

苏州大学第一届中文信息学会句法分析评测技术报告

李军辉, 周国栋

苏州大学 自然语言处理实验室 苏州 215006
{lijunhui, gdzhou}@suda.edu.cn

摘要: 本文主要介绍了苏州大学自然语言处理实验室参加第一届中文信息学会句法分析评测的系统。本实验室共参加了 4 个任务评测: 汉语词性标注、汉语基本块分析、汉语功能块分析和句法结构树识别。本文将简要介绍本小组的系统实现以及评测结果, 并针对结果加以适当的分析。

关键字: 词性标注, 基本块, 功能块, 句法分析

Soochow University Report for the 1st China Workshop on Syntactic Parsing

Junhui Li, Guodong Zhou

Natural Language Processing Lab, Soochow University, Suzhou, China 215006
{lijunhui, gdzhou}@suda.edu.cn

Abstract: This paper gives an overview of Soochow Univ NLP lab technical report for the 1st China Workshop on Syntactic Parsing. Our lab participated in four evaluation tasks: Chinese POS tagging, chunking, functional chunking, and syntactic parsing. This paper presents the implement of the four systems as well as the results we have achieved.

KeyWords: POS tagging, basic chunk, functional chunk, syntactic parsing

1 引言

本文将对苏州大学自然语言处理实验室参加 2009 年第一届中文信息学会句法分析评测的系统进行描述。本次评测共包括了 5 项子任务: 汉语词性标注(task1)、汉语基本块分析(task2)、汉语功能块分析(task3)、汉语事件描述单元识别(task4)以及句法结构树识别(task5)。在本次评测中, 本实验室参加除 task4 外的其他 4 项任务, 本文主要介绍各个参评系统的实现以及在各任务上取得的性能。

文章的第二节给出了各个参评系统的实现, 第三节描述了评测中使用的数据、处理过程及讨论。第四节对本文进行总结。

2 参评系统描述

2.1 词性标注(task1)系统描述

词性标注的目标是准确识别句子中各个词语的词类标记。其输入为经过正确切分的句子词语序列。

2.1.1 标注流程

如图 1 所示, 设 $W=w_1 w_2 \dots w_n$ 为包含 n 个单词的句子, 令 $CF=CF_1 CF_2 \dots CF_n$, 其中

$CF_i (1 \leq i \leq n)$ 为第 i 个词的上下文特征向量。因此词性标记的问题，可以视为在给定词序列的情况下，搜寻词性标记序列 $T = t_1, t_2, \dots, t_n$ ，使得 $P(T|CF)$ 值最大。假设各个词进行词性标注是相互独立的，则有：

$$T^* = \arg \max_T P(T|CF) = \prod_{i=1}^n P(t_i|CF_i)$$

输入: $w_1 w_2 w_3 \dots w_n$

输出: $t_1 t_2 t_3 \dots t_n$

图 1 模型输入和输出

在词性标注时，当前词的词性标记需要用到前面已标注词的词性标记信息。为此，我们定义一个大小为 K 的堆 **Heap**，用来保存中间结果。即当对词 w_i 进行词性标注时，堆中保存了 K 个词序列 $w_1 w_2 \dots w_{i-1}$ 的标注结果。当完成对句子中最后一个词词性标注后，返回堆中最优标记序列作为结果。

2.1.2 实验特征

在词性标注过程中，每一个词的标注都被看作是一个事件，因此由当前词及它的上下文环境来确定一个事件的特征集合。根据影响当前词标注的各种因素，我们定义了如下 25 个常用特征，作为基本特征，如表 1 所示：

表 1 词性标注基本特征集合

单项特征 (13 个)	
word(-2), word(-1), word(0), word(1), word(2)	
pos(-2), pos(-1)	
coarse_pos(-2), coarse_pos(-1)	
prefix(word(0)), suffix(word(0))	
num(word(0))	
includeDot(word(0))	
二项组合特征 (6 个)	
pos(-2)&pos(-1) , pos(-2)&word(-2) , pos(-1)&word(-1) , pos(-1)&word(0) , word(0)&word(1), coarse_pos(-2)&coarse_pos(-1)	
$n(n > 2)$ 项组合特征 (6 个)	
pos(-1)&word(-1)&word(0), pos(-2)&pos(-1)&word(0), word(0)&word(1)&word(2), word(-2)&word(-1)&word(0)	
pos(-2)&word(-2)&pos(-1)&word(0), pos(-2)&word(-2)&pos(-1)&word(-1)&word(0)	

表 1 中使用到的模板函数定义如下：

word(i)：窗口 i 的词

pos(i)：窗口 i 的词性标记

coarse_pos(i)：窗口 i 的大类词性标记，例如，词性 $n, n0, nP, nR, nS, nT$ 的大类都为 n

prefix(word)：词 word 的第一个字

suffix(word)：词 word 的最后一个字

num(word)：词 word 包含的字数

includeDot(word)：词 word 中是否包含有 '·'，这通常出现于外译姓名中

除表 1 定义的基本特征外，我们还使用了以下两类特征：

后邻接词词性特征

从表 1 制定的特征中可以发现，在按从左至右的顺序分别为句子中每个词进行词性标注时，可以获得当前词左侧词的词性标注信息；但对当前词的右侧词，只能获取其词信息。为了获取后邻接词的词性信息，我们定义了如下的特征：后邻接词的任意可能词性（即在训练集中曾出现的词性），以及其与当前词的组合特征。

低频词特征

如上述分析，由于在特征选择中，设置阈值会使得低频词的一些有用特征被舍弃，这也是低频词在测试集中准确率较低的一个原因。为了能够保留低频词在训练集中的词性信息，同时不给模型带来噪音特征，我们在表 1 制定的基本特征基础上，添加以下特征，称之为低频词特征：

- a) 如果当前词在训练集中出现次数为 1，并设其词性为“X”，则添加特征“aX”。例如词“手感”在训练集中出现 1 次，并且其词性为名词(n)，则在抽取特征时，如果当前词为“手感”，则添加特征“an”；¹
- b) 如果当前词在训练集中出现次数为 2、3，则分别针对其出现的词性，添加特征“b%d%s”，其中%d 指出现次数，%s 为词性。例如，词“编译”在训练集中出现次数为 3 次，其中 1 次作为动词(v)，2 次作为名词(n)，则在抽取特征时，如果当前词为“编译”，则添加特征“b1v”和“b2n”；

2.2 基本块识别(task2)系统描述

基本块主要描述句子中直接相邻的、以名词、动词、形容词等实义词为中心聚合形成具有特定语义内容的词语序列，其中一般不包括各种功能词，例如连词、语气词、标点符号等。

2.2.1 标注流程

如前所述，功能词通常并不包含在基本块内，为了处理方便，我们为各种功能词分别制定了功能词-基本块，例如连词-基本块，标点符号基本块等。如下所示，“的/uJDE”、“对/p”和“./wE”分别单独构成基本块 buJDE、bP 和 bwE。

处理前：[vp 出席/v 晚会/n] 的/uJDE [np 各界/n 人士/n] 对/p [np 此/rN] [vp 感到/v 欢欣鼓舞/v] 。/wE
处理后：[vp 出席/v 晚会/n] [buJDE 的/uJDE] [np 各界/n 人士/n] [bP 对/p] [np 此/rN] [vp 感到/v 欢欣鼓舞/v] [bwE 。/wE]

类似于 Kudo & Matsumoto (2001)的做法，通过引进 4 个边界符号：B, I, E, S，可以把基本块识别问题转化为标注问题，其中 B-X 表示基本块 X 的开始；I-X 为基本块 X 的内部；E-X 为基本块 X 的外部；S-X 指单独构成基本块 X。例如在上面例子中，为能够正确识别出“各界/n 人士/n”为名词基本块 np，单词“各界/n”和“人士/n”的标记必须分别为 B-np 和 E-np。

基本块的标注流程类似于词性标注，在标注过程中，当前词的基本块标注需要用到前面已标注词的基本块信息。为此，我们定义一个大小为 K 的堆 Heap，用来保存中间结果。即当对词 w_i/p_i 进行基本块标注时，堆中保存了 K 个词序列 $w_1/p_1 w_2/p_2 \dots w_{i-1}/p_{i-1}$ 的标注结果，并且任意两个相邻的标记之间不会产生冲突（例如，相邻基本块标记 B-np 和 S-vp 产生冲突）。当完成对句子中最后一个词的基本块标注后，返回堆中最优标记序列作为结果。

2.2.2 实验特征

¹我们也尝试过按 b)方式处理在训练集中出现次数为 1 的词，但效果并不理想。我们认为：出现次数为 1 的词，词性具有较强的偶然性，应该与其它低频词区别对待。

根据影响当前词基本块标注的各种因素，在定义特征空间时，将考虑的信息包括：词信息、词性信息和基本块标记信息（只限于当前词前面词的基本块信息）。参照 Ratnaparkhi (1999)的做法，本系统所使用的特征如表 2 所示：

表 2 基本块识别特征集合

单项特征 (10 个)			
chunkandpostag(0),	chunkandpostag(0*),	chunkandpostag(1),	chunkandpostag(1*),
chunkandpostag(2),	chunkandpostag(2*),	chunkandpostag(-1),	chunkandpostag(-1*),
chunkandpostag(-2),	chunkandpostag(-2*)		
二项组合特征 (8 个)			
chunkandpostag(0, 1),	chunkandpostag(0*, 1),	chunkandpostag(0, 1*),	
chunkandpostag(0*, 1*),	chunkandpostag(-1, 0),	chunkandpostag(-1*, 0),	
chunkandpostag(-1, 0*),	chunkandpostag(-1*, 0*)		
n(n>2)项组合特征 (4 个)			
chunkandpostag(0, 1*, 2*),	chunkandpostag(0*, 1*, 2*)		
chunkandpostag(0, -1*, 1*),	chunkandpostag(0*, -1*, 1*)		

表 2 中使用到的模板函数定义如下：

chunkandpostag(i)：word(i)，即窗口 i 的词

chunkandpostag(i*)：当 i<0 时，指窗口 i 的词性标记和基本块标记；当 i>=0 时，指窗口 i 的词性标记。

2.3 功能块识别(task3)系统描述

汉语功能块主要描述句子中反映不同事件内容的基本信息单元。它们一般担当句子中的主语、谓语、宾语、状语、定语、中心语等位置。该任务的目标是识别出句子中不同层次的功能块，覆盖自顶向下进行事件句式拆分而形成的各个基本信息单元，形成进行进一步的事件骨架树分析的最小功能块描述序列。该任务的输入为经过正确词语切分和词性标注处理的汉语句子，输出为不同层次的功能块组合形成的线性序列。

2.3.1 标注流程

通过对 task2 和 task3 的数据分析发现，功能块单元通常位于基本块的上层，即功能块通常由若干个基本块构成²，例如在下面的句子中，基本块 “[vp 出席/v 晚会/n]” 正好单独构成述语块 P，介词 “对/p” 和基本块 “[np 此/rN]” 共同构成状语块 D。同时，仍然存在着部分功能词（例如标点符号、连词等）并不包含于任何功能块。因此，我们常用与基本块处理相同的办法，人为地添加功能块标记，使得每个单词都属于某个功能块。如下所示：

处理前：[P 出席/v 晚会/n]的/uJDE [H 各界/n 人士/n][D 对/p 此/rN][P 感到/v 欢欣鼓舞/v]。/wE
 处理后：[P [vp 出席/v 晚会/n]] [FujDE [bujDE 的/uJDE]] [H [np 各界/n 人士/n]] [D [bp 对/p] [np 此/rN]] [P [vp 感到/v 欢欣鼓舞/v]] [FbwE [bwE 。 /wE]]

在确保每个功能块都是由若干个基本块和功能词-基本块构成后，可以简单地把基本块的识别看作为基于基本块识别结果的 Chunking。为此，与基本块识别的做法类似，通过引进 4 个边界符号：B, I, E, S，可以把功能块识别问题转化为标注问题，其中 B-X 表示功能块 X 的开始；I-X 为功能块 X 的内部；E-X 为功能块 X 的外部；S-X 指单独构成功能块 X。

² 在准备训练数据时，若某个功能块并不是由基本块和功能词-基本块构成，则跳过该句

2.3.2 实验特征

根据影响当前基本块的功能块标注的各种因素，在定义特征空间时，将考虑的信息包括：基本块中心词及其词性信息、基本块标记信息、功能块标记信息（只限于当前词前面词的功能块信息）、已识别的功能块信息等。本系统所使用的特征如表 3 所示：

表 3 功能块识别特征集合			
单项特征 (17 个)			
word(0), word(1), word(2), word(-1), word(-2)			
pos(0), pos(1), pos(2), pos(-1), pos(-2)			
chunk(0), chunk(1), chunk(2), chunk(-1), chunk(-2)			
funChunkTag(-1), funChunk(-1)			
二项组合特征 (12 个)			
chunk(0)&chunk(1),	chunk(0)&pos(1),	chunk(0)&pos(0),	pos(0)&chunk(1),
word(0)&chunk(1),	chunk(0)&chunk(-1),	chunk(0)&pos(-1),	chunk(0)&pos(0),
pos(0)&chunk(-1),	word(0)&chunk(-1),	funChunkTag(-1)&funChunkTag(-2),	funChunk(-1)&funChunk(-2)
$n(n>2)$ 项组合特征 (4 个)			
chunk(0)&chunk(1)&chunk(2), word(0)&chunk(1)&chunk(2),			
chunk(0)&chunk(1)&chunk(2)&chunk(3), chunk(1)&chunk(2)&chunk(3)			

表 3 中使用到的模板函数定义如下：

word(i): 窗口 i 的基本块的中心词

pos(i): 窗口 i 的基本块中心词的词性

chunk(i): 窗口 i 的基本块类别

funChunkTag(i): 窗口 i 的基本块的功能块标记

funChunk(i): 当前基本块左侧已形成的第 i 个功能块类型

2.4 句法树结构识别系统描述

2.4.1 标注流程

本系统使用的层次句法分析模型与其他基于移进/归约序列的句法分析模型类似，都是通过预测“动作”来逐步构建句法树。如果一个连续的动作序列 A 能够构建句法树 T ，则称动作序列 A 为句法树 T 的推导。值得注意的是，通过使用“动作序列”来表示句法树，看不到 PCFG 模型中经常提到的各类语法规则，句法树的分值也被分解为推导中各个动作的分值。

在自底向上的构建句法树过程中，根据已有的动作序列 $\{a_1, \dots, a_N\}$ ，预测下一个可能的动作 a_{N+1} ，产生一个新的动作序列 $\{a_1, \dots, a_N, a_{N+1}\}$ ，并且任意一棵句法树都有一个确切的推导。整个句法分析的过程可分解为三个过程：词性标注(POS Tagging)、基本组块识别(Basic Chunking)和复杂组块识别(Parsing)。对某给定的句子，需分别执行词性标注和基本组块识别各一次，紧接着，循环执行复杂组块识别过程直至识别出根结点。以下是各过程的描述，其中词性标注和基本组块识别过程的更详细描述可参考 Ratnaparkhi (1999)。

词性标注：对输入的句子词串 $S=(word_1, word_2, \dots, word_n)$ ，从左至右分别预测每个词的词性，输出词性标注结果 $S=(word_1/pos_1, word_2/pos_2, \dots, word_n/pos_n)$ 。因此，本过程中的动作类别集合为词性标记集合。

基本组块识别：基本组块指的是子结点均为词性结点的组块。基本组块的识别以词性标注结果为输入。从左至右，为每个单词/词性标记对赋予基本组块识别标记。基本组块识别标记类别包括 **Start_X**, **Joint_X** 和 **Other** 三类，其中 **X** 为任意的组块类别。基本组块识别标记被用于基本组块的检测，如果 $\text{word}_m/\text{pos}_m$ 被标注为 **Start_X**，并且 $\text{word}_{m+1}/\text{pos}_{m+1}, \dots, \text{word}_{m+i}/\text{pos}_{m+i}$ 均被标记为 **Joint_X**，则序列 $\{\text{word}_m/\text{pos}_m, \dots, \text{word}_{m+i}/\text{pos}_{m+i}\}$ 被组合成一个基本组块 **X**。

复杂组块识别：与基本组块不同的是，复杂组块指的是至少有一个子结点不是词性结点的组块。从左到右，分别为每个单元（此单元即可以是基本组块、也可以是复杂组块或词性结点）赋予复杂组块识别标记。复杂组块类别包括 **Begin_X**, **Middle_X**, **End_X**, **Single_X** 和 **Other** 共五类，其中 **X** 为任意的组块类别。复杂组块识别标记被用于复杂组块的检测，如果连续的单元被标注为 **Begin_X**, **Middle_X**, ..., **Middle_X**, **End_X**，则此连续单元被组合为一个复杂组块 **X**；如果某单元被标注为 **Single_X**，则此单元单独构成复杂组块 **X**。

不难分析，假设一个包含 n 个词的句法树对应的推导为 $\{a_1, \dots, a_N\}$ ，则动作序列 $\{a_1, \dots, a_n\}$ 为词性标记序列， $\{a_{n+1}, \dots, a_n\}$ 为基本组块识别标记序列， $\{a_{2n+1}, \dots, a_N\}$ 为复杂组块识别标记序列。

在本任务中，由于输入的对象为经过正确切分和词性标注处理的事件描述单元，因此，本系统只需要针对输入的句子进行基本组块识别和复杂组块的识别。

2.4.2 实验特征

如前分析，本文提出的句法分析器在预测下一个动作时，是以已有的动作序列为依据的。使用条件概率 $PX(a|b)$ 来表示在已有动作序列为 b 的前提下，下一个动作为 a 的概率。然而，动作序列 b 所构成的上下文包含了大量的信息，参考 Ratnaparkhi(1999) 制定的特征和根据多次实验结果，最终分别为基本组块识别和复杂组块识别制定了表 4 和表 5 所示的特征模板：

表 4. 基本组块识别过程使用的特征集合

<p>单项特征 (8 个)</p> <p>$\text{word}(-2), \text{word}(-1), \text{word}(0), \text{word}(1), \text{word}(2),$ $\text{pos}(0), \text{pos}(1), \text{pos}(2)$</p>
<p>二项组合特征 (8 个)</p> <p>$\text{word}(-1)\&\text{word}(0), \text{word}(0)\&\text{word}(1), \text{action}(-2)\&\text{pos}(-2), \text{action}(-1)\&\text{pos}(-1),$ $\text{pos}(0)\&\text{word}(1), \text{word}(-1)\&\text{pos}(0), \text{word}(0)\&\text{pos}(1), \text{pos}(0)\&\text{pos}(1)$</p>
<p>$n(n>2)$ 项组合特征 (2 个)</p> <p>$\text{action}(-1)\&\text{pos}(-1)\&\text{word}(0), \text{action}(-1)\&\text{word}(-1)\&\text{pos}(0)$</p>

表 5. 复杂组块识别过程使用的特征模板

<p>单项特征 (18 个)</p> <p>$\text{cons}(i), \text{cons}(i^*), \text{cons}(i^{**}), -2 \leq i \leq 3$</p>
<p>二项组合特征 (8 个)</p> <p>$\text{cons}(i, i+1), \text{cons}(i^*, i+1), \text{cons}(i, i+1^*), \text{cons}(i^*, i+1^*), i=-1, 0$</p>
<p>$n(n>2)$ 项组合特征 (6 个)</p> <p>$\text{cons}(0, 1^*, 2^*), \text{cons}(0, 1, 2^*), \text{cons}(0, 1^*, 2), \text{cons}(0^*, 1^*, 2^*),$ $\text{cons}(0, 1^*, 2^*, 3^*), \text{cons}(0^*, 1^*, 2^*, 3^*)$</p>

表 4 和表 5 中使用到的模板函数定义如下：

- $\text{word}(i)$: 窗口 i 的词
- $\text{pos}(i)$: 窗口 i 的词性标记
- $\text{action}(i)$: 窗口 i 的基本组块识别标记
- $\text{cons}(i)$: 窗口 i 组块的中心词
- $\text{cons}(i^*)$: 窗口 i 复杂组块识别标记+组块的类别，若 $i \geq 0$ ，则省略复杂组块识别标记
- $\text{cons}(i^{**})$: 窗口 i 复杂组块识别标记+组块的类别+组块中心词词性，若 $i \geq 0$ ，则省略复杂

组块识别标记

2.4.3 中心成分的确

在上述句法分析模型中，需要为每个组块确定其中心成分，并且每个组块有且只能有一个中心成分。为此，我们需要重新为训练集中各个组块标记其中心成分，并且在测试过程中，对每个新产生的组块结点，也需要为其标记中心成分。在此，我们采用基于规则和统计的方法为每个中心词确定规则，具体地：

1). 如果句法规则没有出现在训练集中

类似于 Collins(1999)的做法，我们采用表 6 制定的规则获取中心成分，其中 0 表示从右至左，1 表示从左至右

表 6 中心成分提取规则

父结点	中心成分规则
np	0n nP nR nS nT vN aN 0
vp	1v vC vJY vM vSB vB 1
ap	0a b 0
bp	0b 0
mp	0qN qV qC 0
sp	0f rS fS 0
tp	0t nT qT fT 0
dp	0d dN 0
dj	0vp dj 0ap 0
dlc	0
fj	0vp dj fj 0ap 0
mbar	0m 0
jp	0
yj	0
pp	1p 0pp 1
zj	0vp dj fj zj 0ap 0

2). 如果句法规则出现在训练集中

如果规则 $P \rightarrow C_1 C_2 \dots C_n$ 在训练集中出现过，则选择在训练集中，作为此条规则中心成分次数最多的子结点为中心成分，若这样的子结点有两个或更多，选择最左边的子结点为此规则的中心成分。

3 实验及结果分析

3.1 实验设置

本文的实验数据均采用组办方提供的评测数据，未使用其他数据资源。在实验过程，我们常用了 OPENNLP 最大熵工具包³和 SVMlight 分类器⁴。对每个分类任务，采用的特征阈值为 3，即过滤出现次数为 1 或 2 的特征。在使用最大熵分类器训练时，迭代次数为 100 次，而在使用 SVMlight 训练时，c 值的设置为其默认值的 4 倍。由于 SVMlight 是一个二元分类器，我们对其采用 one Vs. others 的办法，将其封装为多类分类器。具体地，在词性标注、基本块和功能块识别任务中，我们采用

³ The OpenNLP Maximum Entropy Package. <http://maxent.sourceforge.net/>

⁴ SVM-Light Support Vector Machine. <http://svmlight.joachims.org/>

的是 SVMlight 分类器，而在句法树结构识别任务中，由于训练数据较大，我们选择了训练速度较快的最大熵分类器。

3.2 实验结果

3.2.1 词性标注评测结果

表 7. 词性标注评测结果

	Overall F1	Small-class average F1
基本特征	92.98	81.13
+后邻接词词性特征	93.27	81.16
+低频词特征	93.15	81.00
所有特征	93.41	81.13

词性标注评测结果如表 7 所示。从实验结果中可以看出，后邻接词词性特征和低频词特征都在一定程度上提高了系统的性能。我们提交的最终结果为同时集成了后邻接词词性特征和低频词特征所取得的结果。

3.2.2 基本块和功能块识别评测结果

表 8 给出了基本块和功能块的评测结果。

表 8. 基本块和功能块评测结果

	Recall	Precision	F1
基本块评测结果 (boundary + type)	92.41	92.32	92.36
功能块评测结果 (基于正确基本块)	88.65	89.47	89.06
功能块评测结果 (基于自动基本块)	84.78	84.67	84.72

功能块描述的是句子中反映不同事件内容的信息，在结构上要复杂于基本块，同时，功能块通常由若干个连续的基本块构成。因此，功能块的识别较基本块识别更加困难。表 8 同时也反映了基本块的自动识别对功能块识别的性能影响。

3.2.3 句法树结构识别评测结果

句法分析评测结果如表 9 所示。

表 9. 句法树结构识别评测结果

Recall	Precision	F1
85.07	84.93	85.0

4 总结

本文对苏州大学自然语言处理实验室参加第一届中文信息学会句法分析评测的系统进行了描述。本次评测中，本实验室参加的任务包括：词性标注、基本块识别、功能块识别和句法树结构识别共四项。我们将各个任务分别转化为分类任务，并根据各任务的特点分别制定有效的特征集。具体地，在词性标注任务中，我们发现右邻接词的词性信息和低频词的词性信息能够帮助改善词性标注性能；在功能块识别任务中，是基于基本块识别基础上进行的，我们将在下步分析是否有必要这样做；在句法树结构识别任务中，我们发现中心成分的识别对系统的性能影响较大，对此，我们还

有很大的提升空间，尤其是融合开源的句法分析器（如 Berkeley Parser, Chiarniak Parser）等。此外，还需要向国内外同行学习，不断改善现有的系统。

参考文献

- [1] Eugene Charniak. 2000. A maximum-entropy-inspired parser. In Proceedings of NAACL 2000.
- [2] Michael John Collins. 1999. Head-driven statistical models for natural language parsing. Ph.D. thesis. University of Pennsylvania.
- [3] Taku Kudo and Yuji Matsumoto. 2001. Chunking with Support Vector Machines. In Proceedings of NAACL 2001.
- [4] Li Junhui, Zhou Guodong, Zhu Qiaoming, and Qian Peide. Syntactic Parsing with Hierarchical Modeling. In Proceedings of AIRS 2008.
- [5] Salvo Petrov and Dan Klein. 2007. Improved inference for unlexicalized parsing. In Proceedings of NAACL 2007.
- [6] Adwait Ratnaparkhi. 1999. Learning to parse natural language with maximum entropy models. *Machine Learning*, 34(1-3):151-175.
- [7] Adwait Ratnaparkhi. 1996. A maximum entropy model of part-of-speech tagging. In Proceedings of EMNLP 1996.
- [8] 周雅倩, 郭以昆, 黄萱菁, 吴立德. 2003. 基于最大熵方法的中英文基本名词短语识别. *计算机研究与发展*. Vol. 40, NO. 3.
- [9] 李军辉, 周国栋, 朱巧明, 钱培德. 2009. 一种改进的中文层次句法分析模型研究. 第十届全国计算语言学学术会议. 2009.