统计和规则相结合的汉语最长名词短语自动识别

代翠 周俏丽 蔡东风[1]

（沈阳航空工业学院知识工程中心，辽宁 沈阳 110034）

Email：xiaodai151@yahoo.com.cn

摘要：本文提出了一种基于统计和规则相结合的汉语最长名词短语识别方法，首先使用条件随机场模型进行标注识别，再结合最长名词短语的边界信息和内部结构信息对错误识别结果进行后处理。实验结果表明，本文提出的统计和规则相结合的自动识别方法是有效的，开放测试结果 F 值达到了 90.2%。

关键词：条件随机场，名词短语，最长名词短语

Automatic Recognition of Chinese Maximal-length Noun Phrase Based on Statistics and Rules

DAI Cui, ZHOU Qiao-li, CAI Dong-feng

(Knowledge Engineering Research Center, Shenyang Institute of Aeronautical Engineering, Shenyang, Liaoning 110034)

Email：xiaodai151@yahoo.com.cn

Abstract: The research proposes an automatic recognition method of Chinese maximal-length noun phrase based on statistics and rules. Firstly it uses conditional random fields (CRF) model to label and recognize the corpus, then a post-processing step based on the boundary information and inner structure knowledge of maximal-length noun phrase is used to correct the wrong labeled result. The experiment results show the method is efficient for identifying Chinese maximal-length noun phrase, of which the F-value reaches 90.2 % in open test.

Key words: conditional random fields, noun phrase, maximal-length noun phrase

1. 引言

名词短语的识别是自然语言处理中一项重要的子任务。它的识别结果可以简化句子结构，降低句法分析的难度和复杂度，为进一步的短语分析、句法分析等提供基础。从组成结构上看，句子中的名词短语可分为以下3类：(1) 最短名词短语(minimal noun phrase，简称mNP)：不包含其他任何名词短语的名词短语；(2) 最长名词短语(maximal noun phrase，简称MNP)：不被其他任何名词短语所包含的名词短语；(3) 一般名词短语(general noun phrase，简称GNP)：所有不是mNP和MNP的名词短语[1]。名词短语的识别难度随长度的增加而增加，其中最长名词短语的识别最为困难，但它的自动识别具有更为重要的意义。

如下句中“珠宝商泰勒先生”与“一个新陈列的橱柜”为句中的两个MNP。

[珠宝商 泰勒 先生]$_{MNP}$ 正 在 欣赏 [一个 新 陈列 的 橱柜]$_{MNP}$ 。

从句法功能上来看，MNP一般出现在句子的主语或宾语的位置，所以只要识别出句子中所有的MNP，就可以很容易把握句子的整体结构框架，从而很快构建出句子的完整句法树（森林），因此MNP的识别和分析有助于浅层句法分析[2]。此外，识别MNP对于信息抽取、机器翻译等应用领域也有重要意义。

在基于实例的机器翻译系统中，翻译模板反映的是一句话的结构组成，且模板中的常量一般为句子中的谓语成分，变量则是由命名实体或名词短语组成，因此只要识别出句子中的所有MNP，就很容易能得到句子对应的句型模板，进而有助于机译中复杂长句的翻译。

作者简介：代翠（1983-），女，研究生，主要研究方向为自然语言处理、最长名词短语识别；周俏丽（1977-），女，硕士，主要研究方向为自然语言处理、句法分析；蔡东风（1958-），男，博士，教授，主要研究方向为人工智能理、自然语言处理。

特定的名词短语（产品名、事件、场所等）和动词短语（事件描述、事实陈述）的识别对于信息抽取过程有重要的意义。其中对于大部分名词实体的识别，可以通过识别和分析 MNP 来解决。

本文的其它部分组织如下：第二部分介绍MNP识别的相关研究工作，第三部分介绍了汉语MNP的识别，第四部分为条件随机场统计模型的介绍，第五部分详细介绍本文MNP识别方法，第六部分为实验结果及相关分析，最后给出本文工作的总结和下一步工作展望。

2. 相关研究工作

近几年来，国内外研究人员在 MNP 的自动识别方面进行了许多有益的探索，提出了一些行之有效的识别方法，方法主要有基于规则的方法和基于统计的方法。

在较早的研究中，Bourigault[3]介绍了一个法语的术语抽取系统 LEXTER，主要是利用一个 MNP 边界标记的规则库来识别句子中的最长名词短语；Voutilainen 的 MNP 抽取工具 NPtool[4] 实现了英文 MNP 的识别，实验系统建立在两个规则库、一个带有词性标记与中心词信息的词表的基础上，利用两种有限状态分析机制（NP-否定机制和 NP-肯定机制）来发现英文中可能的 MNP；Kuang-hua Chen 等人[5]基于语料库进行英文 MNP 识别，利用基于统计的组块分析和基于规则的有限状态机制相结合的方法来发现句子中的 MNP。这些有益的探索为 MNP 识别积累了宝贵的经验，但它们大都以规则方法为基础，开发周期长、投入大，而且与特定语种紧密相关。

在中文处理领域，李文捷等人[6]最早研究了 MNP 的识别问题，通过统计边界分布概率信息自动识别句中 MNP 的起始和终止位置，在 30 篇新闻报道语料中，开放测试正确率达到了 71.3%。周强等[1]提出的基于内部结构组合的汉语 MNP 识别算法，正确率和召回率分别为 85.4% 和 82.3%。该算法前期要对语料进行基于统计的组块识别，在此基础上利用规约规则来发现可能的 MNP，且前期组块分析产生的错误也会影响 MNP 的识别结果。冯冲，陈肇雄等[7]利用条件随机场建立统计模型来识别句子中的复杂最长名词短语，实验中仅利用了统计模型，且仅针对复杂 MNP，封闭测试正确率和召回率为 75.4%和 70.6%。

从这几年来的一些研究实践来看，汉语 MNP 自动识别效果并不是很理想。本文在分析现有研究方法的基础上，提出一种基于错误驱动的统计与规则相结合的汉语 MNP 自动识别方法。

3. 汉语最长名词短语识别

与外文相比，汉语 MNP 的识别更加困难，这是由汉语句法成分特有的套叠现象[8]所决定的，因为汉语中的任何句法成分都可以不经过任何形态变化，只需加上一个结构助词“的”，就可以充当一个 NP 的定语，这就大大增加了汉语 MNP 自动识别的难度。如下面的例句所示：

[一 个 于 [[半 个 世纪] 之 后] 重新 聚集 在 [[西南 联大] 旗帜] 下 的 奉献 活动]$_{MNP}$ 开始 了 ！

句中的 MNP“一个于半个世纪之后重新聚集在西南联大旗帜下的奉献活动”的核心成分“奉献活动”的定语正是通过助词“的”形成的长定语，这种通过助词“的”构成的复杂定语使得基于规则的方法建立规则库变得更加困难，而基于统计的方法对于这种长距离关联特征“一 个”与“奉献 活动”也不能取得正确的识别结果。在这种情况下，考虑将两种方法有效结合起来，在已有统计模型识别结果的基础上，通过制定一定规模的规则再次对识别结果进行补充识别和修正部分错误情况。

在此，将 MNP 识别问题形式化地看作是一个标注问题[7]，即对于输入的句子 S= w_1/p_1 w_2/p_2 ..... w_N/p_N，（N 为词数，w_i 表示第 i 个词，p_i 为相应的词性标记，$1 \leqslant i \leqslant N$），给出一个与之对应的标注序列 $T^* = t_1 t_2 \cdots t_N$，满足 $T^* = \arg\max_T p(T \mid S)$，其中 $t_i \in \{B, I, O\}$，B 代表当前词

为一个 MNP 的首词，I 表示当前词属于 MNP（除 MNP 首词外），O 表示当前词不属于 MNP。这样，MNP 识别就可以转化为一个典型的序列数据的标注任务。

4. 条件随机场模型

条件随机场(Conditional Random Fields, CRF)是 John Lafferty 等人于 2001 年提出的一种基于统计的序列标注和分类模型，也是在给定输入节点条件下计算输出节点的条件概率的无向图模型[9]。它不需要输出独立性及序列数据严格独立等假设，因为大多数序列数据不能被表示成一系列独立事件。CRF 采用一种概率图模型，具有表达长距离依赖性和交叠性特征的能力，能够较好地解决标注（分类）偏置等问题，并求得全局的最优解。CRF 已经广泛应用到词性标注、组块识别和命名实体识别等任务中，并且相对于最大熵等机器学习方法取得了很好的效果。

定义 $x = x_1 \cdots x_N$ 为给定的输入观测值序列，即无向图模型中 N 个输入节点上的值，如当前输入的中文词序列；定义 $y = y_1 \cdots y_N$ 为输出的状态序列，即无向图模型中 N 个输出节点上的值，如输出的标记序列。CRF 定义从输入 x 得到序列 y 的条件概率定义为：

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_{i=1}^{N} \sum_{k} \lambda_k f_k (y_{i-1}, y_i, x, t) \right)$$

其中，$Z(x)$ 为一个归一化函数，使得在给定输入上的所有可能的状态序列的概率和为1，$f_k (y_{i-1}, y_i, x, t)$ 表示一个特征函数，λ_k 是与 f_k 相关的权重参数，反映了特征函数所代表的事件发生的可能性，可在训练中得到。

本文选择条件随机场作为 MNP 识别的统计模型，是因为作为条件模型它能够综合利用对字、词、词性等多层次的资源，同时，对于长程关联有很好的描述能力，并能避免其他模型中存在的标注偏置问题，能够符合 MNP 识别的需要。

5. 统计和规则相结合的汉语 MNP 识别

图 1 为统计和规则相结合的汉语 MNP 识别系统流程。该系统可以分为两部分：（1）基于 CRF 的 MNP 自动识别器；（2）基于规则的后处理模块。首先系统在训练语料的基础上根据特征选择和参数估计建立 CRF 统计模型，对于未经标注的测试语料，进行基于 CRF 的 MNP 识别，得到初步识别结果；然后采用基于规则的方法对识别结果进行后处理，得到最终的识别结果。

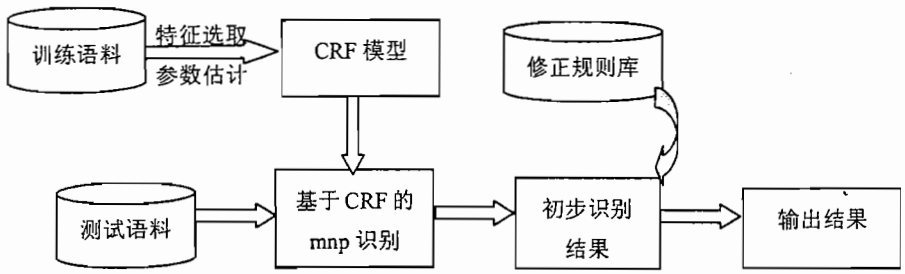

图 1 系统流程图

5.1 基于 CRF 的 MNP 自动识别

基于 CRF 的自动识别器中，特征函数的选取对识别性能起着关键性的作用。对于 MNP 识别问题，边界分部信息和内部结构组合知识能够为 MNP 的识别提供强有力的支持。特征函数也

是围绕这些因素选取的，一个句子经过分词标注后可以表示为“w_1/p_1 w_2/p_2 ...w_i/p_i...w_n/p_n”，w_i 为当前词，对应词性为 p_i，反复实验后，选取下表（表 1）中的 template3 作为进行下一步实验的特征模板。

下表（表 1）为在不同特征模板下训练的 CRF 模型自动识别 MNP 的效果：

表 1 特征模板对比实验

	特征说明	p	r	f
template1	前后各两个及当前的词	73.9%	68.2%	70.9%
template2	前后各两个及当前的词和词性	83.6%	82.5%	83.0%
template3	前后各两个及当前的词和词性、复合特征	85.8%	85.3%	85.5%
template4	前后各三个及当前的词和词性	83.4%	82.4%	82.9%
template5	前后各三个及当前的词和词性、复合特征	**86.4%**	**86.4%**	**86.4%**
template6	前后各四个及当前的词和词性	83.2%	81.8%	82.5%
template7	前后各四个及当前的词和词性、复合特征	85.8%	84.4%	85.1%

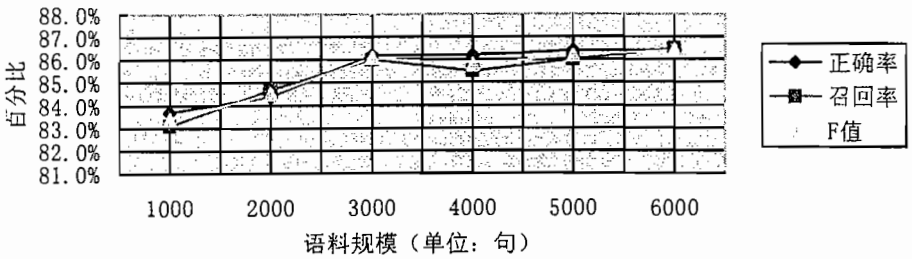

图2 训练语料数量与CRF性能比较

从上表 1 可以看出，随着特征的增加，系统性能并没有提高，这可能与两方面的原因有关，一是在选择特征时，因为 MNP 的复杂性，不能保证特征的有效性；二是与训练语料有关。但从上图 2 CRF 性能对比中可以看出当训练语料达到一定规模（3000 句）时，随语料的增加 CRF 识别性能已不再成比例上升，说明当达到一定规模时，语料数量对于识别性能的影响已不是很明显。因此对于以上情况，通过分析错误实例，进一步引入语言学知识进行后处理。

5.2 基于规则的后处理模块

通过分析错误实例，发现部分 CRF 的错误标注结果与实际情况明显不符，对于这些错误可以制定规则来修正。为保证 CRF 初步标注结果中正确的标注，在此修正规则仅是对 CRF 标注结果进行补充识别和纠正错误标注。主要针对以下比例大且易于用规则修正的五种情况进行查缺补漏：

（1） 简短整句被识别为一个 MNP：利用基于规则的方法重新识别；

例：[其中/r 一/m 位/q 飞行员/nc 想/vg 超过/vg 飞行/vg 物/ng] 。/wj

修正后：[其中/r 一/m 位/q 飞行员/nc] 想/vg 超过/vg 飞行/vg 物/ng 。/wj

（2） 缩短 MNP：对于具有明显左边界信息的缩短 MNP 进行补充识别；

例：那些/r 主张/vg [“/wyl 妇女/nc 解放/ng ”/wyr 的/usde 人/ng] 未免/d 要求/vg 过/d 高/a 。/wj

修正后：[那些/r 主张/vg “/wyl 妇女/nc 解放/ng ”/wyr 的/usde 人/ng] 未免/d 要求/vg 过/d 高/a 。/wj

（3）含有固定搭配的 MNP：利用固定搭配词表修正；

例：在/p 那儿/r 他们/r 学/vg 到/vq 了/ut 在/p [冰雪/ng 地/ng 里/f 生活/vg 的/usde 本领/ng] 。/wj

修正后：在/p 那儿/r 他们/r 学/vg 到/vq 了/ut [在/p 冰雪/ng 地/ng 里/f 生活/vg 的/usde 本领/ng] 。/wj

（4）漏识别的 MNP：基于规则的方法进行查漏识别；

例：在/p [那/r 次/q 聚会/ng] 上/f 我/r 身体/ng 不太/d 好/a 。/wo

修正后：在/p [那/r 次/q 聚会/ng] 上/f [我/r 身体/ng] 不太/d 好/a 。/wo

（5）简单并列结构：基于并列词表并结合标注结果进行修正。

例：我们/r 只有/vg [一/m 个/q 共同/b 的/usde 目的/ng] ，/wo [一/m 个/q 共同/b 的/usde 想法/ng] 。/wj

修正后：我们/r 只有/vg [一/m 个/q 共同/b 的/usde 目的/ng ，/wo 一/m 个/q 共同/b 的/usde 想法/ng] 。/wj

针对以上五种情况，规则库和处理方法说明如下：（1）、（4）中基于规则的方法是通过统计 MNP 词性串建立的简单有限自动机；（2）中的左边界信息是通过统计训练语料中的 MNP 左边界信息并结合普遍左边界错误情况获得的规则库；（3）中固定搭配词表为手工整理的 774 条固定搭配词，包括左右匹配标点等；（5）中并列词表既包括连词词表也包括标点符号连接并列成分的情况。

后处理模块中，基于规则的后处理流程如下所示：

背景知识：有限自动机 FSM，边界信息规则库 Blist，固定搭配词表 Dlist，并列词表 Clist 及简单并列规则

输入：CRF 识别 MNP 后的标注文件

基本操作：

① 输入句子序列 $w_1/p_1/t_1\ w_2/p_2/t_2 \ldots\ldots w_n/p_n/t_n$ （w_i：词，p_i：词性，$t_i \in \{B、I、O\}$，$1 \leqslant i \leqslant n$），序列以句号、感叹号、问号作为分割点。

② 若序列满足任意 t_k!=O，$k \in [1,n-1]$，利用 FSM 重新标注，否则执行③

③ 若 $t_i \ldots\ldots t_j$ 满足 t_i=B，t_{j+1}!=I 且任意 t_k=I，$k \in [i+1, j]$，则记为 MNP_{ij}

若 $t_i \ldots\ldots t_j$ 满足 t_{i-1}!=O，t_{j+1}!=O，且任意 t_k=O，$k \in [i, j]$，则记为 NNP_{ij}

分别执行以下操作直至序列结束：

a) 若 t_i=O，且 $w_i/p_i \in$Blist，从 i 开始规约直至 t_j=B 形成新的 MNP

b) FSM 处理 NNP_{ij} 进行漏识别标注

c) 若(w_i And w_j)∈Dlist，且 t_i=O&&t_j!=O 或 t_i!=O&&t_j=O，利用 Dlist 修正为 w_i、w_j 同在 MNP 内或外

④ 对于每个 $w_k/p_k/t_k$ （$1 \leqslant k \leqslant n$）若 t_k=O，且 $w_k/p_k \in$Clist，若存在 MNP_{ik}、MNP_{kj} 则形成 MNP_{ij}

基本流程：对于文件的所有序列执行以上操作，直至文件结束

---

6. 实验结果及分析

6.1 实验准备

本文 MNP 识别实验语料由哈工大提供的 10000 句汉语树库转化而来，经过处理提取其中的 7330 句作为实验语料，每句含有词性标注信息，MNP 标注信息，平均句长为 14（个词）。随机选取 6330 句为训练语料，包含 10281 个 MNP，剩余 1000 句为测试语料，包含 1605 个 MNP。

对于 MNP 识别性能的评价沿用自然语言处理中常见的评价方法：

MNP准确率：P=N3 / N2

MNP召回率：R=N3 / Nl

以及综合反映二者的指标 $F=(\beta^2+1)\times P\times R/(R+\beta^2\times P)$，$\beta=1$

其中：N1：语料中实际的MNP数量

N2：系统自动识别出的MNP数量

N3：系统正确识别出的MNP数量

系统对“正确的标记”采用了严格的定义，即当且仅当 MNP 的左右边界都被正确识别。

6.2 最长名词短语识别实验

实验中，将最长名词短语分为两类：词数<5 的简单最长名词短语（Simple MNP，SMNP）和词数≥5 的复杂最长名词短语（Complex MNP，CMNP）。表 2 为两种方法下进行的 MNP 识别实验结果对比，并分别列出了 SMNP 及 CMNP 的识别结果。

表 2 MNP 识别效果对比

方法	分类	N1	N2	N3	P	R	F
CRF 模型	SMNP	1172	1174	1076	92.5%	91.8%	92.2%
	CMNP	433	432	310	70.1%	71.6%	70.9%
	合计	1605	1606	1386	**86.3%**	**86.4%**	**86.3%**
CRF 模型+规则方法	SMNP	1172	1163	1102	94.8%	94.0%	94.4%
	CMNP	433	444	346	77.9%	79.9%	78.9%
	合计	1605	1607	1448	**90.1%**	**90.2%**	**90.2%**

从开放测试结果来看，在仅用 CRF 模型时 MNP 识别正确率和召回率分别为 86.3%、86.4%，而 CRF 模型+规则方法中 MNP 识别正确率和召回率分别为 90.1%、90.2%，系统总体性能提高了大约 4 个百分点，表明系统在后期加入规则方法进行辅助修正的方法是有效的。

在仅用 CRF 统计模型时，系统开放测试中 CMNP 正确率和召回率分别为 70.1%和 71.6%，而[7]中封闭测试结果为 75.4%和 70.6%，CRF 模型+规则方法中 CMNP 的正确率和召回率比仅用 CRF 统计模型时分别提高了大约 8 个百分点，表明该方法对于 CMNP 的识别是有效的。对于 SMNP 的识别在两种方法中均显示了比较高的性能，因此对于复杂最长名词短语的研究是下一步研究的重点。

为进一步刻画系统对于 MNP 左右边界的识别能力，同时测试了左边界 LB、右边界 RB 的准确率 P_B，召回率 R_B 和二者的综合 F_B。下表为测试结果：

表 3 左右边界测试

	N1	N2	N3	P_B	R_B	F_B
LB	1605	1607	1472	90.9%	91.0%	90.9%
RB	1605	1607	1529	95.1%	95.3%	95.2%

从上表 3 可以看出，右边界识别正确率要高于左边界，反映了汉语 MNP 识别的难点在于确定其左边界，因此如何确定复杂定语的左边界位置将是今后研究的一个重点。

6.3 错误实例分析

通过对实验中的 123 个识别错误实例的内部结构组合进行分析，发现其中句法歧义结构占了很大比例。考虑结构组合“V NP 的 NP”，对于不同的词语有不同的分析结构，如“毫无/vg[接受/vg 那/r 邀请/ng 的/used 意向/ng]”与“接受/vg[那/r 人/ng 的/used 道歉/ng]”。如下面的例句中均含有类似的轻歧义结构，导致识别错误。

她/r B 毫无/vg O 接受/vg O 那/r B 邀请/ng I 的/used I 意向/ng I 。/wj O

她/r B 惊奇/a O 得/usdf O 瞪/vg O 大/a B 眼睛/ng I 。/wj O

我/r B 附上/vg O 贴/vg O 有/vg O 邮票/ng B 并/c O 写/vg O 好/a B 地址/ng I 的/used I 回

/vg I 邮/vg I 信封/ng I 。/wj O

目前文中的识别算法，仅仅依靠词性规则不能对类似情况进行正确识别，从而导致了类似结构识别正确率和召回率的下降。以后，主要针对歧义结构进行研究，尝试引入语义信息来提高他们的识别性能，即能提高整体的识别性能。

7. 结论

最长名词短语的自动识别作为一项重要的应用基础研究，不仅有助于浅层句法分析，并且对于自然语言处理领域中的许多应用研究，如信息检索、信息抽取、机器翻译等，都具有重要的实践意义。本文在分析现有方法的基础上，通过分析对比不同方法的识别情况，提出了一种基于统计和规则相结合的 MNP 自动识别方法。统计和规则方法并用，对文本进行 MNP 识别，提高了识别的正确率和召回率。

在今后的研究中，将进一步扩大语料规模，完善后处理规则库，并从以下两个方面提高自动识别的性能：

1） 集中于复杂 MNP 的识别研究；

2） 针对歧义结构等未解决的问题，尝试引入语义信息来提高自动识别的性能。

参 考 文 献

[1] 周强, 孙茂松, 黄昌宁. 汉语最长名词短语的自动识别. 软件学报, 2000,11(2):195-201.

[2] 孙宏林, 俞士汶. 浅层句法分析方法概述. 当代语言学, 2000,2(2):74-83.

[3] Bourigault D.Surface grammatical analysis for the extraction of terminological noun phrases. In: Boitet C ed. Proceedings of the 15th International Conference on Computational Linguistics (COLING'92). Nantes: Academic Press, 1992. 977～981.

[4] Voutilamen A.NPTool, a detector of English noun phrases. In: Church K ed. Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives. Columbus: Association for Computational Linguistics, 1993. 48～57.

[5] Chen Kuang-hua, Chen Hsin-hsi. Extracting noun phrases from large-scale texts: a hybrid approach and its automatic evaluation. In: Proceedings of the 32nd Annual Meeting of Association of Computational Linguistics. New York: Association for Computational Linguistics, 1994. 234～241.

[6] 李文捷, 周明, 潘海华, 等. 基于语料库的中文最长名词短语的自动提取. 见: 陈力为,袁琦主编 计算语言学进展与应用 北京: 清华大学出版社, 1995,119～124.

[7] 冯冲, 陈肇雄, 黄河燕, 等. 基于条件随机域的复杂最长名词短语识别. 小型微型计算机系统, 2006,6(27): 1134-1139.

[8] 陆俭明. 汉语句法成分特有的套叠现象. 陆俭明自选集. 郑州: 河南教育出版社, 1993,174～192.

[9] John Lafferty, Andrew McCallum, Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]. Proc of ICML, 2001. 282-289.