采用无标注语料和词"粘连"剔除策略的韵律短语识别

钱揖丽[1,2] 蔡滢滢[1]
(山西大学计算机与信息技术学院 太原 030006)[1]
(山西大学计算智能与中文信息处理教育部重点实验室 太原 030006)[2]

摘 要 针对人工标注韵律结构获取大规模语料的困难和问题,利用标点符号能够表示停顿的性质,提出一种采用无标注语料和词"粘连"剔除策略的韵律短语识别方法。对标点符号划分等级,并在利用其模拟韵律边界时对其赋予不同的权重。基于无标注语料构建最大熵模型,并采取Top-K方法实现句子韵律短语边界的自动预测。通过计算相邻语法词词性间的互信息对句子进行"粘连"处理,生成"粘连"单元,并对出现在其内部的韵律边界进行剔除,实现韵律短语的自动识别。实验结果表明,获取无标注语料时对标点进行分级利用及采用"粘连"剔除策略能够明显提升模型性能,该方法能够获得较好的识别效果。

关键词 无标注语料,韵律短语边界,最大熵(ME),互信息
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.2.011

Recognition of Prosodic Phrases Based on Unlabeled Corpus and "Adhesion" Culling Strategy

QIAN Yi-li[1,2] CAI Ying-ying[1]
(School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China)[1]
(Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan 030006, China)[2]

Abstract Obtaining large-scale annotated corpus manually is very difficult and has some disadvantages. Based on the pause role of punctuation, this paper proposed a prosodic phrase recognition method which uses unlabeled corpus and "adhesion" culling strategy. In the method, punctuation is graded and given different weights when it is used to simulate the prosodic boundaries. For recognizing prosodic phrase boundaries automatically, a max entropy model is constructed based on an unlabeled corpus and a Top-K method is also used. According to the mutual information of two contiguous part of speech tagging, words are bundled into adhesion units and the prosodic boundaries appear in it are eliminated. The experimental results show that hierarchical use of punctuation and "adhesion" culling strategy can improve the performance of the model significantly. The method can obtain better recognition results.

Keywords Unlabeled corpus, Prosodic phrase boundary, Maximum entropy(ME), Mutual information

1 引言

可懂度和自然度是衡量文语转换(Test-to-Speech, TTS)水平的两个重要指标。目前,可懂度已达到相当高的水平,但自然度尚待提高。在文本分析处理中,韵律短语的划分对机器合成语音的自然度有着至关重要的影响。人们在进行语言表达时,由于生理上的约束和语义传达上的需要,通常会将一个句子划分为若干个长短不一的短语,短语与短语之间又以长短不一的无声段隔开。这些无声段似乎是有声语言的"隐形标点符号"[1]。从表层上看,这些短语非重叠、非嵌套,使语言表达的节奏感增强;从深层上看,这些短语暗示了词语之间语义上的亲疏远近不同。我们将这些停顿长短不同的短语形成的结构称为韵律结构。

关于韵律结构的等级划分,目前最为公认的是从高到低分为3个等级,即语调短语、韵律短语、韵律词。国内外已经提出了许多韵律结构预测方法。比如,较早的有基于语法信息的方法[2]、基于概率频度的方法[3]、利用规则学习算法自动归纳预测规则的方法[4]、建立分层统计模型的方法[5]、基于语块和韵律短语的长度约束规则的方法[6]等;近期的有条件随机场方法[7]、基于句子语块结构的方法[8]、分析韵律结构与语法语义之间关系并应用于预测的方法[9]、基于词类序列的方法[10]、基于语法树高度的方法[11]、基于Viterbi解码的方法[12]、基于整句相似度计算的方法[13]、基于约束模型的方法[14]、基于依存句法分析结果抽取深层句法特征进行韵律层级自动预测的方法[15]、基于句法特征的方法[16]、基于改进的决策树的算法[17]等等。以上研究存在一个共同的特点——全部需要人工进行韵律结构标注的语料库。人工标注语料库在中国语言研究和发展中发挥着不可磨灭的作用,但它们的

到稿日期:2015-03-17 返修日期:2015-06-01 本文受国家自然科学青年基金项目(61005053,61100138),山西省青年科技研究基金资助项目(2012021012-1),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)资助。
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,E-mail:qyl@sxu.edu.cn;蔡滢滢(1989—),女,硕士生,主要研究方向为自然语言处理。

规模受到人工标注巨大成本的限制，且标注结果极易受到标注人的主观影响，出现标注不一致的情况。为此，文献[1]曾提出一种利用标点符号位置模拟韵律边界的思想，用每一个语法词边界邻接标点符号的可能性大小近似地度量该位置作为韵律边界的概率，并实现了基于二叉树结构的韵律结构自动预测模型。但之前的工作存在一定的不足之处，即在利用9种中文标点符号时采取的方法是一视同仁、统一对待，对于不同标点的停顿作用没有做进一步的区分，因此处理颗粒度较大，比较笼统、粗糙，不够细腻。

针对上述问题，本文首先对基于标点的无标注语料获取进行改进，依据9种标点表示停顿的不同情况，采用较小颗粒度对其进行分级、区分利用；其次，基于大规模无人工韵律标注语料建立统计模型，并采用Top-K方法实现句子韵律短语的自动预测；最后，再利用互信息衡量相邻语法词词性间的紧密程度，"粘连"词对，并据其对韵律短语的预测结果进行噪声"剔除"处理，从而得到最终的韵律短语识别结果。

2 无标注语料的获取

2.1 标点的分级处理

文献[1]依据《标点符号用法》，并基于大规模语音文本材料语料库，对出现在9种汉语标点符号处的语音无声段进行了统计和分析，结果如表1所列。

表1 标点符号处语音无声段的平均长度

标点符号	无声段的平均时长(ms)
。	1794
……	1205
！	1132
？	1088
；	1066
：	882
，	738
——	711
、	628

其基本思想是：用能表示停顿的标点符号模拟韵律短语的边界，并认为每个语法词边界作为韵律边界的概率的大小可以用该处出现标点符号的概率的大小来衡量。因此，可以基于标点符号获取大规模的无标注语料，从而避免人工进行韵律结构标注的问题和困难。

在之前的工作中，对上述9种标点的处理方法是粗粒度、大统一，不加区分地将它们归为同一类。而从表1可以看出，虽然9种标点都能表示停顿，但是它们所表示的停顿有很大的不同。从无声段时长来看，句号"。"处的无声段最长(达1794ms)，而顿号"、"处的无声段最短(仅628ms)，差别很大。因此，应该将上述标点符号按照停顿长短不同区分对待，归类为不同等级。本文先将其分割为两个等级，则共有8种可能的分割方法，如图1所示(标点从左到右按照平均停顿时长从大到小依次排列)。

. ① …… ② ！ ③ ？ ④ ； ⑤ ： ⑥ ， ⑦ —— ⑧、

图1 标点符号的可能分割

其中，分割位置Seg的可能取值为①，…，⑧。最佳分割位置的选择将在5.2节详细介绍，本文最终选择的分割位置为⑦，之前的标点归为第1级，之后的归为第2级。

2.2 语料获取

本文所用训练语料为无韵律标注语料——分等级处理标点符号后，再进行自动分词和词性标注。语料来源于文学作品、《人民日报》、科技网文、杂志文摘等，内容涉及文学、金融、法律、管理、科技、生活、教育等多个方面。共含227815个句子，6587742个字，包含两级标点共801017个。将1级标点统一用"★"替换，2级标点用"●"替换。

训练语料示例：

他/r 在/p 追求/v 美/a 的/u 过程/n 丝毫/m 不/d 感/Vg 痛苦/a ●/w 绝望/a ★/w 恰恰相反/c ★/w 他/r 充分/a 享受/v 着/u 选择/v 的/u 自由/a ●/w 热爱/v 的/u 快乐/a ★/w

3 基于互信息的语法词"粘连"

3.1 互信息计算

互信息原是信息论中衡量两个信号之间的关联程度的一种尺度，后来引申为计算两个随机变量间关联程度的统计函数。在自然语言处理中，互信息被描述为计算两个字或词之间关联程度的量度。本文基于大规模语料库(仅做了自动分词和词性标注处理)，利用互信息对任意两个词性标记的邻接情况进行了统计和度量，并据此将联系较为紧密的语法词对"粘连"起来，形成"粘连单元"。这里定义的互信息计算公式为：

$$MI(x,y)=\log\frac{p(x,y)}{p(x)p(y)}\approx\log\frac{C(xy)\times N}{C(x)\times C(y)} \quad (1)$$

其中，$p(x)$、$p(y)$分别表示词性标记x、y在语料库中出现的概率；$p(xy)$表示词性标记x、y共现的概率；$C(x)$、$C(y)$分别为词性标记x、y在语料库中出现的次数；$C(xy)$为词性标记x、y在语料中共现的次数；N为语料库中语法词的总数。

当$MI(x,y)\gg 0$时，表示x、y关联程度强；当$MI(x,y)\approx 0$时，表示x、y关联程度较微弱；$MI(x,y)\ll 0$时，表示x、y几乎不存在关联关系。

3.2 语法词"粘连"

语法词"粘连"的基本思路是：首先，基于约含1200万个语法词的分词、词性标注语料，对任意相邻词性对之间的互信息进行统计和计算；其次，选取合适阈值，筛选得到候选"粘连"词性对集；然后，测试各候选"粘连"词性对的出错率，将出错率较高的剔除，得到最终的"粘连"词性对集；最后，采用一定的"粘连"算法，将满足条件的语法词"粘连"起来，形成"粘连"单元。

经测试，本文选取的互信息阈值为0.5，最终获得271对"粘连"词性对。

"粘连"算法描述如下：

Input：待处理的任意子句 $S=W_1/T_1\ W_2/T_2\cdots W_n/T_n$，$W_i(1\leqslant i\leqslant n)$是第i个语法词，$T_i(1\leqslant i\leqslant n)$是第i个语法词的词性；"粘连"词性对集T。

Output：子句S的"粘连"单元标注结果S′。

Step 1：For i=1 to (n−1)

If $(T_i,T_{i+1})\in T$，则将W_i、W_{i+1}粘连起来，标记为【$W_i/T_i\ W_{i+1}/T_{i+1}$】

Step 2：For i=1 to (n−2)//初"粘连"结果的合并扩展

For j=i+1 to (n−1)

If 【$W_i/T_i\cdots W_j/T_j$】and 【$W_j/T_j\ W_{j+1}/T_{j+1}$】，则将其合并标注为【$W_i/T_i\cdots W_{j+1}/T_{j+1}$】

Step 3：输出"粘连"单元标注结果S′。

例如句子：S＝一个/m 学者/n 的/u 学术/n 地位/n

不/d 是/v 靠/v 花边新闻/n 就/d 能/v 确立/v 起来/v 的/u

初"粘连"结果如图 2 所示，合并扩展后最终的"粘连"标注结果 S' 如图 3 所示。

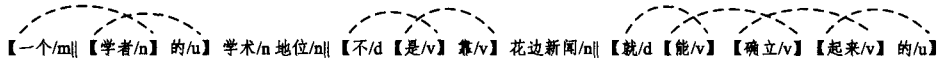

图 2 初步"粘连"结果

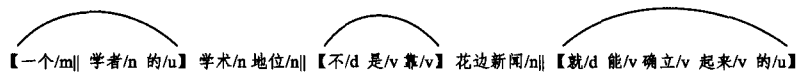

图 3 最终"粘连"结果

在"【 】"内的若干语法词为一个"粘连单元"。我们认为"粘连单元"内的语法词结合紧密，其中不出现韵律短语边界。

4 韵律短语自动识别

4.1 处理流程

基于 2.2 节中获取的无韵律结构标注语料(仅做了自动分词、词性标注和标点替换处理)，采用最大熵方法和词"粘连"剔除策略的韵律短语自动识别处理流程如图 4 所示。

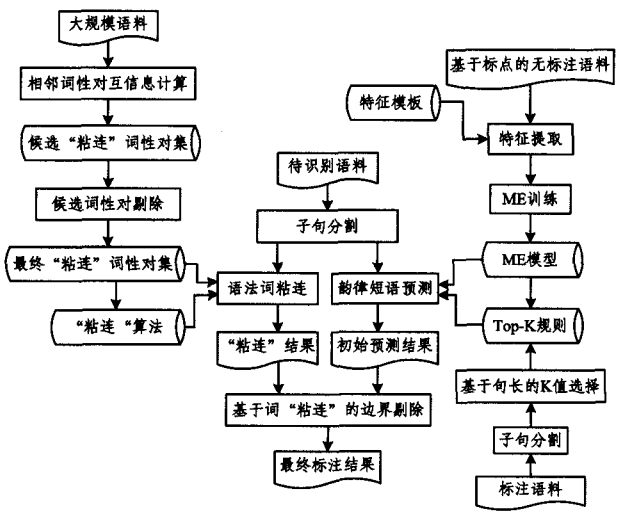

图 4 韵律短语自动识别流程

加工处理流程主要包括以下几个部分：

(1)ME 模型的构建

选择特征，建立特征模板，并基于利用标点符号获取的无标注语料训练生成最大熵模型，用于韵律短语边界的自动预测。

(2)Top-K 参数的确定

针对不同的句子长度，确定参数 K 的不同取值，并采用 Top-K 方法将待识别句子中计算概率最大的 K 个语法词边界预测为韵律边界。

(3)"粘连"词性对集获取

基于大规模分词、词性标注语料，统计和计算任意相邻词性对间的互信息，并选取合适阈值，筛选得到候选"粘连"词性对集；然后，剔除候选集中出错率较高的"粘连"词性对，得到最终的"粘连"词性对集。

(4)韵律短语边界预测

在对待识别语料进行子句分割的基础上，利用构建的最大熵模型完成韵律短语边界的初步预测。

(5)基于词"粘连"的边界剔除

基于"粘连"词性对集和"粘连"算法，对待识别语料进行"粘连"处理和标注，并依据标注结果对韵律短语边界初预测结果进行噪声剔除，获得最终的韵律短语识别结果。

4.2 模型特征选择

最大熵模型是一个概率模型。近年来，它在自然语言处理的多个领域中都取得了很好的应用效果，如词性标注、文本分类、短语识别、浅层句法分析及机器翻译等。因此，本文采用最大熵方法建立统计模型。

对于统计模型来说，特征的选取直接影响着模型的性能。特征选取过多、过少都会带来一系列的问题，最终降低模型的性能。另外，上下文语境窗口大小的选择也至关重要，过大的窗口涵盖较多上下文信息，但会引起数据稀疏问题；过小的窗口又会造成信息损失。本文选用基于频次的特征选择方法(Count Cutoff Feature Selection，CCFS)，并经过反复实验，最终选用的特征类型有：词、词性、边界类型、词长、距离，以及其组合，并选定窗口大小为"6+1"。模型选用的由 19 个原子特征和复合特征共同构成的特征模板集如表 2 所列。

表 2 模型特征模板

特征类型		符号表示	含义
原子特征	词	W−i(i=1,2,3)	当前词前面第 i 个词
		W0	当前词
		W+i(i=1,2,3)	当前词后面第 i 个词
	词性	P−i(i=1,2,3)	当前词前面第 i 个词的词性
		P0	当前词的词性
		P+i(i=1,2,3)	当前词后面第 i 个词的词性
	词长	L0	当前词所含的音节数
	边界类型	LastMark	前一个边界的类型
	距离	Dis	与上一边界的距离(音节数)
组合特征	词性组合	p(i−3)p(i−2)p(i−1)pi (i=0,1,2,3)	含当前词的四词串的词性
	词性词长组合	p(i−2)L(i−2)p(i−1) L(i−1)p0L0(i=0,1,2)	含当前词的三词串的词性、词长

4.3 基于无标注语料的韵律短语边界预测

基于构建的最大熵模型，采用 Top-K 原则(K 取决于待识别子句的长度)，分情况对句子中的韵律短语边界进行预测和标注。

4.3.1 Top-K 中参数 K 的选择

输入任意经过自动分词和词性标注的句子 $Sent=W_1W_2\cdots W_n$，$W_i(1\leqslant i\leqslant n)$是句子 $Sent$ 中的第 i 个语法词。利用上文构建的最大熵模型能够计算出每个语法词边界作为韵律短语边界的概率，包含 n 个语法词的句子 $Sent$ 中共存在 $n-1$ 个这样的潜在边界。常用的方法是选择一个统一的阈值，将满足条件的所有边界都预测为韵律短语边界。但实际上，句子 $Sent$ 中韵律边界的个数与句子的长度直接相关，一般来说，句子越长，其中包含的韵律边界数就越多。因此，本文采用 Top-K 方法选择 $Sent$ 中概率最大的 K 个语法词边界预测为韵律边界，并且句长不同，K 的取值也不相同。

基于取自 1998 年《人民日报》并进行了韵律标注的约

15000 个子句，本文对不同子句长度 *Length*（语法词个数）下韵律边界个数 *Sign* 的分布情况进行了考察，得到的结果如图 5 所示。

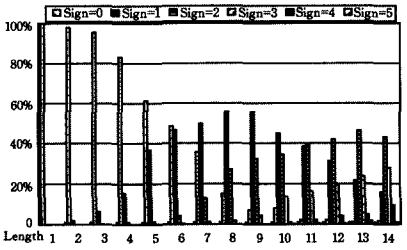

图 5　不同子句长度 *Length* 下韵律边界个数 *Sign* 的分布

分析图 5 可知：(1) *Length*≤4 时，子句中几乎没有出现韵律短语边界，所以此时取 $K=0$；(2) 5≤*Length*≤6 时，子句中出现 2 个边界的概率很小，而无边界和一个边界所占比列不相上下，此时取 $K=1$；(3) 7≤*Length*≤9 时，虽然占最大比例的仍是 1 个边界，但此时出现 2 个边界的比例显著增加，为了保证召回率，此时取 $K=2$；(4) 同理，10≤*Length*≤14 时，取 $K=3$。

另外，由于 *Length*≥15 的子句在语料中仅占 4%左右，总结规律不易，类推上述取值情况，令 $K=$取整(句长/3)。一定程度扩大 K 的取值在提高召回率的同时也会带来负面影响，但在后续的“剔除”处理中能将噪声边界剔除，起到平衡作用。

4.3.2　预测算法

算法的基本思想是：用每个语法词边界出现标点的概率大小估计该处作为韵律短语边界的概率大小，出现标点的概率越大，作为韵律边界的概率也越大。而每个语法词边界出现标点的概率则用出现 1 级标点和出现 2 级标点的综合概率来衡量。

预测算法描述如下：

设：第 i 个语法词边界处出现 1 级标点和 2 级标点的概率分别为 P_{i1}、P_{i2}。

Input：待标注的任意子句 $Sent=W_1W_2\cdots W_n$，W_i ($1\leqslant i\leqslant n$) 是第 i 个语法词。

Output：进行了韵律短语标注的序列 Sent′。

Step 1：For i=1 to (n−1)

利用最大熵分别计算 W_i 处出现 1 级、2 级标点的概率 P_{i1}、P_{i2}；

计算其作为韵律短语边界的概率 $P_i=\lambda_1P_{i1}+\lambda_2P_{i2}$，$\lambda_1$，$\lambda_2$ 为权重。

Step 2：计算 Sent 的长度 Length，并选定参数 K 的取值。

Step 3：对 P_i ($1\leqslant i\leqslant n-1$) 进行排序，并将概率最大的 K 个语法词边界预测为韵律短语边界。

Step 4：输出标注了韵律边界的序列 Sent′。

例如：

待识别子句为：一个/m 学者/n 的/ u学术/n 地位/n 不/d 是/v 靠/v 花边新闻/n 就/d 能/v 确立/v 起来/v 的/u

依据 Top-K 概率的预测结果为：

一个/m‖ 学者/n 的/u 学术/n 地位/n‖ 不/d 是/v 靠/v 花边新闻/n‖ 就/d 能/v 确立/v 起来/v 的/u

(注：length=14，故 K 取值为 3)

4.4　基于词“粘连”的边界剔除

在上述基于最大熵模型和 Top-K 方法的韵律短语识别结果中存在一定的噪声边界，因此，采用第 3 节中获取的词“粘连”规则，对待识别语料中的“粘连”单元进行识别和标注；并对包含在“粘连”单元内部的边界预测结果进行剔除，以此保证预测结果的正确率。

例如，上述例句的“粘连”单元标注及剔除后的最终识别结果如图 6 所示。

图 6　最终的识别结果

其中，虚线弧标示最大熵模型的韵律短语预测结果；双线弧标示“粘连”单元；最外侧实线弧标示将“粘连”单元中的边界剔除后最终的韵律短语识别结果。在上例中，位于语法词“一个/m”后面的预测边界位于“粘连”单元内部，故会被剔除。

5　实验与结果分析

5.1　测试语料及评价指标

本文所用测试语料是来源于 1998 年《人民日报》的 750 个长句，共有 21072 个语法词，平均句长约为 29.27 个语法词。语料经过分词、词性标注以及人工韵律结构标注，共标注了 5186 个韵律短语边界，平均每句含有 7.91 个韵律短语。

实验采用的评价指标为：准确率(P)、召回率(R)和 F 值(F)。

$P=$模型正确识别的韵律边界个数/模型识别出的韵律边界总数×100%

$R=$模型正确识别的韵律边界个数/人工标注的韵律边界总数×100%

$F=2\times P\times R/(P+R)\times 100\%$

5.2　标点分割位置 Seg 及两级标点权重 λ_1、λ_2 的确定

对于①，…，⑧每一个可能的标点分割位置，分别训练得到最优的 λ_1、λ_2 (设 $\lambda_1+\lambda_2=1$) 分配，并计算其韵律短语识别的准确率(P)、召回率(R)及 F 值(F)。实验结果如表 3 所列。

表 3　不同参数下的最优结果比较

标点分割位置及参数取值	P	R	F
Seg 取①，$\lambda_1=0.1$，$\lambda_2=0.9$	82.53%	78.65%	80.55%
Seg 取②，$\lambda_1=0.2$，$\lambda_2=0.8$	82.68%	78.44%	80.50%
Seg 取③，$\lambda_1=0.1$，$\lambda_2=0.9$	85.76%	81.00%	83.32%
Seg 取④，$\lambda_1=0.3$，$\lambda_2=0.7$	87.34%	82.79%	85.00%
Seg 取⑤，$\lambda_1=0.2$，$\lambda_2=0.8$	87.36%	82.81%	85.03%
Seg 取⑥，$\lambda_1=0.8$，$\lambda_2=0.2$	90.86%	87.89%	89.35%
Seg 取⑦，$\lambda_1=0.9$，$\lambda_2=0.1$	92.92%	91.30%	92.10%
Seg 取⑧，$\lambda_1=0.9$，$\lambda_2=0.1$	92.18%	89.88%	91.01%

依据表 3，本文将标点分割位置 Seg 选为⑦，λ_1 取值 0.9，λ_2 取值 0.1，此时模型的性能最优，其韵律短语识别准确率为 92.92%，召回率为 91.30%，F 值为 92.10%。

5.3　词“粘连”剔除策略的影响

为了考察词“粘连”剔除策略的影响，对韵律短语识别过

程中各个阶段的实验数据进行了统计和分析，结果如表4所列。

表4 词“粘连”剔除策略的影响

	最大熵模型的预测结果	“粘连”剔除情况	最终模型的识别结果
边界总数	6382	1286	5096
正确个数	4863	1158	4735
错误个数	1519	128	361
准确率	76.20%	90.05%	92.92%

依据表4可知：(1)利用最大熵模型预测的6382个边界中，只有4863个正确边界，预测准确率为76.20%；(2)通过词“粘连”处理，共剔除了1286个边界，其中1158个剔除正确，剔除准确率为90.05%；(3)最终模型共识别出5096个韵律边界，4735个正确，识别准确率为92.92%；(4)词“粘连”剔除策略使得模型的准确率由76.20%提高到92.92%，提高了16.72%。

5.4 标点分级处理的影响

为了考察将标点符号分等级处理所带来的影响，将本文模型(标点分级)与之前模型[1](标点不分级)的性能进行了比较，结果如图7所示。

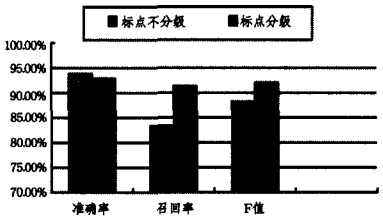

图7 标点分级对模型性能的影响

分析图7可知：(1)将标点分级利用后，模型的准确率略有下降，这是因为在利用最大熵模型进行边界预测时，Top-K中选择的参数K偏大，所以造成句子切分过细、噪声增加；(2)与之前相比，标点分级模型的召回率获得了大幅的提高，这是因为通过剔除操作，能够以很高的正确率将噪声边界剔除；(3)标点分级处理且采用“粘连”剔除策略后，模型的综合性能有了大幅的提升。

结束语 在语音合成中，正确划分韵律结构对于提高机器合成语音的自然度具有重要的意义和作用。本文摒弃传统的采用人工标注语料的方法，基于标点符号表示停顿的性质，利用不同标点分级模拟韵律结构边界，获取大规模无标注语料。基于无标注语料，利用最大熵方法和Top-K原则对汉语句子中的韵律短语边界进行预测。利用互信息计算相邻语法词间的“邻接紧密度”，对其进行“粘连”处理，并依据“粘连”结果对预测出的韵律边界进行“修剪”、“剔除”，从而提高韵律短语的自动识别效果。实验结果表明，标点符号的分级利用、基于“粘连”的剔除策略都能够明显提升模型的性能，本文方法能够获得较好的识别效果。

采用基于标点符号获取的无标注语料，能够极大地降低手工工作量，减少人力消耗，避免人工标注的问题和困难，对于相关研究工作的开展具有积极的意义和实用价值。今后将在进一步细化标点符号的分级、探索更好的预测方法等方面进行更加深入的研究和实验。

参 考 文 献

[1] Qian Yi-li, Xun En-dong. Prediction of Speech Pauses Based on Punctuation Information and Statistical Language Model[J]. Pattern Recognition and Artificial Intelligence, 2008, 21(4): 541-545(in Chinese)
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

[2] Cao Jian-fen. Prediction of Prosodic Organization Based on Grammatical Information [J]. Journal of Chinese Information Processing, 2003, 17(3): 41-46(in Chinese)
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

[3] Zheng Min, Cai Lian-hong. Statistical model based on probability frequency for Mandarin prosodic structure prediction[J]. Journal of Tsinghua University(Science and Technology), 2006, 46(1): 78-81(in Chinese)
郑敏，蔡莲红. 基于概率频度的普通话韵律结构预测统计模型[J]. 清华大学学报(自然科学版)，2006，46(1)：78-81

[4] Zhao Sheng, Tao Jian-hua, Cai Lian-hong. Rule-learning Based Prosodic Structure Prediction[J]. Journal of Chinese Information Processing, 2002, 16(5): 30-37(in Chinese)
赵晟，陶建华，蔡莲红. 基于规则学习的韵律结构预测[J]. 中文信息学报，2002，16(5)：30-37

[5] Ostendorf M, Veilleux N. A hierarchical stochastic model for automatic prediction of prosodic boundary location[J]. Computational Linguistics, 1994, 20(1): 27-54

[6] Atterer M, Klein E. Integrating linguistic and performance-based constraints for assigning phrase breaks[C]// Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2002: 1-7

[7] Dong Yuan, Zhou Tao, Dong Cheng-yu, et al. Prosodic Structure Prediction Based on Conditional Random Field Model[J]. Journal of Beijing University of Posts and Telecommunications, 2009, 32(5): 36-40(in Chinese)
董远，周涛，董乘宇，等. 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，2009，32(5)：36-40

[8] Qian Yi-li, Feng Zhi-ru. Identification of Chinese Prosodic Phrase Based on Chunk and CRF[J]. Journal of Chinese Information Processing, 2014, 28(5): 32-38(in Chinese)
钱揖丽，冯志茹. 基于语块和条件随机场(CRFs)的韵律短语识别[J]. 中文信息学报，2014，28(5)：32-38

[9] Wang Yong-xin, Cai Lian-hong. Syntactic Information and Analysis and Prediction of Prosody Structure[J]. Journal of Chinese Information Processing, 2010, 24(1): 65-70(in Chinese)
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，2010，24(1)：65-70

[10] Pei Yu-lai, Qiu Jin-ping, Wang Hong-jun, et al. Chinese sentence prosodic structure prediction based on the sequence of the parts of Speech[J]. Journal of Tsinghua University(Science and Technology), 2009(S1): 1339-1343(in Chinese)
裴雨来，邱金萍，王洪君，等. 基于词类序列的汉语语句韵律结构预测[J]. 清华大学学报(自然科学版)，2009(S1)：1339-1343

[11] Yang Hong-wu, Wang Xiao-li, Chen Long, et al. Predicting Chinese prosodic phrase with height of syntax tree[J]. Computer Engineering and Applications, 2010, 46(36): 139-143(in Chinese)
杨鸿武，王晓丽，陈龙，等. 基于语法树高度的汉语韵律短语预测[J]. 计算机工程与应用，2010，46(36)：139-143

[12] Yang Chen-yu, Zhu Li-xin, Ling Zhen-hua, et al. Automatic Phrase boundary labeling for a Mandarin TTS corpus using the Viterbi decoding algorithm[J]. Journal of Tsinghua University (Science and Technology), 2011, 51(9): 1267-1281(in Chinese)
杨辰雨，朱立新，凌震华，等. 基于 Viterb 解码的中文合成音库韵律短语边界自动标注[J]. 清华大学学报(自然科学版)，2011，51(9)：1276-1281

[13] Li Jian-feng, Hu Guo-ping, Wang Ren-hua. New Prosody Phrase Prediction Model Based on Whole Sentence Similarity Computing[J]. Journal of Chinese Computer Systems, 2006, 27(10): 1935-1938(in Chinese)
李剑锋，胡国平，王仁华. 基于整句相似性计算的韵律短语预测模型[J]. 小型微型计算机系统，2006，27(10)：1935-1938

[14] Dong Hong-hui, Tao Jian-hua, Xu Bo. Chinese Prosodic Phrasing with a Constraint based Approach[J]. Journal of Chinese Information Processing, 2007, 21(1): 54-59(in Chinese)
董宏辉，陶建华，徐波. 基于约束模型的韵律短语预测[J]. 中文信息学报，2007，21(1)：54-59

[15] Shao Yan-qiu, Hui Zhi-fang, Han Ji-qing, et al. A Study on Chinese Prosodic Hierarchy Prediction Based on Dependency Grammar Analysis[J]. Journal of Chinese Information Processing, 2008, 22(2): 116-123(in Chinese)
邵艳秋，穗志方，韩纪庆，等. 基于依存句法分析的汉语韵律层级自动预测技术研究[J]. 中文信息学报，2008，22(2)：116-123

[16] Yang Hong-wu, Zhu Ling. Predicting Chinese Prosodic boundary based on syntactic features[J]. Journal of Northwest Normal University (Natural Science), 2013, 49(1): 41-45(in Chinese)
杨鸿武，朱玲. 基于句法特征的汉语韵律边界预[J]. 西北师范大学学报(自然科学版)，2013，49(1)：41-45

[17] Zhang Yuan-ping, Ling Zhen-hua, Dai Li-rong, et al. Improved decision tree based method for English prosodic phrase boundary Prediction[J]. Application Research of Computers, 2012, 29(8): 2921-2925(in Chinese)
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

(上接第 34 页)

[9] Yang X Q, Su J, Zhou G D, et al. An NP-cluster based approach to coreference resolution[C]// Proceedings of the 20th international conference on Computational Linguistics. Morristown: ACL, 2004: 226-232

[10] Denis P, Baldridge J. A Ranking Approach to Pronoun Resolution[C]// International Joint Conferences on Artificial Intelligence. Hyderabad: ACL, 2007: 1588-1593

[11] Rahman A, Ng V. Narrowing the modeling gap: a cluster-ranking approach to coreference resolution [J]. Journal of Artificial Intelligence Research, 2011, 40(1): 469-521

[12] Kong Fang, Zhu Qiao-ming, Zhou Guo-dong, et al. Coreference Resolution Based on Center Theory [J]. Computer Science, 2009, 36(6): 219-222 (in Chinese)
孔芳，朱巧明，周国栋，等. 基于中心理论的指代消解研究[J]. 计算机科学，2009，36(6)：219-222

[13] Gao Jun-wei, Kong Fang, Zhu Qiao-ming, et al. Research of Chinese Noun Phrase Anaphora Resolution: A SVM-based Approach[J]. Computer Science, 2012, 39(10): 231-234(in Chinese)
高俊伟，孔芳，朱巧明，等. 基于 SVM 的中文名词短语指代消解研究[J]. 计算机科学，2012，39(10)：231-234

[14] Kong Fang, Zhou Guo-dong. Pronoun Resolution in English and Chinese Languages Based on Tree Kernel[J]. Journal of Software, 2012, 23(5): 1085-1099(in Chinese)
孔芳，周国栋. 基于树核函数的中英文代词消解[J]. 软件学报，2012，23(5)：1085-1099

[15] Xi Xue-feng, Zhou Guo-dong. Pronoun Resolution Based on Deep Learning[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1): 100-110(in Chinese)
奚雪峰，周国栋. 基于 Deep Learning 的代词指代消解[J]. 北京大学学报(自然科学版)，2014，50(1)：100-110

[16] Raghunathan K, Lee H, Rangarajan S. A multipass sieve for coreference resolution [M]. MIT, Massa chusetts: ACL, 2010: 492-501

[17] Lee H, Peirsman Y, Chang A, et al. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task [C]// Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task. Oregon: ACL, 2011: 28-34

[18] Lee H, Chang A, Peirsman Y, et al. Deterministic coreference resolution based on entity-centric, precision-ranked rules[J]. Computational Linguistics, 2013, 39(4): 885-916

[19] Zhang X C, Wu H. Chinese coreference resolution via ordered filtering[C]// Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning: Shared Task. Jeju: ACL, 2012: 95-99

[20] Kong Fang, Zhu Qiao-ming, Zhou Guo-dong, et al. Anaphoricity Determination for Coreference Resolution in English and Chinese[J]. Journal of Computer Research and Development, 2012, 49(5): 1072-1085(in Chinese)
孔芳，朱巧明，周国栋，等. 中英文指代消解中待消解项识别的研究[J]. 计算机研究与发展，2012，49(5)：1072-1085

[21] Marc V, John B, John A, et al. A model theoretic coreference scoring scheme[C]// Proceedings of the 6th Message Understanding Conference. Stroudsburg: ACL, 1995: 45-52

[22] Amit B, Breck B. A lgorithms for scoring coreference chains [C]// Proceedings of LREC. Granada: ACL, 1998: 563-566

[23] Luo X Q. On coreference resolution performance metrics[C]// Proceedings of HLT-EMNLP Stroudsburg. ACL, 2005: 25-32

[24] Nong Y. The Handbook of Data Mining [M]. Cleveland CRC Press, 2001: 247- 277

[25] Marta R, Eduard H. BLANC: Implementing the Rand Index for coreference evaluation[J]. Natural Language Engineering, 2011, 17(4): 485-510

[26] Doddington G, Mitchell A, Przybocki M. The automatic content extraction (ACE) program-tasks, data, and evaluation [DB/OL]. 2012-5-11. http:// www. comp. nus. edu. sg/~ rpnlpir/ proceedings/ lrec-2004 /pdf/ 5. pdf