文章编号：1003-0077（2014）00-0000-00

基于语块和条件随机场（CRFs）的韵律短语识别

钱揖丽[1,2]，冯志茹[1]

（1.山西大学 计算机与信息技术学院，山西 太原 030006；
2.山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006）

摘要：该文提出一种基于汉语语块这一浅层句法信息，并利用条件随机场模型的中文文本韵律短语边界预测方法。首先介绍语块的定义和标注算法，然后在进行了语块结构标注以及归并处理的语料上，利用 CRFs 算法生成相应模型对韵律短语进行识别。实验结果表明，基于语块信息的 CRFs 韵律短语识别模型的识别效果优于不利用语块结构的模型，其 F 值平均能够提高约十个百分点。

关键词：韵律短语；边界预测；语块结构；条件随机场

中图分类号：TP391　　**文献标识码：**A

Identification of Chinese prosodic phrase based on Chunk and Conditional Random Fields

QIAN Yili[1,2] , FENG Zhiru[1]

(1. School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China;
2. Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: A Chinese prosodic phrase prediction method is proposed, based on Chinese Chunk which reflects shallow syntactic information, and making use of the conditional random fields (CRFs) model. Start from providing the knowledge of the Chunk definition and its tagging algorithm, then based on the Chunk annotated corpus, it use CRFs algorithm to generate the appropriate model to predict prosodic phrase boundary. The experimental results show that, after labeling the structure of Chunk the F-sore of the CRFs model for prosodic phrase identification increased nearly 10% than before.

Key words: prosodic phrase; boundary prediction; Chinese chunk; CRFs

1 引言

语音设备的广泛普及使得人们对语音合成的自然度和清晰度有了更高的要求。韵律结构的划分是影响合成语音自然度的重要因素之一，并对机器合成语音的质量起着决定性的作用。

目前最为公认的汉语语音合成系统中韵律结构从低到高分为三个级别，分别为：韵律词、韵律短语和语调短语。级别越高，边界处的停延越长。由于韵律词往往与语法词相对应，语调短语则相当于一个较为完整的分句，所以其中韵律短语的预测最难，也最为重要，其预测结果直接影响着最终合成语音的自然度。

针对韵律短语识别问题，国内外的研究者们提出了许多方法。最早的预测方法主要是使

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学青年基金资助项目（61005053，61100138），山西省青年科技研究基金资助项目（2012021012-1），山西省高校科技开发项目（20091001），山西省自然科学基金资助项目（2011011016-2），山西省回国留学人员科研资助项目（2013-022）。

作者简介：钱揖丽（1977—），女，博士，副教授，硕士生导师，主要研究方向为自然语言处理；冯志茹（1988—），女，硕士研究生，主要研究方向为自然语言处理。

用语言学规则[1]，但是这种方法的复用度低，很容易受到人为因素的限制；紧接着出现了基于统计的预测方法，如基于二叉树[2,3]、马尔科夫模型[4]、最大熵模型[5]、决策树[6]等等，这些方法使用的模型特征大多为词、词性等词法特征，或者使用语法特征，但其语法特征依赖于人工标注；随后，为了更好地预测韵律结构，提出了将规则和统计相结合的方法，用规则去约束机器识别的结果，或者添加一定的规则然后再进行机器训练，这些研究工作使得韵律结构的划分问题取得了一定的进展。

通过对大量语料的分析可以看出，韵律结构和句法结构之间存在着一定的联系。韵律结构是以句法结构为基础的[7]。由于语块本身可以反映出一定的句法信息，且人们在朗读或说话时往往是以语块为基本单位的，语块的切分可以把句法上相关的词进行整合，所以本文将语块结构这种非递归嵌套的浅层句法结构应用于韵律短语的预测，提出了一种基于语块这种浅层的句法信息，并利用条件随机场（Conditional Random Fields，CRFs）对韵律短语进行预测的方法。该方法在总结普遍的语块标注规则，并实现语块归并的基础上，利用 CRFs 方法抽取相应的特征模型训练、构建模型用于韵律短语的识别。实验结果显示，语块信息能够为韵律短语的识别做出贡献，利用语块信息能够取得更好的韵律短语识别效果。

2 语块的定义及处理

在韵律短语的边界处有着较为明显的停顿，而人们在正常说话或朗读的时候，往往会在联系紧密的句法短语之后停歇。从下图 1 中可以看出，句法结构和韵律结构之间存在一定的联系。

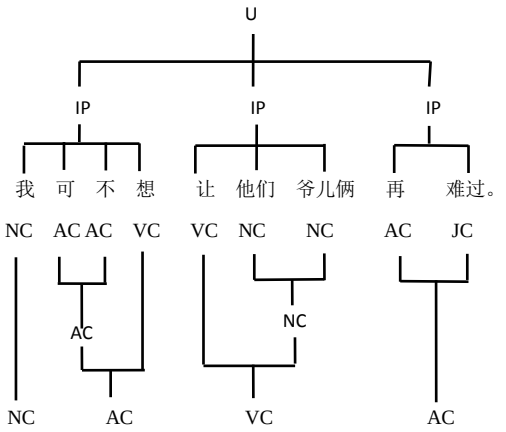


图 1 韵律短语与语块关系图

上图中，U 表示整个句子，IP 表示韵律短语，NC、AC、VC、JC 分别表示名词语块、副词语块、动词语块和形容词语块。从图中可以看出，韵律短语边界出现在语块的边界处，语块内部没有出现韵律短语边界。而且，通过对大量语料的分析可知，韵律短语边界往往出现在连接紧密度较弱的语块之间。所以，本文根据这一特性，提出将语块应用于韵律短语边界的预测中。

2.1 语块的定义

语块标注将标准的短语结构分成两部分：直接成分短语以及成分短语之间的句法依存，它可看作是理解自然语言的一个中间过程。Abney[8]定义的英文 chunk 是介于词汇与句子之间的具有非递归特征的块，其包括中心词的前置修饰成分不包括后置附属成分。中文语块定义为被标记了句法功能的非递归、非嵌套、不重叠的相邻词序列。通过对语块的研究，结合韵律短语的特点，本文定义了 8 种类型的语块结构（见表 1）。这 8 种语块都是汉语中常见的短语类型。

表 1 本文定义的语块类型

语块描述	语块类型	示例
名词语块	NC	资信/n 程度/n
动词语块	VC	组织/v
形容词语块	JC	密切/a 的/u
副词语块	AC	随后/d
介词语块	PC	在/p
连词语块	CC	而且/c
数量词语块	QC	１１/m 米/q
方位语块	LC	以下/f

2.2 初语块的标注

本文中初始语块的标注思想是：根据汉语中的句法特征总结归纳出各类语块的特征，然后利用有限状态自动机对文本中的句子进行正则匹配，从而完成语块的初始标注。

初语块的标注算法如下：

Input：未进行语块标注的文本集合 ***S***，语块标注规则集合 ***R***。

Output： 标注好语块结构的文本集合 ***T***。

Procedure of Reco:

Initial ***T***=Ø, *num*=1；

R={NC,VC,QC,PC,LC,JC,AC,CC}，

S={s_1, s_2, ……, s_n}；

从初始文本集中读取一段文本 s_i；

While（$s_i \neq$ Ø）{

匹配数据集合 ***F***= Ø；

While（*num*<=len(***R***)）{

从左到右扫描文本 s_i；

If（s_i 中含有与 *R*[*num*]匹配的项）{

将匹配的文本块 f_i 加入到匹配数据集合 ***F***；

将 f_i 标注为 *R*[*num*]型的语块【*R*[*num*] f_i】；}

num++；

}

将标注了语块的文本重新赋值给 s_i；

将 s_i 加入到语块标注结果集 ***T*** 中；

}

例如：

整整/d 一/m 天/q 的/u 交流/vn 洽谈/vn，５０５/m 名/q 博士/n 研究生/n 中/f 有/v １８６/m 人/n 达成/v 来/v 唐山/ns 工作/vn 的/u 意向/n

上述例句的语块初标注结果为：

【AC 整整/d】【QC 一/m 天/q 的/u】【VC 交流/vn】【VC 洽谈/vn】,【QC ５０５/m 名/q】【NC 博士/n 研究生/n】【LC 中/f】【VC 有/v】【QC １８６/m】【NC 人/n】【VC 达成/v】【VC 来/v】【NC 唐山/ns】【VC 工作/vn 的/u】【NC 意向/n】

2.3 语块的归并

通过对比初语块结构和韵律结构，统计和分析两者之间的潜在联系，发现：由于汉语句法结构方面的特性，汉语语块之间往往存在着较为紧密的联系。如动宾结构中，动词语块和其后的名词语块结合紧密；介词结构中，介词语块和其后的语块结合紧密，等等。所以，结合汉语句法结构的特点，基于取自 1998 年《人民日报》的 3200 个句子，对任意类型初始语

块间连接的紧密程度进行了考察，统计结果如下表 2 所示。

表 2 各语块间的结合紧密度

C_j \ C_{j+1}	NC	VC	JC	LC	AC	CC	QC	PC
NC	85.66%	73.44%	85.24%	98.29%	49.07%	12.56%	63.61%	26.71%
VC	91.93%	86.42%	77.98%	99.22%	72.80%	12.30%	87.75%	76.84%
JC	96.94%	91.25%	88.56%	95.83%	55.75%	20.21%	84.21%	67.74%
LC	73.42%	61.40%	69.01%	70.00%	65.41%	93.75%	61.63%	2.50%
AC	96.31%	98.83%	99.57%	87.50%	96.02%	75.00%	96.77%	90.76%
CC	94.57%	94.82%	95.92%	94.74%	83.77%	63.64%	87.32%	83.03%
QC	95.06%	80.65%	95.37%	98.51%	52.83%	16.67%	89.92%	39.71%
PC	97.93%	96.73%	98.48%	97.92%	95.29%	100%	95.67%	83.64%

相邻语块的结合紧密度被定义为一个条件概率，用于描述语块间不出现韵律短语边界的概率。概率越大，说明两个语块结合得越紧密。

满足规则 $R[k]$ 时不出现韵律短语边界的条件概率：

$$P(L=0 \mid R[k]) = \frac{C(L=0 \mid R[k])}{Count[k]} (1 \le k \le 64)$$

其中，$L=0$ 表示相邻语块 C_j 和 C_{j+1} 之间不出现韵律短语边界；$R[k]$ 表示第 k 条规则，描述相邻语块 C_jC_{j+1} 的类型序列；$Count[k]$ 表示满足规则 $R[k]$ 的实例总数；$C(L=0 \mid R[k])$ 表示满足规则 $R[k]$ 且在语块间未出现韵律短语边界的实例个数。

将关系紧密的初语块进行归并，能够更有利地反映句子的韵律结构。根据上表 2 归纳得到的初语块归并规则如下表 3 所示：

表 3 初语块的归并规则

初语块类型序列	归并后语块类型	归并示例
mC + LC (mC 表示任意语块)	LC	【NC 过程/n】【LC 中/f】 ⇨ 【LC 过程/n 中/f】
VC + NC	VC	【VC 组建/v】【NC 民营/b 企业/n 贷款/n】 ⇨ 【VC 组建/v 民营/b 企业/n 贷款/n】
PC +zC (zC 为除 PC 外的其他语块)	PC	【PC 以/p】【NC 速写/n 形式/n】 ⇨ 【PC 以/p 速写/n 形式/n】
AC +yC (yC 为除 CC 外的其他语块)	AC	【AC 随后/d】【VC 举行/v 了/u】 ⇨ 【AC 随后/d 举行/v 了/u】
JC + NC/VC	JC	【JC 多/a 所/q】【NC 国际/n】 ⇨ 【JC 多/a 所/q 国际/n】
CC + NC/VC/JC	CC	【CC 和/c】【VC 总结/v 了/u】 ⇨ 【CC 和/c 总结/v 了/u】
QC + NC/JC	QC	【QC 许多/m】【JC 宝贵/a】【NC 经验/n】 ⇨ 【QC 许多/m 宝贵/a 经验/n】
xC + mC (xC 为以“的”结尾的任意语块)	xC	【NC 人类/n 文明/n 的/u】【NC 大门/n】 ⇨ 【NC 人类/n 文明/n 的/u 大门/n】

3 基于 CRFs 的韵律短语边界预测

3.1 条件随机场 CRFs 模型

CRFs 是一个条件概率序列无向图模型，在给定一个观测序列条件下，CRFs 能够定义出关于整个类别标记的单一联合概率分布，从而找到全局的最优解。CRFs 不仅避免了许多模型中需要将观测对象与其他对象进行独立性假设的缺点，还能够有效的使用上下文信息，避免了类别标注偏差问题[9,10]。

在给定待识别韵律短语边界观测序列 $X=(X_1,X_2,...,X_n)$的条件下，对应的韵律短语边界标注结果序列为：$Y=(Y_1,Y_2,...,Y_m)$。

于是，CRFs 定义的条件概率为：

$$P(Y\mid X)=\frac{1}{Z(X)}\exp\left\{\sum_k l_k f_k(y_{i-1},y_i,X,i)\right\}$$

其中，$Z(X)$为归一化因子，它保证整个状态序列的概率之和为 1。

$$Z(X)=\sum_{y\in Y}\exp\left\{\sum_k l_k f_k(y_{i-1},y_i,X,i)\right\}$$

在 CRFs 中(X,Y)的确定是由局部特征转移函数 $f_k(y_{i-1},y_i,X,i)$和特征函数权重l_k共同确定的。$f_k(y_{i-1},y_i,X,i)$是关于待标注韵律短语边界观测序列的特征函数，它有两种形式，分别用来表示无向图 G = (V , E)点的状态特征和点与点之间边的转移特征。特征函数中，y_{i-1}、y_i是标注了是否是韵律短语边界的结果标签，X是输入的待标注的文本序列，i是文本序列的某个位置。

在给定训练样本集合特征转移函数之后，便可以从训练样本中训练学习得到 CRFs 模型。对于任意输入的待标注韵律短语边界观测序列 X，经过 CRFs 训练之后便会给出其相应的韵律短语标注序列Y，其中最优的标注序列就是使得条件概率取最大值的标注结果：

$$Y^*=\arg\max_Y P_\lambda(Y\mid X)$$

CRFs 超强的推理能力可以得到序列之间存在的任意关系，训练得到的模型能够得到非常丰富的信息。

3.2 特征的选取

对于 CRFs 来说，建立的模型能否高效地对韵律短语进行预测，选取合适的特征至关重要。特征的种类越多，则 CRFs 可以从训练语料中学习到知识就越多；但是，若特征太多不仅会使系统的复杂度增加，而且相关性不大的特征有时还会降低模型的性能。所以，通过对已有文献的研究，并结合语料的特点和多次的反复试验，本文最终选用的特征类型有：语块内容、语块类型、语块包含的词数以及语块的字数。并将语块内容的距离长度拓展为 1，其

余的距离长度拓展为 2。基于语块和 CRFs 的韵律短语预测模型所选用的特征模板如下表 4 所示。

表 4 韵律短语预测模型的特征模板

特征类型	符号表示	含义
Chunk	Chunk-1	前一个语块
	Chunk	当前语块本身
	Chunk+1	后一个语块
Type	Type-2	前面第二个语块的类型
	Type-1	前面第一个语块的类型
	Type	当前语块的类型
	Type+1	后面第一个语块的类型
	Type+2	后面第二个语块的类型
WLen	WLen-2	前面第二个语块所含词的个数
	WLen-1	前面第一个语块所含词的个数
	WLen	当前语块所含词的个数
	WLen+1	后面第一个语块所含词的个数
	WLen+2	后面第二个语块所含词的个数
Clen	CLen-2	前面第二个语块所含字的个数
	CLen-1	前面第一个语块所含字的个数
	CLen	当前语块所含字的个数
	CLen+1	后面第一个语块所含字的个数
	CLen+2	后面第二个语块所含字的个数

除了上述原子特征之外，上下文之间的相互联系也会对韵律短语的预测起到一定的影响，所以本文中将不同类型的原子特征进行了组合，根据多次实验的结果，模型采用了表 5 中所示的组合特征。

表 5 模型采用的组合特征

符号表示	含义
ChunkType	当前语块及其类型
ChunkClen	当前语块及其所含字的个数
TypeClen	当前语块类型及其所含字的个数
WlenClen	当前语块所含词的个数及其所含字的个数

3.3 韵律短语识别模型

本文利用语块信息并采用条件随机场方法建立了韵律短语的识别模型。实验模型的构建以及韵律短语的识别流程如下图 2 所示。

（1）识别模型的构建

基于经过了分词、词性标注、韵律标注和初始语块标注的训练预料，在分析韵律短语和语块之间关系的基础上，归纳总结规则并进行初语块的归并，然后抽取并构建特征模板训练生成 CRFs 韵律短语识别模型。

（2）韵律短语的识别

对于待识别的语料，首先进行自动分词和词性标注，然后利用正则匹配的方法进行语块的自动标注及归并，最后利用上述 CRFs 模型完成韵律短语的自动识别和标注。

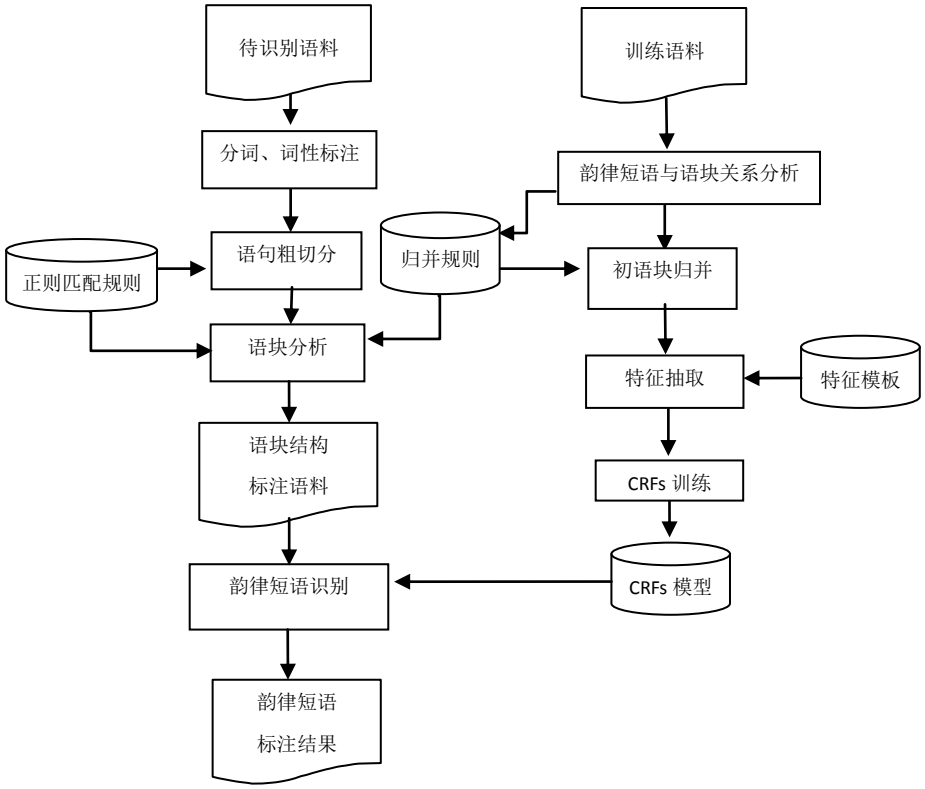


图 2 基于语块和 CRFs 的模型构建以及韵律短语识别流程

4 实验结果和分析

4.1 实验语料准备

实验使用的语料约 20 万字，是从富士通标注的 1998 年《人民日报》中随机抽取的 3200 个长句，经过了分词、词性标注以及人工韵律结构标注，平均每句含有 34.61 个词，10.36 个韵律短语。为了使实验更具说明性，由程序随机生成 10 组实验语料，每组中 2800 句作为训练语料，400 句作为开放测试语料。

4.2 实验评价指标

韵律短语识别的评价指标采用常用的精确率(P)、召回率(R)和 F 值(F)。

P = 机器正确标注的韵律短语个数 / 机器标注的韵律短语总数 ×100%

R= 机器正确标注的韵律短语个数 / 人工标注的韵律短语总数 ×100%

P = 2×P×F /（F＋P）×100%

4.3 韵律短语识别结果与分析

基于 10 组实验语料，利用本文 3 中介绍的方法进行韵律短语的自动识别，获得的开放测试结果如下表 6 所示。

表 6 10 组实验开放测试结果

	1	2	3	4	5	6	7	8	9	10	平均
精确率	0.8981	0.8717	0.9026	0.9173	0.9040	0.9091	0.8824	0.8817	0.9010	0.8984	0.8966
召回率	0.8269	0.8611	0.8345	0.8059	0.8275	0.8142	0.8443	0.8576	0.8312	0.8384	0.8342
F 值	0.8611	0.8663	0.8672	0.8580	0.8640	0.8590	0.8629	0.8695	0.8647	0.8673	0.8640

从表 6 可以看出，基于语块信息和条件随机场模型进行韵律短语识别，10 组实验的平均识别精确率为 89.66%，召回率为 83.42%，F 值为 86.4%。

另外，为了考察语块信息对于韵律短语识别的贡献，本文构建了一个不利用语块信息的 CRFs 韵律短语识别模型。借鉴前人的研究工作，该模型使用词、词性、词长为原子特征，并将原子特征距离长度拓展为 2（即当前词前后各两词），同时将原子特征组合构成复合特征（词+词性，词+词长），并设置距离长度为 1。不利用语块信息的 CRFs 模型所用的特征模板及其含义如下表 7 所示。

表 7 不利用语块信息的 CRFs 训练模板

符号表示	具体含义
Word（±2）	当前词（前后各两个词）
POS （±2）	当前词的词性（前后各两个词的词性）
Wlen（±2）	当前词的长度（前后各两个词的长度）
Word（±1）POS（±1）	当前词及其词性（前后各一个词及其词性）
Word（±1）Wlen（±1）	当前词及其词长（前后各一个词及其词长）

同样利用上述 10 组实验语料做开放测试，并将基于语块信息的 CRFs 模型与不使用语块的 CRFs 模型的韵律短语识别结果进行 F 值的对比，结果如图 3 所示。

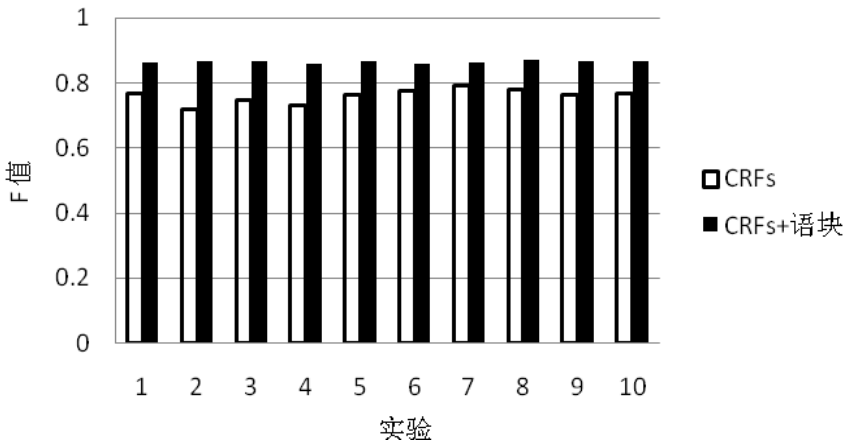


图 3 是否使用语块模型的 F 值结果比较

从图 3 可以看出，引入语块结构之后，CRFs 模型在韵律短语识别效果上有了明显的提升，其 F 值提高了 10%左右。从实验结果可以看出，语块这一浅层句法信息，能够在韵律短语识别中得到应用并做出贡献。

另外，在相同的语料集上利用不同的方法进行韵律短语识别，其识别结果与本文方法的对比情况如表 8 所示。

表 8 相同语料下不同方法的识别结果对比

方法	精确率	召回率	F-值
二叉树方法[2]	0.8261	0.7920	0.8080
最大熵方法	0.6743	0.7476	0.7090
本文方法	0.8966	0.8342	0.8640

从表 8 的测试结果可以看出，与其他方法相比，基于语块和 CRFs 的韵律短语识别方法，在识别精确率、召回率和 F 值上都有明显的提高。

5 结论

本文利用语块这种浅层句法信息建立了 CRFs 模型用于韵律短语的自动识别。首先使用有限状态自动机，利用正则匹配的方法，对语料进行了初始语块结构的识别和标注；然后，基于初语块间结合紧密度的调查，制订了归并规则对初语块进行了处理；最后，利用 CRFs 方法构建了韵律短语的识别模型。实验结果表明，基于语块信息的 CRFs 韵律识别方法优于不利用语块结构的模型，其 F 值平均能够提高约十个百分点。同时，在相同语料集上利用不同方法进行韵律短语识别的实验结果表明，本文方法优于其他两种方法。

目前，是利用正则匹配的方法进行语块结构的识别，而韵律结构比较灵活多变，往往不能像句法结构那么规则，不可避免地，少数韵律短语的边界会出现在语块结构的内部。今后的工作将针对这些问题进行深入的研究和改进，从而进一步提高韵律短语的识别效果。

参考文献：

[1] 曹剑芬. 基于语法信息的汉语韵律结构预测[J]. 中文信息学报, 2003, 17(3): 41-46.

[2] 荀恩东,钱揖丽,郭庆,等. 应用二叉树剪枝识别韵律短语边界[J]. 中文信息学报, 2006, 20(3): 1-5.

[3] 钱揖丽,荀恩东. 基于标点信息和统计语言模型的语音停顿预测[J]. 模式识别与人工智能, 2008, 21(4): 541-545.

[4] Taylor P, Black A W. Assigning phrase breaks from part-of-speech sequences[J]. Computer Speech & Language, 1998, 12(2): 99-117.

[5] 李剑锋, 胡国平, 王仁华. 基于最大熵模型的韵律短语边界预测[J]. 中文信息学报, 2004, 18(5): 56-63.

[6] 王永鑫, 蔡莲红. 语法信息与韵律结构的分析与预测[J]. 中文信息学报, 2010 (1): 65-70.

[7] 曹剑芬. 汉语韵律切分的语音学和语言学线索[C]//新世纪的现代语音学—第五届全国现代语音学学术会议论文集, 北京: 清华大学出版社, 2001: 176-179.

[8] Abney S. Prosodic structure, performance structure and phrase structure[C]//Proceedings of the workshop on Speech and Natural Language. Association for Computational Linguistics, 1992: 425-428.

[9] 董远,周涛,董乘宇,王海拉. 条件随机场模型在韵律结构预测中的应用[J]. 北京邮电大学学报,2009,05:36-40.

[10] 包森成. 基于统计模型的韵律结构预测研究[D]. 北京邮电大学, 2009.

[11]]杨鸿武,朱玲. 基于句法特征的汉语韵律边界预测[J]. 西北师范大学学报(自然科学版),2013,01:41-45.

[12] 李素建, 刘群. 汉语组块的定义和获取[C]//语言计算与基于内容的文本处理—全国第七届计算语言学联合学术会议论文集, 北京:清华大学出版社, 2003:110-115.

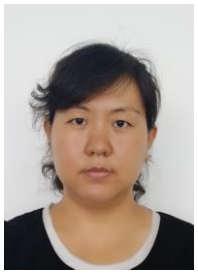
[13] 周强, 李玉梅. 汉语块分析评测任务设计[J]. 中文信息学报, 2010, 24 (1): 123-128.

[14] S. P. Abney. Parsing by chunks. In Berwick R C, Abney S P, and Tenny C(editors), Principle-based parsing: computation and psycholinguistics[M]. Kluwer Academic Publishers, Boston, 1991: 257 - 278.

[15] 周游, 刘方舟. 语调短语预测中长度约束模型的对比研究[J]. 清华大学学报: 自然科学版, 2013 (6): 787-790.

[16] 张元平, 凌震华, 戴礼荣, 等. 一种改进的基于决策树的英文韵律短语边界预测方法[J]. 计算机应用研究, 2012, 29(8): 2921-2925.

[17] Tjong Kim Sang E F, Buchholz S. Introduction to the CoNLL-2000 shared task: Chunking[C]//Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7. Association for Computational Linguistics, 2000: 127-132.



钱揖丽（1977—），女，博士，副教授，硕士生导师，主要研究领域为自然语言处理。

E-mail: qyl@sxu.edu.cn

冯志茹（1988—），女，硕士研究生，主要研究领域为自然语言处理。

E-mail: fengzhiru0321@126.com