提高.它在 *SIGHAN MSRA* 语料命名实体识别评测中也位于前列[7].在 863 语料上的各项具体评测指标列于图 3.



	LOC			ORG			Proper Nouns		
	F1	Prec	Rec.	F1	Prec	Rec.	F1	Prec	Rec.
SIGHAN	90.88	92.63	89.19	71.61	90.46	59.26	88.76	92.92	84.95
单项最佳	82.51	87.02	79.50	60.81	74.35	57.41	80.13	82.21	78.15
总体最佳	80.65	81.84	79.50	47.78	61.00	39.27	77.15	79.43	75.00

图 3 863 简体语料上的实验结果

5.4 小规模常用提示字特征对训练代价的影响

和大规模列表特征集相比,小规模用字列表避开了大规模列表收集困难,训练代价过高的缺点.为了考察小规模常用提示字特征对训练代价的影响,我们用 863 简体语料进行了训练时间对比实验,如图 4 所示.与基准系统相比,它所增加的训练代价非常小.在 Windows XP 环境下 256MB 内存,2.4GHz CPU 台式机上,基于 TXT + POS 特征集的字模型的训练时间约为 11.2 小时,而 TXT + POS + LIST 特征集的训练时间约为 11.9 小时,小规模列表没有带来明显的训练开销,仅增加训练时间约 6.25 %.与包含了人名用字、地名和机构名中心词和指界词列表、名称列表等 37 个特征的大规模特征集 POS + TXT + aLIST(即文献[4]中的 ALL 特征集)的训练时间 73 小时相比,所节省的训练时间相当可观.

上述三个实验所需虚拟内存分别为 700MB、750MB 和 1400MB.由于机器内存仅 256MB,虚拟内存比例较

** http://www.sighan.org/bakeoff2006/

*** http://www.chineseldc.org/resourse.asp

**** 引入人名是出于直接引用标准评测程序输出的需要

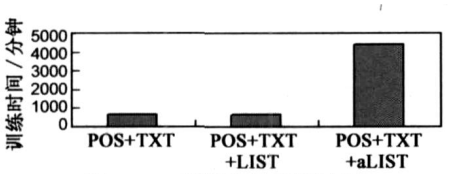


图 4　863 语料上的训练时间

大,这些数据不能完全说明训练速度的比率,但也表明了在计算资源受限的情况下,小规模特征对训练速度提升的作用是非常明显的(例如在该机器上引入后可将 POS + TXT + aLIST 特征集训练时间下降 83.71 %).

5.5　错误分析

尽管线性链(linear-chain)结构的条件随机场模型在线性序列分析上达到全序列最优,但它毕竟建立在一阶马尔科夫过程之上,缺乏对结构性信息的把握,因而在中文命名实体识别中的错误主要来自于长距离的语法和语义层面,有三种主要情形.

1. 并列的命名实体中部分未能识别或类型不一致.又分为两种,有共用成分和无共用成分. 如

×[花旗集团 ORG]、宝洁和[百时美施贵宝公司 ORG]

×[多米尼加共和国 LOC]、[安提瓜 PER]、[巴布达 PER]、[多米尼克国 LOC]

2. 对同一文本串多次出现给以不同的标注. 如

×天气好并不表示能看到[梅里十三峰 LOC],曾有个[波兰 LOC]的女孩为了一睹[梅里 PER][十三 NUM]峰的真容,……

3. 复杂结构未能识别. 如

×根据[伊拉克 LOC]临时政府内政部提供的数据,……

4. 语义信息把握不足. 如

×依据[美台 ORG]情报合作协定,……

6　结论与将来的工作

作为自然语言处理的关键技术之一,命名实体识别是一项非常复杂、需要综合多方面信息的处理技术. 条件随机场为命名实体识别提供了一个特征灵活、全局最优的标注框架,但存在收敛慢,训练时间长的问题. 在这个框架下,我们主要针对中文地名和机构名提出了将常见尾字作为命名实体识别的重要提示特征的思想.

一方面,小规模常用提示字特征具有一定的词类特征互补性,将基于文本的整体 F1 值提高 3 个百分点,特别是地名和机构名得到了显著的提高;当它与词类特征相结合时,可提高 4 - 8 个百分点. 另一方面,小规模常用字列表的引入,以较小的训练代价有效地提升了系统的性能,避开了常见的词语提示与名称列表、字号列表收集困难的缺点.

此外,我们的实验数据还表明,POS 对机构名识别有显著的帮助. 由此可以认为,更大范围的环境模式分析可能有助于机构名的识别. 这是因为机构名内部结构最为复杂,内部成分不稳定,从而需要借助外部环境的知识. 这应该成为我们下一步工作的方向.

另一个方向与前面所分析的错误类型对应. 针对并列和重现中出现的错误,以及语义不一致的现象,引入语法语义修正规则,或添加其它相关的特征.

由于分词、词类标注、浅层分析等与命名实体识别非常相似,本文提出的尾字提示特征也可推广应用到这些处理中.

参考文献:

[1] 刘非凡,赵军,吕碧波,等. 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.
F Liu, J Zhao, B Lv, et al. Study on product named entity recognition for business information extraction [J]. Journal of Chinese Information Processing, 2006, 20(1): 7 - 13. (in Chinese)

[2] J Lafferty, A McCallum, F Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data [A]. Proc of International Conference on Machine Learning [C]. San Francisco: Morgan Kaufman, 2001. 282 - 289.

[3] 俞鸿魁,张华平,刘群,等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报,2006,27(2):87 - 94.
H Yu, H Zhang, Q Liu, et al. Chinese named entity identification using cascaded hidden markov model [J]. Journal on Communications, 2006, 27(2): 87 - 94. (in Chinese)

[4] Y Feng, L Sun, J Zhang. Early results for chinese named entity recognition using conditional random fields model, HMM and maximum entropy [A]. IEEE Natural Language Processing & Knowledge Engineering [C]. Beijing: BUPT Publishing House, BUPT, 2005. 549 - 552.

[5] Z Kou, C William, M Robert. High-recall protein entity recognition using a dictionary [J]. Bioinformatics, 2005, 21 (Supplement 1): 266 - 273.

[6] H Zhao, C Huang, M Li. An improved chinese word segmentation system with conditional random field [A]. Ng HT, Kwong OY, eds. Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing [C]. Sydney: SIGHAN, July 2006. 162 - 165.

[7] Y Feng, L Sun, Y Lv. Chinese word segmentation and named entity recognition based on conditional random fields models [A]. The Third International Chinese Language Processing Bakeoff [C]. Sydney: SIGHAN, 2006. 181 - 184.

[8] 姜维,王晓龙,关毅,等. 基于多知识源的中文词法分析系统[J]. 计算机学报. 2007,30(1):137 - 145.
W Jiang, XL Wang, Y Guan, et al. Research on chinese lexical

analysis system by fusing multiple knowldge sources [J]. Chinese Journal of Computers, 2007, 30(1): 137 - 145. (in Chinese)

[9] 蒋建民,郭宏蕾,胡岗,等. 基于正则化 Winnow 算法的中文命名实体识别[A]. Proceedings of the 20th International Conference on Computer Processing of Oriental Languages [C]. 沈阳: ICCOL, 2003. 50 - 56.
J Jiang, H Guo, G Hu, et al. Chinese named entity recognition by regularized winnow algorithm [A]. Proceedings of 20th International Conference on Computer Processing of Oriental Languages [C]. Shenyang: ICCOL, 2003, 50 - 56. (in Chinese)

[10] 郎君,秦兵,刘挺,等. 中国人名性别自动识别[A]. 第三届学生计算语言学研讨会 [C]. 北京: SWCL, 2006. 166 - 171.
J Lang, B Qin, T Liu, et al. Gender recognition of chinese person name [A]. the Third Student Computational Linguistics [C]. Beijing: SWCL, 2006. 166 - 171. (in Chinese)

[11] 863 基础资源与评测. 命名实体评测结果报告[A]. 2004 年度 863 计划中文信息处理与智能人机交互技术评测 [OL]. http://www.863data.org.cn/. 2005
863 program. Results on named entity recognition [A]. The 2004 HTRDP Chinese Information Processing and Intelligent Human-Machine Interface Technology Evaluation [OL]. http://www.863data.org.cn/. 2005 (in Chinese)

作者简介:

冯元勇 男,1973 年生于湖南临湘,现为中国科学院软件研究所博士研究生,主要研究方向为命名实体识别与共指消解.
E-mail: ComerFeng@gmail.com

孙 乐 男,1971 年出生于陕西省,博士,现为中国科学院软件研究所副研究员,主要研究方向为自然语言处理,文本信息检索.

张大鲲 男,1980 年生于黑龙江哈尔滨,现为中国科学院软件研究所博士研究生,主要研究方向为统计机器翻译模型.

李文波 男,1975 年生于内蒙古临河,现为中国科学院软件研究所博士研究生,主要研究方向为文本分类与机器学习.