文章编号：1003-0077（2017）

由粗到精的哈萨克语短语结构句法分析研究

梁金莲[1,2,3]　古丽拉·阿东别克[1,2,3]

（1.新疆大学 信息科学与工程学院，新疆 乌鲁木齐市 830046；2.新疆多语种信息技术实验室；3.国家语言资源监测与研究少数民族语言中心哈萨克和柯尔克孜语文基地）

摘要：本文针对哈萨克语短语结构句法分两个阶段采用由粗到精的方法进行哈萨克语句法分析研究。第一阶段使用粗略的句法分析器生成 20 个最佳候选树；第二阶段采用感知机的方法训练，提取特征信息，并对第一阶段生成的 20 个最佳候选树进行重排序，最终解析结果是第一阶段产生的候选树的结果和重排序结果按照比例选取。此方法在两个阶段不仅可以获取到句子的结构信息，还可以提取到详细的特征信息，可以最大限度的对句子进行解析，获得了较好的句子解析结果，其句法分析的正确率为 71.4%。

关键词：句法分析，PCFG，重排序；

中图分类号：TP391　　**文献标识码：**A

A Kazakh Syntactic Structure Parsing from Coarse to fine

LIANG Jinlian[1,2,3],Gulila ALTENBEK[1,2,3]

(1. College of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang 830046, P.R. China ;
2. Xinjiang Laboratory of Multi-Language Information Technology ; 3. The Base of Kazakh and Kirghiz Language of National Language Resource Monitoring and Research Center on Minority Languages)

Abstract : In this paper, the syntactic analysis of the Kazakh phrase structure is used in the method of coarse to fine by two stages. The first stage uses the rough parser to generate 20 best analytical sets. In the second stage, the perceptron method is used to train and extract the feature information, and the 20 best analyzes of the first stage are reranking, and the best result after reranking is selected as the final result. This method can not only obtain the structural information of the sentence in two stages, but also extract the detailed feature information, which can analyze the sentence to the maximum and obtain the better sentence analysis result, the result of syntactic analysis is 71.4%

Key words: Syntactic analysis； PCFG； reranking

1　引言

自然语言处理的过程一般分为词性分析、句法分析和语义分析。句法分析是自然语言处理中关键的环节[1]。目前句法分析主要的方法包括基于规则的方法和基于统计的方法。基于统计的句法分析在处理歧义等方面有较好的效果，相对也

收稿收稿日期：　　；**定稿日期**：

基金项目：国家自然科学基金（61363062），其他（NMLR201601）

作者简介：梁金莲（1990-），女，硕士，研究方向为自然语言处理;古丽拉·阿东别克（1962-），女，教授，博士生导师，主要研究方向为自然语言处理，人工智能。

比较灵活。目前基于统计的方法成为主流[2]。

目前哈萨克语的研究已经进行到句法分析阶段。完成了词法分析[3]的研究。其中，文献[4]中提出一种规则与最大熵结合的方法对哈萨克语基本动词短语进行识别：文献[5]提出了一种基于条件随机场模型的哈萨克语的基本短语自动识别方法：文献[6]采用基于规则自动识别及人工标注的方法建立基本名词短语标注语料库。在句法分析阶段也进行了相关的研究。文献[3] 中采用 PCFG 方法，结合自底向下的 Vitrtbi 算法实现一种有自学习能力的哈萨克语句法分析器：文献[7]中根据概率上下文无关文法模型和 Chart 算法特点，将概率引入 Chart 算法，提出一种 PChart 算法，实现一种基于 PChart 算法的哈萨克语句法分析器。 在句法分析研究中，无论是基于统计的方法还是基于规则的方法，都不能完全解决句法分析的问题，将两者结合起来，才有可能最大限度的解决句法分析存在的问题。概率上下文无关文法（Probabilistic Context Free Grammars, 简称 PCFG），是统计与规则相结合的方法。自然语言是一种上下文有关文法，在用上下文无关文法对自然语言进行处理过程中，必然会产生歧义。PCFG 只能捕捉到句子的结构和规则，不能捕捉到上下文的信息，因此，对语言的描述是粗粒度的。该文提出一种 PCFG[8]与感知机相结合的方法进行句法分析。感知机在进行训练过程中，可以捕获到句子的上下文信息。利用感知机捕获到的信息，对 PCFG 产生的解析候选树进行重排序，进一步提高哈萨克语句法分析的效果，进而弥补 PCFG 的不足之处。该文提出的方法分为两个阶段，在第一阶段，采用 PCFG 方法，对输入的每个待解析的句子，粗略的产生 20 个概率最高的句子候选集，由于句子长度的差异，有些句子的最佳候选集长度小于 20。在感知机训练过程中，将训练得到的参数，以及提取特征得到的特征模板，对第一阶段生成的 20 个最佳候选集进行重排序。将重排序的结果和 PCFG 得到的结果按照一定比例选取，得到解析结果作为重排序最终的解析结果。实验表明，使用 PCFG 和感知机相结合的方法，可以得到比较理想的句法分析的效果。

2 PCFG 和感知机

由于 PCFG 的句法分析不能捕捉到句子的上下文信息，在消歧能力方面有限，感知机可以通过自学习，能捕捉到句子中细粒度的信息。可以弥补 PCFG 的不足。因此，该文采用 PCFG 和感知机相结合的方法，对哈萨克语进行句法分析。

2.1 PCFG 模型

PCFG 模型是句法分析中研究比较广泛和充分的模型之一。它是一种统计和规则相结合的方法。CFG 是获取语言中的句法规则，由非终结符、词汇表、开始字符以及规则的产生式集合构成[9]。PCFG 则是在此规则中增加了概率参数，通过计算概率，预测可能性最大的句法结构。

对于一个输入句子 S，通过统计的方法得解析树。其得到最优解的 Tbest 的如公式 1 所示：

$$T_{best} = \arg\max_{T} P(T|S) = \arg\max_{T} \frac{P(T,S)}{P(S)} = \arg\max_{T} P(T,S) \quad (1)$$

其中 T 表示候选树， P(T, S）是有候选树 T 中所有规则的概率的乘积。如公式 2 所示：

$$P(T,S) = \prod_{i-1}^{n} P(RHS_i|LHS_i) \quad (2)$$

通过计算句子 S 中所有可能的 T 中的概率 P(T, S)，选出概率最大的值。在计算概率时，需给定三个假设：祖先无关性假设、位置不变性假设以及上下文无关性假设。

2.2 感知机算法

神经网络由一个或者多个神经元组成。而一个神经元包括输入、输出和“内部处理器”。神经元从输入端接收信息，通过“内部处理器”将这些信息进行一定的处理，最后通过输出端输出。单层感知器（Single Layer Perceptron）是最简单的神经网络。它包含输入层和输出层，而输入层和输出层是直接相连的。

康奈尔大学教授 Frank Rosenblatt 1957 年提出“感知机（Perceptron)”，它是第一个用算法来精确定义神经网络，第一个具有自组织学习能力的数学模型，是日前许多新的神经网络模型的始祖。

单层感知机训练：

第一步：函数输出数量相等的感知机会以小的初始值开始。

第二步：选取训练集中的一个例子作为输入，计算感知机的输出。

第三步：对于每一个感知机，如果其结果和该例子的结果不匹配，调整初始值

第四步：继续采用训练集中的例子，重复输入，进行匹配，调整参数。

在该文中，重排序的训练过程以及句法分析解码阶段都采用感知机算法。感知机算法[10, 11]如

表 1 所示：

表 1 感知机算法

Inputs: Training examples (xi; yi)

Initialization: Set $\bar{\alpha} = 0$

Output: Parameters $\bar{\alpha}$

Algorithm:

For $t = 1 \ldots T, i = 1 \ldots n$

Calculate $z_i = \arg\max_{z \in GEN(x_i)} \Phi(x_i, z) \cdot \bar{\alpha}$

if $(z_i \neq y_i)$ then $\bar{\alpha} = \bar{\alpha} + \Phi(x_i, y_i) - \Phi(x_i, z_i)$

其中，$\overline{GEN}(X_i) = GEN(X_i) - \{y_i\}$ 中$\{\boldsymbol{y_i}\}$表示标准解析树，$GEN(X_i)$表示句子的候选集合，$\overline{\mathrm{GEN}}(\mathrm{x_i})$ 表示所有候选解析中的错误解析。

3 重排序

本文的句法分析主要分为两个阶段，第一阶段，采用 PCFG 的方法，将待分析的每个句子产生 20 个概率最高的候选解析列表，第二阶段，使用感知机重排序的方法，对第一阶段产生的 20 个概率最高的候选解析序列进行重排序，将两者得分按照比例相加，选出得分最高的候选树，作为句法分析最终的结果。该文中的句法分析流程如图 1 所示：

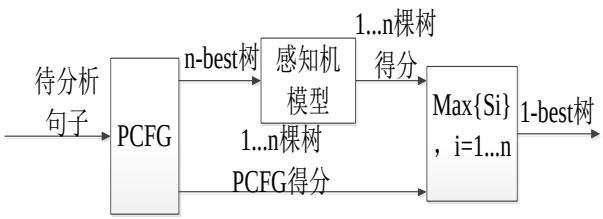

图 1 句法分析流程图

文中的待分析句子指的是要进行句法分析的句子，在 PCFG 解析阶段，待分析的句子作为输入，经过解析，每句待分析的句子将产生 20 个最佳句子候选集，以及各自的句子的得分。有些句子较简 单，产生的句子少于 20 个。将产生的 20 个候选句子作为感知机模型的输入，感知机模型根据训练得到的特征模板和参数$\bar{\alpha}$计算候选树中各结点的分数，从而达到给候选树重排序的目的。最终候选树的得分，根据 PCFG 的得分，以及感知机对候选树重新计算结点之后的得分按照一定比例相加，得分最高的候选树作为句法分析最终的解析结果。该文中使用两者结合的重排序进行句法分析比使用单一的 PCFG 进行句法分析的结果要好。

3.1 哈萨克语的 20 个最佳解析

在这个阶段，对于每个输入的字符串 S，采用 n-best 解析算法[8]，返回 n 个最高概率的解析 $Y(s) = \{y_1(s), \ldots, y_n(s)\}$，以及根据解析器产生概率模型的每个解析 y 的概率 P(y)。本文的实验数量 n 的分析是为 20。但是有些简单的句子，实际上得到的解析列表集少于 20 个。

目前哈萨克语的有 10 种词性标注和 5 中短语标注集[7]。如表 2 所示：

表 2 哈萨克语标注集

标识	说明	标识符	说明
n	名词	ono	模拟词
num	数词	conj	连词
int	感叹词	S	句子
v	动词	NP	名词短语
adj	形容词	VP	动词短语
adv	副词	AdjP	形容词短语
pron	代词	NumP	数词短语
part	助词	AdvP	副词短语

对于给定的哈萨克语的句子 S，通过 PCFG 得 20 个候选集，例如，输入句子形如：

ءبىز ۇرىمجىگە ۇشاقپەن ۇشىپ كەلدى .

则该句子其中一个候选树如图 2 所示：

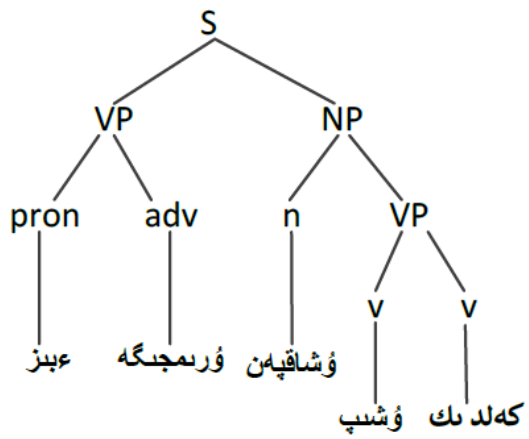

图 2 哈萨克语句法树

3.2 训练

在重排序阶段中，首先使用感知机算法对语料进行训练，再将 PCFG 阶段每个句子产生的 20 个候选解析树进行解码，解码的过程就是对每个结点进行重新算分，重新排序。

在用感知计算法进行训练的过程，对输入 x

∈χ 有一个映射 y∈y ，在句法分析中 χ 是一个未处理的句子集合，y 是一个 χ 句子中的标准句法树的集合。下面首先给出如下四个假设：

假设一：训练样本（xi,yi） i = 1...n

假设二：定义函数 GEN，GEN(x)是列举了输入 x 所有可能的句法树集合

假设三：Φ 表示每一组(x,y)∈χ ×y →特征向量 Φ(x,y)∈Rd 的映射一个参数向量 $\bar{\alpha}$∈Rd

假设四：一个参数向量$\bar{\alpha}$∈Rd

以上四个假设内容 GEN,Φ ,$\bar{\alpha}$定义了一个输入 x→输出 F(x)的映射：

如公式 3 所示：s

$$f(x) = \arg\max_{y \in GEN(x)} \Phi(x,y) \cdot \bar{\alpha} \quad (3)$$

其中 $\Phi(x,y)\cdot\bar{\alpha}$可以表示为$\sum_s \alpha_s \Phi_s(x,y)$，$\bar{\alpha}$是待训练参数，初始值为 0，Φ(x,y)是未处理句子的候选树的映射。$\bar{\alpha}$ 是在训练语料中，通过比较标准树和训练得到的候选树得分，进行参数调整得到的。

在感知机训练中需要特征模板，特征模板[8]如表 3 所示：

表 3 感知机训练特征模板

N-grams	features
Unigrams	$W_0,T_f,W_f,W_c,T_c,W_b,T_b$.
Bigrams	$W_0T_f,W_0W_f,W_0W_c,W_0T_c,T_fW_f,T_fW_b,$ $T_fT_b,W_fW_b,W_fT_b,W_0W_b,W_0T_b$.
Trigrams	$W_0T_fW_f,W_0W_cT_c,T_fW_bT_b,W_fW_bT_b,$ $W_bT_fT_b,T_bW_fT_f,W_bT_fW_f,T_fW_cT_c,$ $W_fW_cT_c$.

其中，W_0 表示当前结点的词，T_f 表示父结点的词性，W_f 表示父结点的词，W_c 表示子结点的词，T_c 表示子结点的词性，W_b 表示兄弟结点的词，T_b 表示兄弟结点的词性。

在感知机训练过程中，对于输入的训练句子，都有相对应的标注好的句子。首先，感知机对于输入的训练句子进行解析，产生所有可能的句子结构，对产生的每个候选句子的结点进行计算，并将每个结点计算的得分相乘，得到整个候选树的总的得分，将候选树的总得分与标准句子进行比较，当候选树的分数和标准的句法树分数一致，参数$\bar{\alpha}$不需要调整，停止训练；或者迭代到一定的次数，训练停止，得到参数$\bar{\alpha}$，训练过程是对参数$\bar{\alpha}$进行不断的调整。重新输入下一个句子，下一个句子的初始参数的值则使用上一句的参数作为初始值，直到将训练语料中所有的句子训练完成，得到一个$\bar{\alpha}$和特征[12]模板。将得到的参数和模板，用于重排序过程当中。

3.3 重排序

在 PCFG 阶段，对于给定一个句子 S，可以产生 20 个最佳解析的候选集，将这 20 个候选树作为感知机重排序的输入。在重排序阶段，感知机重新计算每个父结点的分数，最后将每个父结点的分数相乘，得到每个候选树的得分[13]。最终的句法分析的结果，以 PCFG 的结果和重排序的结果按照一定比例，选出最佳的句法分析。图 3 显示了一个候选树计算结点的例子。

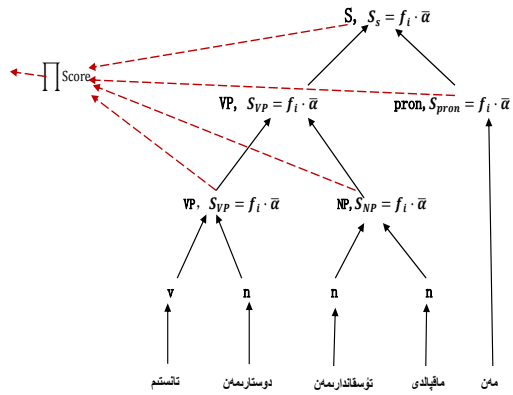

图 3 感知机计算结点示意图

计算父结点得分如公式 4 所示：

$$S_{(p)} = f_i \cdot \bar{\alpha} \quad (4)$$

其中，$\bar{\alpha}$是感知机在训练语料得到的参数，f_i 表示特征向量。在计算整个候选树的时候，是对每个父结点的分数相乘，得到整个候选树的得分。如公式 5 所示：

$$S = \sum S_{(p)} \quad (5)$$

最终的评分参考 PCFG 和感知机重排序之后的两者的得分，按照一定的比例求和如公式 6 所示：

$$S = S_{PCFG} + t \cdot S_P \quad (6)$$

其中S_{PCFG}是 PCFG 的得分，S_P是感知机重排序之后的得分，t 是权重系数。

4 实验

本文的数据来源于新疆中小学哈萨克语文课文。这些原始语料以短语形式标注过，例如名词短语标注为 NP，动词短语标注为 VP 等，本语料中哈萨克语的短语标注分为 5 类，分别为名词短语、动词短语、形容词短语、数词短语和副词短语。

4.1 实验设置

使用的原始语料以及语料中的句法树的形式，如下所示. 原始语料：

ۇرىمجىدە ماڤپالدىكّ ۇيىندە ون كۇندەي دەمالدىق

句法树形式：

(S

(NP (adv ۇرد مجدە))

(NP (N1 ماڤ پالددك) (adv ۋيە ندە)))

(VP (numP (num ون) (n كۇندەي)) (v دەمالددق))

(. .))

评价指标采用常用的 PARSEVAL 评价体系评测句法分析，标记正确率 (Labeled Precision, LP)、标记召回率(Labeled Recall, LR)、F1 值作为评价标准。PARSEVAL 是应用比较广泛的短语结构分析器性能评价方法，标记正确率是句法分析结果中正确短语数占结果中短语总数的比例，如公式 7 所示；标记召回率是句法分析结果中正确短语数占标准树中短语总数的比例，如公式 8 所示;F1 值是综合正确率和召回率两者的综合指标，计算公式如公式 9 所示。

$$LP = \frac{\text{分析得到的正确短语数}}{\text{分析得到的短语数}} \times 100 \quad (7)$$

$$\mathrm{LR} = \frac{\text{分析得到的正确短语数}}{\text{标准树的短语数}} \times 100\% \quad (8)$$

$$\mathrm{F1} = \frac{2 \times \mathrm{LP} \times \mathrm{LR}}{\mathrm{LP} + \mathrm{LR}} \times 100\% \quad (9)$$

4.2 实验结果

首先利用 PCFG 对句子进行解析。其中以一个句子的5个候选树为例进行说明，如图4所示：

-92.7973

(S1 (S (VP (pron ىمز) (adv ۇرىمجىدە) (NP (n ۋشاڤپەن) (VP (v ۋشىپ) (v كەلدىك))) (. .)))

-101.623

(S1 (S (VP (pron ىمز) (adv ۇرىمجىدە) (NP (n ۋشاڤپەن) (VP (v ۋشىپ) (v كەلدىك)) (. .))))

-106.685

(S1 (S (VP (pron ىمز) (adv ۇرىمجىدە) (NP (NP (n ۋشاڤپەن)) (VP (v ۋشىپ) (v كەلدىك))) (. .)))

-112.486

(S1 (S (VP (pron ىمز) (adv ۇرىمجىدە) (NP (n ۋشاڤپەن) (VP (v ۋشىپ) (v كەلدىك)))) (. .)))

-113.156

(S1 (S (VP (pron ىمز) (adv ۇرىمجىدە) (CC ۋشاڤپەن) (VP (v ۋشىپ) (v كەلدىك))) (. .)))

图 4 5个句法解析候选树

文中的 PCFG 句法分析器产生最佳候选解析树，其次是感知机对训练语料进行训练，再将训练得到的参数用于感知机句法重排序当中，重新计算 PCFG 句法解析器生成的候选树的总分，在感知机重排序阶段，感知机不对 PCFG 生成的句法结构做任何改动，仅通过训练得到的特征模板和参数对 PCFG 生成的句法结构的每个结点重新计算得分。在表 4 中，显示的是由 PCFG 句法分析器的准确率，召回率和 f 值，以及感知机重排序之后，和 PCFG 句法分析器的结果，按照一定比例的解析最终的句法分析的准确率（Precision)，召回率（Recall）和 F-Meature。

表 4 句法分析实验结果

	Precision (%)	Recall (%)	F (%)
PCFG	68.2	66.7	67.4
重排序	71.4	69.6	70.5

表 5 显示的是将测试语料根据不同的句子长度进行测试，该文将句子长度分为三种类型，分别是长度为 1-5 的句子;长度为 6-15 句子和长度大于 15 的句子，结果如下表所示：

表 5 不同句长实验结果

	句子长度	Precision (%)	Recall (%)	F (%)
PCFG	1-5	70.9	69.4	70.1
	6-15	68.4	67.1	67.7
	>15	65.3	64.7	65.0
重排序	1-5	74.1	73.3	73.7
	6-15	72.1	70.1	71.1
	>15	67.9	66.5	67.2

4.3 实验结果分析

从表 4 实验结果可以看出，经过感知机重排序之后的句法分析，相比较单一的使用 PCFG 句法分析器，句法分析的结果有所提高，准确率提高 3.2%，主要原因是 PCFG 对句法解析的过程是粗粒度的，不能捕捉到上下文的信息，因此在句法解析的过程中,捕捉不到上下文的信息,其次，PCFG 在进行句法解析的过程当中，还需要句法规则，因此对语料的规模要求较高，以上因素导致基于 PCFG 的句法分析结果不够理想。从表 5 实验结果可以看出，在句长为 1-5 的实验结果显示，重排序之后准确率提高 3.2%；句长为 6-15 重排序之后准确率提高 3.7%；句长大于 15 重排序之后准确率提高 2.6%。主要原因，在于较短的句子，句法结构不是特别规范，较长的句子结构较复杂，因此，对于句子结构较完整又较规范的句子效果最好。

由实验结果可以明显的看出，重排序的句法分析相比较使用 PCFG 效果要好，但在句法解析

结果中，仍有些句法解析的结果不够理想，分析主要原因如下：

（1） 在 PCFG 解析过程中，有些句子并没有匹配正确的规则，因此，产生的候选树的结果并不是特别理想

（2） 有些句子的结构比较难，还有些句子的结构，不是严格的按照句法规则，PCFG 进行句法解析的过程中存在一定的难度

（3） 语言是比较复杂的。在感知机训练中得到的参数$\bar{\alpha}$，在重排序的过程当中，计算结点，使用同一个参数，不能很好的体现每个结点的信息。

5 结束语

本文描述了由粗到精的哈萨克语短语结构句法分析。主要由 PCFG 解析器对每个待分析的句子进行解析，生成 20 个最佳候选树，然后由感知机进行训练得到参数以及特征模板，再对生成的 20 个最佳候选树进行重排序。PCFG 对语言的描述是粗粒度的，该文的重排序的方法是细粒度的，弥补了其不能捕捉到上下文信息的不足。在 PCFG 进行句法解析的过程当中，需要大量的语料，以及需要的语料题材多样性，因此，之后的工作之一是对语料以及语料题材进一步扩大。语言是复杂的，在感知机训练阶段，重排序过程当中，使用相同的参数，并没有将语言的特性很好的表现出来，在后续的重排序的的过程当中，可以考虑使用不同的参数，结合每个结点的信息，进行参数训练，对生成的候选句法树进行重排序。

参考文献

[1] 吴伟成，周俊生，曲维光．基于统计学习模型的句法分析方法综述[J]．中文信息学报，2013, 27(3):9-19.

[2] 刘挺，马金山．汉语自动句法分析的理论与方法[J]．当代语言学，2009(2):100-112.

[3] 尚文清，古丽拉·阿东别克，牛娜,等．基于 PCFG 模型的哈萨克语句法分析[J]．现代计算机:专业版，2015(5):7-10.

[4] 古丽扎达·海沙，古丽拉·阿东别克．哈萨克语动词短语自动识别研究与实现[J]．计算机工程与应用，2015, 51(2):218-223.

[5] 汪泱，古丽拉·阿东别克，户冰心,等．基于条件随机场的哈萨克语基本短语自动识别[J]．计算机工程与设计，2014(10):3602-3607.

[6] 孙瑞娜，古丽拉·阿东别克．哈萨克语基本名词短语自动识别研究与实现[J]．中文信息学报，2010, 24(6):114-119.

[7] 尚文清，古丽拉·阿东别克，牛娜,等．基于 PChart 算法的哈萨克语句法分析[J]．计算机工程与设计，2016, 37(3):832-836.

[8] Charniak E, Johnson M. Coarse-to-fine n -best parsing and MaxEnt discriminative reranking[C]// ACL 2005, Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, Usa. DBLP, 2005:173-180.

[9] Kasami T. An Efficient Recognition and Syntax-Analysis Algorithm for Context-Free Languages[J]. 1966.

[10] Collins M, Roark B. Incremental parsing with the perceptron algorithm[C]// Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004:111.

[11] Martí, Nez C, Prodinger H. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms[C]// Acl-02 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2002:1-8.

[12] Charniak E. A maximum-entropy-inspired parser[C]//Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. Association for Computational Linguistics, 2000: 132-139.

[13] Socher R, Bauer J, Manning C D, et al. Parsing with Compositional Vector Grammars[C]//ACL (1). 2013: 455-465.