基于最小语言学资源的资源受限领域命名实体识别

涂兆鹏　姜文斌　刘群　林守勋
智能信息重点实验室
计算技术研究所，中国科学院

{tuzhaopeng, jiangwenbin, liuqun, sxlin}@ict.ac.cn

廖剑　吴克文
B2B 国际站算法组
阿里巴巴（中国）网络技术有限公司

{jian.liaoj, kewen.wukw}@alibaba-inc.com

摘要

如何在资源匮乏的大规模数据（如互联网纯文本数据）上识别命名实体是一个重要的问题。为解决该问题，我们使用简单的词典资源自动标注数据，然后将命名识别问题转化为基于最大熵马尔可夫模型的序列标注问题。我们使用两种方法搜索结果：1）输出标注序列，并使用重排序方法对 k-best 结果进行重排序；2）使用变种的前向-后向算法计算出候选命名实体的概率，并使用阈值过滤。实验表明，使用后种方法可以极大地提高命名实体识别的召回率和 F 值，并可以更加灵活地根据需求控制准确率与召回率。

1　引言

命名实体，是指人名、机构名、产品名以及其他所有以名称为标识的实体。当前主流的命名实体识别方法都是使用有监督的机器学习方法，依赖人工标注好的训练语料。随着互联网的发展，涌现了越来越多的网页数据。识别互联网数据中的命名实体对其他应用有很大的帮助，比如通过识别命名实体并加以翻译从而提高最终的翻译性能(Jiang et al., 2007; Zhao et al., 2008; Yang et al., 2008)。

如何识别存在大规模生语料、资源匮乏领域的命名实体，也引起了广大研究者的兴趣。如果人工标注大规模的数据，将是一个极其消耗人力和物力的工作，很不现实。传统的做法是使用人工维护的词典或者使用规则方法识别网络数据中的命名实体。但是，互联网数据中的命名实体具体如下特点：

1. 新出现的命名实体多。在互联网中，由于新名词层出不穷，所以出现了很多与之相关的新的命名实体。使用词典方法无法识别这些新词，而且人工维护词典也需要很高的人力成本。
2. 形式灵活多变。命名实体识别任务的困难之处在于歧义问题，即一个单词可能出现在命名实体的不同位置。比如电商领域中，单词 *screen* 可以出现在命名实体的不同位置：

   (a) **screen** guard mirror for iphone4S
   (b) large touch **screen** panel
   (c) high quality led advertising **screen**

   由于命名实体的形式灵活多变，使用规则的方法很难捕获这种规律。此外，规则的

方法需要有相关经验的人员手工构建，不仅效率低下，而且对于多语言语料情况不适用。

针对以上问题，我们首先使用简单的词典资源对大规模语料自动标注，然后将命名实体识别问题转化为传统的序列标注问题。在最大熵马尔可夫模型的框架下，我们可以使用丰富的特征集捕获上下文信息，从而有效地识别新出现的命名实体。我们不仅提供常用的输出句子的标注序列 1-best 结果和 k-best 的重排序结果，同时使用变种的前向-后向算法直接计算出候选命名实体的概率，并使用阈值过滤质量较差的命名实体。实验表明，使用后者不仅能极大地提高召回率和 F 值，也可以更加灵活地根据需要调节准确率与召回率。

本文结构如下，我们首先介绍相关工作，然后章节 2 讲述如何将命名实体识别转化为序列标注问题。章节 3 介绍最大熵马尔可夫模型和我们使用的特征。章节 4 介绍我们使用的两种搜索结果的方法。我们在章节 5 提供实验结果并加以分析，最后进行总结。

2 相关工作

我们的工作主要基于命名实体识别和标注模型两方面工作的基础上发展而来。目前国内外涉及产品命名实体识别的工作较少。Pierre (2002)使用类似字符匹配模型的简单分类器方法，而 Niu et al., (2003)使用自举方法，Liu et al., (2006)使用层级隐马尔可夫模型识别嵌套的产品词。在普通命名实体识别方面，Liu et al., (2007)使用多层嵌套的实体识别方法，利用条件随机场模型和支持向量机模型融合上下文语言学特征进行识别；Yu et al., (2006)使用层叠隐马尔可夫模型识别嵌套的人名、地名和机构名，Feng et al., (2008)使用改进的快速条件随机场方法识别中文命名实体，Qi et al., (2009)模拟语言习得的过程，通过分类器得到含有命名实体的碎片序列，进而抽取命名实体。Zhou et al., (2004)通过建立巨大的词和字串的集合，然后使用构词规则自动识别互联网上的中文新词；Wang et al., (2006)使用多种机器学习算法识别生物学中的命名实体，Liu et al., (2008)借助搜索引擎从搜索查询中抽取复杂的命名实体。

标注模型被成功应用于自然语言处理的多个任务中，比如词性标注(Ratnaparkhi, 1996)，句法分析(Ratnaparkhi, 1998)和中文分词(Xue, 2003)等。对于这些任务，结果是很自然的句子的全部序列标注。而命名实体识别任务只关注于识别出的候选命名实体，所以我们使用变种的前向-后向算法直接计算候选命名实体的概率，实验表明该概率能较好地反映命名实体的质量，从而极大地提高召回率和 F 值。

3 命名实体的序列标注模型

在应用机器学习算法之前，我们首先将语料中标注好命名实体的单词序列转换成标注序列。我们根据单词与命名实体的关系，将单词标注为五个标注 II, LL, MM, RR 和 LR 中的一个。标注详细信息见表 1。

表 1 标注说明

标注	说明
II	与命名实体无关的词语
LL	命名实体的左边界
MM	命名实体的中间词语
RR	命名实体的右边界
LR	单独词语作为命名实体

比如，我们可以将标注好命名实体(a)句子标注(b)（下划线表示识别出的命名实体）：

(a) screen guard mirror for iphone4S
(b) screen/LL guard/MM mirror/RR for/II iphone4S/LR

如果在识别过程中存在歧义问题，则句子中的某些单词会有多个可能的标注。比如，对

表 2 特征说明

特征	说明	示例
W_0	当前单词	led
W_{-2},W_{-1},W_1,W_2	前、后两个单词	high, quality, advertising, screen
$W_{-1}W_0$, W_0W_1	前、后单词与当前词的组合	quality-led, led-advertising
$W_{-2}W_{-1}$, W_1W_2	前、后两个单词组合	high-quality, advertising-screen
$W_{-1}W_1$	前一和后一单词组合	quality-advertising
T_{-2},T_{-1}	前两个词的标注	II, II

于上述句子(a)，前面两个单词有两个标注。下面显示了所有可能的四种标注结果：

(c) screen/LL|II guard/MM|LL mirror/RR for/II iphone4S/LR

和词性标记相似，一个单词的标注会受前面单词的标注影响。比如，当 *screen* 被标注为 LL 时，则其后的单词只能被标注为 MM 或 RR；而当 *screen* 被标注为 II 时，其后的单词只能被标注为 II，LL 和 LR。同样，一个单词的标注也会受该单词的周围单词影响。比如，当 *screen* 后面是 *guard mirror* 时，应该标注为 LL，而如果前面是 *touch*，后面是 *panel* 时，应该标注为 MM。这与词性标注任务相似，即在特定的上下文中，从多个可能的标注中选择正确的标注。我们下一步即是从标注好的语料中训练得到最大熵马尔可夫模型，以对新出现的句子进行自动标注。

4 训练

4.1 最大熵马尔可夫模型

最大熵马尔可夫模型（Maximum Entropy Markov Model, MEMM）是对隐马尔可夫模型（Hidden Markov Model, HMM）的一种改进。MEMM是条件概率模型，结合了隐马尔可夫模型和最大熵模型的特征。它并不去解释观察序列如何被生成，而是当给定观察序列时，努力去预测标号序列，这使得该模型可以使用观察序列的任意特征，包括全局特征、描述非局部交互的特征以及滑动窗口等。

假如我们有一串单词$\{w_1,...,w_n\}$，我们需要使用标注集$T=\{\mathrm{t}_1,...,\mathrm{t}_m\}$对其进行标注，以最大化条件概率$P(\mathrm{t}_{w_1},...,\mathrm{t}_{w_n}\mid w_1,...,w_n)$。在MEMM中，这个概率是马尔可夫转换概率，其中给某单词标注成某标注的概率依赖于单词位置及前面位置的标注（以二阶MEMM为例）：

$$P(\mathrm{t}_{w_1},...,\mathrm{t}_{w_n}\mid w_1,...,w_n)=\prod_{i=1}^{n}P(t_{w_i}\mid t_{w_{i-1}}t_{w_{i-2}},w_i)$$

在 MEMM中，传统HMM的转换函数和观察函数被单个函数$P(t\mid t',w)$所替代，这个函数给出了在给定以前的标注t'和当前的单词w的条件下转移到当前标注t的概率。MEMM从训练数据中学习$P(t\mid t',w)$，它是通过使用最大熵方法来使得该模型最大可能的与训练数据中的特征约束保持一致，这使得$P(t\mid t',w)$具有如下的指数形式:

$$P(t\mid t',w)=P_{t'}(t\mid w)=\pi\prod_{j=1}^{k}\lambda_j^{\,f_j(w,t)}$$

其中，π是归一化因子，$\{\lambda_1,...,\lambda_n\}$是需要被学习的模型参数，而$\{f_1,...,f_n\}$是模型所使用的布尔值特征。每一个特征$f_j$都有对应的参数$\lambda_j$，

也可称为该特征的权重。特征依赖于标注 t 和输入单词 w 的任何特征，如“w 是纯数字”、“w 是未登录词”。因此，MEMM支持长距离的特征依赖。

4.2 特征

MEMM 模型在标注任务中是否成功很大程度上依赖于是否选择了合适的特征。给定 (w,t)，好的特征需要包含信息以能够帮助正确预测标注 t。受(Yang et al., 2008)的启发，对于每个单词，我们使用表 2 所示特征（以句子 *high quality led advertising screen* 中的单词 *led* 为例）。

一般而言，给定 (w,t)，特征常常表示标注 t 和单词 w 上下文信息的共现关系。比如，

$$f_i(w_i,t_{w_i})=\begin{cases}1 & if\ \ w_{i+1}=for \text{ and } t_{w_i}=RR\\ 0 & otherewise\end{cases}$$

这个特征表示如果 w_{i+1} 是单词 *for* 且 w_i 被标注为 RR，则该特征值为 1。

我们使用的特征包含了两类上下文信息。1）词汇化上下文。对于句子中的每个单词，我们会考察当前单词，前面两个单词和后面两个单词。比如，对于单词*iphone4S*，它往往是单独成为一个命名实体，所以被标注为RR。而对于另外的单词，它的标注结果受其上下文信息影响。比如对于单词*screen*，如果其前面是*high quality*等修饰词，则其往往是某命名实体的起始单词，被标注为LL或LR；如果其后面是*for*等介词，则其往往是某命名实体的结束单词，被标注为RR或LR。2）前面出现的标注。这类信息对于预测当前单词的标注是非常有用的。比如，如果前面的单词被标注为LL，则表示前一个单词是某命名实体的起始单词，则当前单词是该命名实体的中间单词或结束单词，应该被标注为MM或RR。当训练过程结束时，这些特征及它们对应的权重将被用来自动标注新出现句子中的命名实体。

5 解码

与传统的词性标注或分词任务不同，我们不需要关注一个句子的全部标注结果，而只是关心其中可能的命名实体。所以对于新出现的某个句子，我们在搜索最优标注结果之外，同时使用变种的前向-后向算法给出所有候选命名实体的概率。

5.1 标注序列结果

搜索最优标注序列

该过程与HMM中确定最佳状态序列任务类似，我们同样使用Viterbi算法。在MEMM模型中，需要对Viterbi算法作适当修改：重新定义 $\delta_i(t)$ 为在给定位置 i 的单词的条件下，单词被标注 t 的概率值，这样可将Viterbi算法中的递归步骤改写如下:

$$\delta_{i+1}=\max_{t\in T}\delta_i(t')P(t\mid t',w_{i+1})$$

然后我们可以通过回溯得到最优标注结果以及 k-best 标注序列。

重排序

通过观察，我们发现有些例子正确命名实体往往不是出现在1-best结果中，而是k-best列表的其他结果中。所以，我们使用自然语言处理任务中常用的重排序方法。我们使用简单的二元分类方法，即将1-best结果标为正例，k-best中其他结果为反例。

我们通过考察结果中抽取的命名实体及其周围的词语，以评估该结果是否是正确的。除标注序列得分及当前标注序列包含候选命名实体数量等全局特征外，对于每个结果中，我们对其中出现的所有命名实体均使用如下特征：1）词汇化上下文。我们通过考察识别出的命名实体以及其周边词语，以判断该命名实体的边

界是否正确。2）命名实体在k-best其他结果中出现的次数。我们假设概率得分能反映命名实体的质量。所以好的命名实体概率得分较高，会更多地出现在其他结果中。当然，这也是下面直接计算命名实体概率的假设。

我们将使用十项交叉方法将训练语料分为十份，以得到每一份的识别10-best结果。再通过与标准答案的对比，得到重排序后的结果，以此构建重排序模型的训练语料。

5.2 带概率的命名实体列表

一个句子中往往包含大量与命名实体无关的单词（标注为II的单词），而这些单词的标注结果会极大地影响到最后的标注序列结果。考虑到我们的任务是识别出句子中可能存在的命名实体，所以我们直接计算候选命名实体的概率，以衡量其质量的好坏。

受文献(Mi and Huang, 2008; Tu et al., 2010)的启发，我们使用变种的前向-后向算法计算某个候选命名实体的概率。首先，我们要将最大熵隐马尔可夫模型的格路（格路是由状态相对于单词位置组成的一个方阵）转化成概率化的超图。由于我们使用最大熵的分数代替了传统HMM的转换函数和观察函数，并且我们使用的是二阶的隐马尔可夫模型（使用前两个单词的标注状态）。所以我们使用以下公式将单词 w_n 的标注经过某条边的分数概率化：

$$p_n(i,j) = P(t_{w_n} = i, t_{w_{n+1}} = j)$$

$$= \sum_{k=1}^{m} P'(t_{w_{n+1}} = j \mid t_{w_n} = i, t_{w_{n-1}} = k, w_n)$$

其中 $P'(t_{w_{n+1}} = i \mid t_{w_n} = i, t_{w_{n-1}} = k, w_n)$ 是经过归一化后的最大熵输出分数。

给定一个候选命名实体 T=$\{t_{w_k}, \ldots, t_{w_l}\}$

$$\alpha\beta(T) = \alpha_k(t_{w_k}) \times \beta_l(t_{w_l}) \times \prod_{i=k+1}^{l} p_i(t_{w_{i-1}}, t_{w_i})$$

其中，前向变量 $\alpha(t_{w_k})$ 表示单词 w_k 以状态 t_{w_k} 结束时总的概率，而后向变量 $\beta(t_{w_l})$ 表示单词 w_l 除标注 t_{w_l} 以外剩余部分的概率之和。

同样，我们需要对传统HMM模型中的内向概率和外向概率的计算过程作适当修改：

$$\alpha_{i+1}(t_j) = \sum_{k=1}^{m} \alpha_i(t_k) \times P(t_j \mid t_k, w_i)$$

$$\beta_i(t_j) = \sum_{k=1}^{m} \beta_{i+1}(t_k) \times P(t_k \mid t_j, w_{i+1})$$

则候选命名实体 $\{t_{w_k}, ..., t_{w_l}\}$ 的概率为：

$$p(\{t_{w_k}, ..., t_{w_l}\}) = \frac{\alpha\beta(\{t_{w_k}, ..., t_{w_l}\})}{\alpha_{n+1}(ROOT)\beta_{n+1}(ROOT)}$$

其中，ROOT节点为虚节点，$\beta_{n+1}(ROOT) = 1$。

我们会得到所有可能命名实体以及对应的分数。我们同样可以按照分数取前k个命名实体或者使用阈值过滤质量较差的命名实体。

6 实验

本节，我们尝试回答以下三个问题：

1 我们的模型能否较好地识别未在训练语料中出现的命名实体？（章节5.1）
2 我们的模型是否能提供较高质量的候选命名实体列表？（章节5.2）
3 我们的模型是否有较好的领域鲁棒性？（章节5.3）

本次实验,我们主要识别阿里巴巴产品英文页面标题中的产品命名实体.我们使用电子、电器、化工、服装和食品五个领域共 1.2M 标题，5.6M 单词。其中各语料标题数基本相当。由于语料库规模较大，所以我们使用人工维护的百万级的产品命名实体词典对其进行自动标注。各语料库及命名实体分布情况见表 3。

注意到，为了保证我们实验的合理性，我们做了以下设置：

表 3　语料中命名实体分布表

语料		电子	电器	化工	服装	食品
训练集	标题数	234,362	246,867	210,671	247,918	246,019
	实体数	198,738	78,388	119,290	221,936	147,800
测试集	标题数	1,428	914	1,062	659	2,103
	实体数	1,270	538	817	518	1,622

表 4　识别新词实验结果

模型		准确率	召回率	F 值
输出标注序列	1-best 结果	0.9247	0.4661	0.6198
	重排序结果	0.9467	0.6035	0.7371
输出命名实体	1-best 结果	0.7496	0.8188	0.7827
	概率≥0.1	0.7482	0.8234	0.7840

(1). 为了考察我们的模型识别新词的能力，我们保证测试集中的产品命名实体不在训练集中出现；

(2). 为了考察命名实体候选方法的好坏，我们在各领域测试集中加入了适量的不包含任何命名实体的标题，以防止不考虑分数而取1-best候选命名实体的结果不可靠；

(3). 为了考察模型在不同子领域上表现如何，所以我们选择了电子，电器，化工，服装和食品这五个领域；

考虑到电商领域的产品命名实体形式灵活，有些命名实体边界尽管和答案不一样，但是也可以接受。对于位置正确但边界错误的情况，我们在原有基础上给予了0.5的折扣分。

6.1　识别新词能力

从表4中我们可以看到，当候选命名实体被识别为命名实体的概率低于识别为非命名实体的概率时,输出标注序列的1-best结果会偏向于将其识别为非命名实体。所以使用1-best标注序列结果存在着高准确率和低召回率的问题。而使用重排序模型可以一定程序上解决该问题，通过使用标注序列的全局信息和候选命名实体的上下文信息，我们可以极大地提高召回率。

由于命名实体的概率能较好地反映命名实体的质量，所以通过输出候选命名实体的1-best结果或高于概率阈值的结果，可以将更多的候选命名实体识别为命名实体，从而提高召回率。我们同时需要注意，一个句子中往往含有多个命名实体，所以直接使用候选命名实体的1-best结果显然不合适。针对于此，我们采用了更加灵活的方法，使用概率阈值对其进行过滤。这种方法更灵活地针对任务要求调节准确率和召回率（章节5.2）。此外，我们通过人工检验结果发现，很多高于阈值的候选命名实体即使未匹配上答案（未在产品命名实体词典中），仍然是较好的命名实体。下面所有实验均是使用直接输出命名实体的方法。

6.2　阈值对命名实体识别性能的影响

本实验考察阈值对命名实验识别性能的影响（见图1）。我们使用了10个不同的阈值：0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8和0.9。显然，阈值越低，命名实体识别的准确率越低，而召回率越高。当阈值为0.01时，准确率为

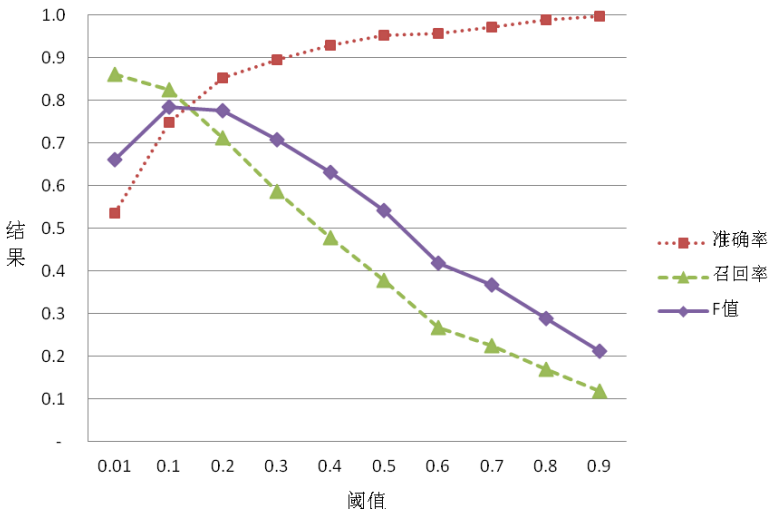

图 **1** 命名实体识别性能随阈值变化曲线

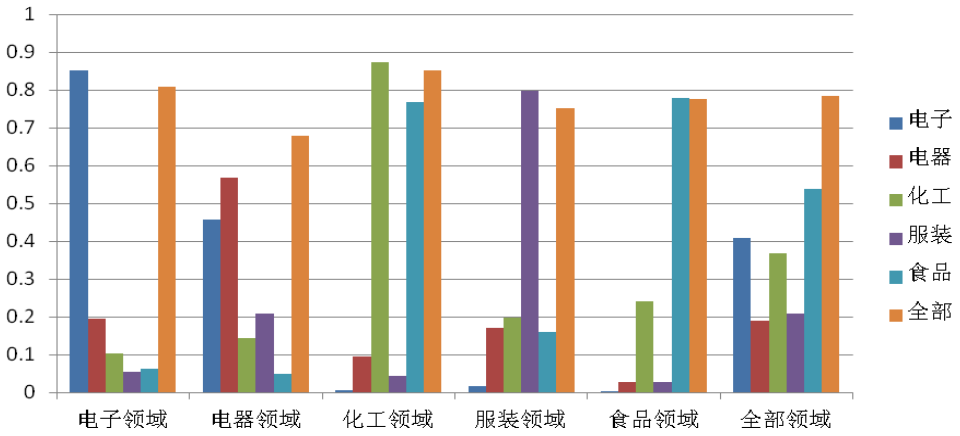

图 **2** 命名实体识别性能受领域影响

53.51%，而召回率为86.08%。随着阈值的增长，准确率不断增加，而召回率不断降低。这也反映了命名实体的概率可以较好地反映命名实体的质量。为了在准确率和召回率间取得平衡，我们在以后的实验中取阈值为0.1。

我们可以看到通过调节阈值，我们可以很方便地调节准确率与召回率。比如，当需要尽可能准确地识别命名实体时，我们可以将阈值设的较高（如0.8）；而当需要尽可能多的候选命名实体时，我们可以将阈值设的较低（如0.01）。

6.3 领域适应性

电子商务领域中存在着很多子领域，比如电子，化工，服装等领域。我们如何训练一个好的模型，使之能在各子领域中均有良好表现？本实验主要考察两个问题：

1）在某一子领域上训练得到的模型在其他领域上表现如何？各领域差别是否较大？

2）使用全部子领域的语料训练得到的模型能否在各子领域上有良好表现？

图2显示了实验结果。从图中我们可以看出，各子领域训练得到的模型在本领域往往取得最好结果，而在其他领域表现很差。这说明各子领域间的差异很大。而在全部语料上训练得到的模型，在各子领域的表现和子领域本身训练得到的模型表现相当。说明使用全部语料可以较有效地克服领域差异性，在各子领域上都能有良好表现。

7 结语

针对资源受限领域生语料较多，但相关资源较少的特点，我们使用简单的实体词典自动标注生语料，并将命名实体识别问题转换为基于最大熵马尔可夫模型的序列标注问题。我们的创新点主要有以下两个方面：

(1). 使用较易获得的语言学资源解决资源受限问题。我们的方法可以广泛应用于其他资源受限领域的命名实体识别任务，比如专利文档，信息安全等领域。

(2). 常用的输出最优标注序列结果及重排序结果外，我们使用变种的前向-后向算法直接计算候选命名实体的概率，并通过阈值控制输出列表。实验表明，使用该方法不仅可以极大地提高召回率及 F 值，同时能更灵活地根据需求控制准确率和召回率。

致谢

本工作受 863 重大项目课题 (2011AA01A207)资助。部分工作是作者在阿里巴巴（中国）网络技术有限公司实习时完成。

参考文献

蒋龙, 周明, 简立峰. 2007. 利用音译和网络挖掘翻译命名实体. 中文信息学报, 21(1): 23-29.

赵军, 杨帆. 2008. 利用单语网页挖掘辅助汉英人名反向音译. 第四届全国学生计算语言学会议.

Fan Yang, Jun Zhao, Bo Zou, Kang Liu, and Feifan Liu. 2008. Chinese-English Backward Transliteration Assisted with Mining Monolingual Web Pages. In Proc. of ACL, pages 541-549.

John M. Pierre. 2002. Mining Knowledge from Text Collections Using Automatically Generated Metadata. In Proc. of PAKM, pages 537-548.

Cheng Niu, Wei Li, Jihong Ding, and Rohini K. Srihari. 2003. A Bootstrapping Approach to Named Entity Classification Using Successive Learners. In Proc. of ACL, pages 335-342.

刘非凡,赵军,吕碧波,徐波,于浩,夏迎炬. 2006. 面向商务信息抽取的产品命名实体识别研究. 中文信息学报, 20(1): 7-13.

刘非凡,赵军, 徐波. 2007. 实体提及的多层嵌套识别方法研究. 中文信息学报, 21(2): 14-21.

俞鸿魁,张华平,刘群,吕学强,施水才. 2006. 基于层叠隐马尔可夫模型的中文命名实体识别. 通信学报, 27(2): 87-94.

冯元勇,孙乐,李文波, 张大鲲. 2008. 基于单字提示特征的中文命名实体识别快速算法. 中文信息学报, 22 (1): 104-110.

齐振宇,赵军,杨帆. 2009. 一种开放式中文命名实体识别的新方法. 第五届全国信息检索学术会议论文集: 587-595.

邹纲,刘洋,刘群,亢世勇. 2004. 面向 Internet 的中文新词语检测. 中文信息学报, 18(6): 1-9.

王浩畅,赵铁军,刘延力, 于浩. 2006. 生物医学文本中命名实体识别的智能化方法. 北京邮电大学学报, 29(2): 54-58.

Hui Liu, Xueyin Hu, Jinglei Zhao, Maosheng Zhong, Ruzhan Lu. 2008. Identification of Complex Named-Entities in Chinese Queries Using WWW. In Proc. of FSKD, pages 180-185.

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In Proc. of EMNLP. pages 133-142.

Adwait Ratnaparkhi. 1998. Maximum Entropy Models for Natural Language Ambiguity Resolution. Ph.D. thesis.

Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. In International Journal of Computational Linguistics and Chinese Language Processing, pages 29-48.

Haitao Mi and Liang Huang. 2008. Forest-based translation rule extraction. In Proc. of EMNLP, pages 206-214.

Zhaopeng Tu, Yang Liu, Young-Sook Hwang, Qun Liu, and Shouxun Lin. 2010. Dependency Forest for Statistical Machine Translation. In Proc. of COLING, pages 1092-1100.