平仄信息对中文词法分析的影响*

孟凡东[1] 徐金安[2] 姜文斌[1] 刘群[1]
[1]中国科学院 计算技术研究所 智能信息处理重点实验室，北京 100190
[2]北京交通大学 计算机与信息技术学院，北京 100044
Email:{mengfandong, jiangwenbin, liuqun}@ict.ac.cn; jaxu@bjtu.edu.cn

摘 要：词法分析，作为自然语言处理领域的基础性研究课题之一，其效果直接影响自然语言处理后续的工作。本文从汉语语音的特色出发，利用机器学习的方法，学习汉语句子中词语的平仄信息，研究平仄信息对词法分析的影响，并分别在人民日报语料和宾州中文树库语料上设计实验，实验结果证明了平仄信息特征确实能够大幅度提高中文词法分析的精度。

关键词：词法分析；平仄信息；计算语言学；

The Influence of the Level and Oblique Tones Information on Chinese Lexical Analysis

Fandong Meng[1] Jin'an Xu[2] Wenbin Jiang[1] Qun Liu[1]
[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Science, Beijing 100190, China;
[2]School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

Email:{mengfandong, jiangwenbin, liuqun}@ict.ac.cn; jaxu@bjtu.edu.cn

Abstract: *Lexical analysis, as one of basic research work in natural language processing (nlp) field, the performance of which directly affects nlp's follow-up work. This paper focuses on Chinese pronunciation characteristics, learns Chinese words' level and oblique tones in sentences and researches the influences they have on Lexical analysis, using machine learning methods. We design experiments on People's Daily corpus and Penn Chinese Treebank corpus, respectively. The results proved that level and oblique tones will do improve the performance of Chinese Lexical analysis by a large margin.*

Keywords: *lexical analysis; level and oblique tones information; computational linguistics;*

1 引言

作为自然语言处理领域的基础性研究课题之一，词法分析的效果直接影响自然语言处理后续的工作。词法分析的难点主要在于切分歧义，对于汉语这种词与词之间没有明显间隙的语言来说尤其显著。

在改进词法分析方面，涌现了大量的出色的工作。其中 Jiang et al.(2009)提出词性标注迁移的方法，利用大语料修正小语料以提高词法分析精度；Zhongguo Li and Maosong Sun (2009)利用标点符号做标记来训练统计模型，增强未登录词的识别能力，提高词法分析性能；Zhongguo Li(2011)提出在词法分析过程中分析词语的内部结构的方法。以上工作在词法分析方面都取得了很好的效果，使词法分析精度有很大提高。

本文从汉语语音的特色出发，研究平仄信息（声调一声或二声为平音，三声或四声为

*本文承中央高校基本科研业务费专项资金项目（2009JBM027）和国家自然科学基金项目（60873167，60736014）的资助。

仄音）对词法分析的影响。古汉语中，古诗词讲究平仄、押韵，例如仄起平落、平仄相间等，诵读起来朗朗上口（如图 1）。现代汉语平仄韵律方面虽然没有古诗词那么明显，但也没有完全摒弃汉语的韵律之风，仍有规律可循。

平 仄 平 平 仄 ， 平 平 仄 仄 平
国 破 山 河 在 ， 城 春 草 木 深

平 仄 仄 平 平 仄 仄 ， 平 平 仄 仄 仄 平 平
沧 海 月 明 珠 有 泪 ， 蓝 田 日 暖 玉 生 烟

图 1 古汉语诗句平仄韵律举例

本文利用机器学习的方法，学习汉语句子中词语的平仄信息，探求其对词法分析的影响。并分别在人民日报语料和宾州树库语料上设计实验，实验结果证明了平仄信息特征确实能够大幅度提高中文词法分析的精度。

本文在第 2 节简要介绍采用的词法分析方法，第 3 节详细阐述利用平仄信息辅助词法分析的思想，第 4 节是实验及结果分析，第 5 节是对本文的总结与展望。

2 中文词法分析方法

本文采用判别式的词法分析方法。将分词和词性标注问题转化为字符（汉字）分类问题。根据 Ng and Low (2004)的方法，分词采用四种位置标记，b 表示词首，m 表示词中，e 表示词尾，s 表示单个汉字独立成词。即一个词只可以被标记成 s（单字词）或 bm*e（多字词）。联合分词与词性标注就是对于每个字，有位置标记和词性标记，例如“e_v”，表示一个动词的词尾。

2.1 分词基础特征模板

根据 Ng and Low (2004)的方法，用C_0表示当前的汉字,C_{-i}表示C_0左边第i个汉字,C_i表示 C_0右边第i个汉字。$Pu(C_i)$用于判断当前汉字C_i是否为分隔符(是就返回 1，否则返回 0)。$T(C_i)$用于判断当前汉字C_i的类别：数字，日期，英文字母，和其它(分别返回 1，2，3 和 4)。

表 1. 基础特征模板

序号	基础特征模板
1	$C_i(i=-2...2)$
2	$C_iC_{i+1}(i=-2...1)$
3	$C_{-1}C_1$
4	$Pu(C_0)$
5	$T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$

表 1 描述了分词和词性标注的基础特征模板。假设当前分析的汉字是 “450 公里”中的“0”，模板 1 生成特征：$C_{-2}=4$ $C_{-1}=5$ $C_0=0$ $C_1=$公 $C_2=$里，模板 2 生成特征：$C_{-2}C_{-1}=45$ $C_{-1}C_0=50$ $C_0C_1=0$公 $C_1C_2=$公里，模板 3 生成特征：$C_{-1}C_1=5$公，模板 4 生成特征：$Pu(C_0)=0$，模板 5 生成特征：$T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)=11144$。

2.2 训练算法

本文采用了 Collins（2002）的平均感知机训练算法。训练的过程就是学习一个从输入$x\in X$映射到输出$y\in Y$的判别模型，X是训练语料中的句子集合，Y是相应的标记结果。Jiang et al.(2009)中使用了$GEN(x)$函数列举输入x的所有候选结果，表示每个训练实例$(x,y)\in X\times Y$映射到特征向量$\Phi(x,y)\in R^d$，对于一个特征向量，$\bar{\alpha}\in R^d$是与其对应的参数向量。对于一个输入的汉字串x,目的是找到一个满足下式的输出结果$F(x)$：

$$F(x) = \arg\max_{y \in GEN(x)} \Phi(x,y) \cdot \vec{\alpha} \quad (1)$$

其中$\Phi(x,y)\cdot\vec{\alpha}$表示特征向量$\Phi(x,y)$和参数向量的内积。本文沿用此方法。

图 2 描述了感知机训练算法。本文使用了“平均参数”技术(Collins, 2002)避免过拟合。

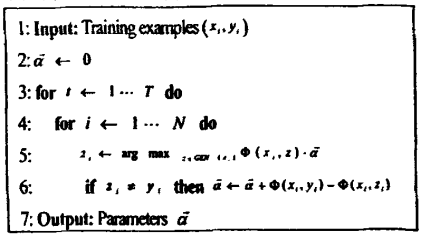

图 2 感知机训练算法的伪代码

3 引入平仄信息

3.1 平仄词典

该方法需要一部标注平仄信息的词典。词典中每一条信息是由词语本身以及词中每个字的平、仄音组成的，具体形式如下：

长江/PP

长颈鹿/PZZ

长久之计/PZPZ

……

其中 P 表示平音（声调为一声或二声），Z 表示仄音（声调为三声或四声）。由于绝大多数词都是二字、三字或四字的，所以暂不考虑单字词和多于四字词的词典及其平仄音。

3.2 平仄信息特征模板

表 2. 平仄信息特征模板

序号	平仄信息特征模板
1	$D_iD_{i+1}(i=-2\ldots1)$
2	$D_iD_{i+1}D_{i+2}(i=-2\ldots0)$
3	$D_iD_{i+1}D_{i+2}D_{i+3}(i=-2,-1)$
4	$PZ_iPZ_{i+1}(i=-2\ldots1)$
5	$PZ_iPZ_{i+1}PZ_{i+2}(i=-2\ldots0)$
6	$PZ_iPZ_{i+1}PZ_{i+2}PZ_{i+3}(i=-2,-1)$

表 2 描述了平仄信息特征模板。若当前词是平仄词典中的词，用 D 表示词典中的词条信息，包括“词本身_平仄信息”，用 PZ 标志平仄信息。下标表示字的相对位置，0 表示当前正考虑的位置，-i 表示相对于当前考虑的左边第 i 个位置，i 表示相对于当前考虑的右边第 i 个位置。例如，当前分析汉字串“是长久之计”的“久”字，平仄信息特征如下：

$D_{-1}D_0$ = 长久_PZ

$D_{-1}D_0D_1D_2$ = 长久之计_PZPZ

$PZ_{-1}PZ_0$ = PZ

$PZ_{-1}PZ_0ZP_1PZ_2$ = PZPZ

由于，“是长”、“久之”、“之计”、“是长久”、“长久只”、“久之计”、“是长久之”等不是平仄词典中词，因此没有形成平仄信息特征。

如果，将平仄词典中词语的平仄信息去掉，并且平仄信息特征模板中的平仄信息也不

予考虑，那么该模板可退化为词典模板。例如，同样还是分析汉字串“是长久之计”的“久”字，词典信息特征如下：

$D_{-1}D_0$ = 长久

$D_{-1}D_0D_1D_2$ = 长久之计

同时利用基础特征模板与平仄信息特征模板便可成功地将平仄信息引入到词法分析中，同时利用基础特征模板与词典特征模板便可引入词典信息。本文分别对这两种方法进行了实验验证。

3.3 训练与解码

训练引入平仄信息的词法分析器，需要利用基础特征和平仄信息特征；训练只用词典信息的词法分析器，需要利用基础特征和词典特征。训练都采用平均感知机算法，解码过程都一样。

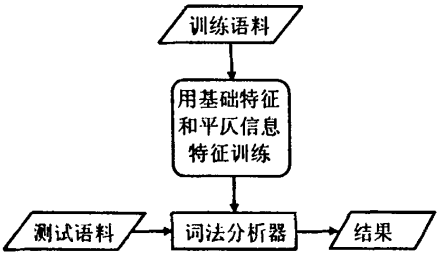

图 3 引入平仄信息的训练与解码流程

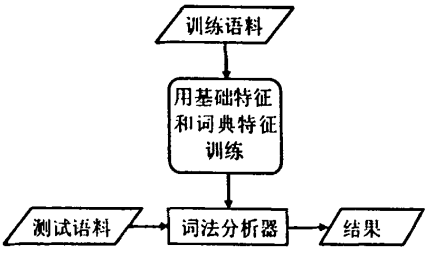

图 4 引入词典信息的训练与解码流程

图 3 描述了引入平仄信息的词法分析器的训练和解码流程。在训练语料上利用基础特征和平仄信息特征训练出一个词法分析器，利用该词法分析器直接处理测试语料，得到结果。图 4 描述了只引入词典信息的词法分析器的训练和解码流程。除了利用的特征不同（基础特征和词典特征），其余的与图 3 的流程完全一样。

4 实验与结果分析

4.1 实验数据、环境和评测方法

本文分别在人民日报语料和宾州中文树库语料上进行了单独分词与联合分词和词性标注的实验，排除训练语料分词和词性标注标准对实验结果的影响。人民日报训练语料与测试语料的句子数分别为 100344 和 19007，宾州树库训练语料与测试语料的句子数分别为 18074 和 348。带平仄信息的词典有 9 万词条，其中只有二字词、三字词和四字词及其平仄信息。

本文采用 F-measure 来评价词法分析精度，F1 = 2PR/(P+R)，其中 P 是准确率，R 是召回率。

4.2 结果与分析

表 3 中 PD 表示人民日报语料，描述了在人民日报语料上的 Baseline 模型的结果以及引入平仄信息和只引入词典信息对分词及词性标注的影响。其中，第一行是 Baseline 模型，是单独的在人民日报语料上利用感知机算法训练的模型。第二行是引入平仄信息后，单独分词和联合分词与词性标注的结果；第三行是只引入词典信息后，单独分词和联合分词与词性标注的结果（即，将平仄词典中的平仄信息去掉，将平仄特征退化为词典特征的结果）。

从第二行与第一行的对比中可以看出，引入平仄信息后，无论是单独分词还是联合分词和词性标注的精度都有大幅度提高。其中，分词的 F_1 提高了 0.56 个百分点，联合分词与词性标注的 F_1 分别提高了 0.44 个百分点（不考虑词性标注）和 0.41 个百分点（考虑词性标注）。可见，引入平仄信息特征确实可以提高词法分析精度。从第三行与第一行的对比中可

以看出，只引入词典信息后，单独分词的 F_1 值相对于 Baseline 模型略有提高，提高了 0.17 个百分点；但是在联合分词与词性标注方面却使得 F_1 值大幅度下降。分别下降 1.21 个百分点（不考虑词性标注）和 1.48 个百分点（考虑词性标注）。可见，单独引入词典信息，不加其他约束条件，不一定会改进词法分析的效果。

表 3. PD 上单独分词、联合分词与词性标注的结果

	语料	分词结果（F_1%）	联合分词和词性标注的结果（F_1 %）	
			不考虑词性标注	考虑词性标注
Baseline	PD	97.27	97.57	94.54
引入平仄信息	PD	97.83 ↑	98.01 ↑	94.95 ↑
只引入词典信息	PD	97.44 ↑	96.36 ↓	93.06 ↓

表 4 中 CTB 表示宾州中文树库语料，描述了在宾州中文树库语料上的 Baseline 模型的结果以及引入平仄信息和只引入词典信息对分词及词性标注的影响。对比实验的顺序设置与表 3 中的一致。

从第二行与第一行的对比中可以看出，引入平仄信息后，无论是单独分词还是联合分词和词性标注的精度都有大幅度提高。其中，分词的 F_1 提高了 0.92 个百分点，联合分词与词性标注的 F_1 分别提高了 0.53 个百分点（不考虑词性标注）和 0.43 个百分点（考虑词性标注）。从第三行与第一行的对比中可以看出，只引入词典信息后，单独分词的 F_1 值相对于 Baseline 模型略有提高，提高了 0.21 个百分点；但是在联合分词与词性标注方面却使得 F_1 值大幅度下降。分别下降 1.16 个百分点（不考虑词性标注）和 1.34 个百分点（考虑词性标注）。F_1 值的上升与下降规律与在 PD 上的一致。

表 4. CTB 上单独分词、联合分词与词性标注的结果

	语料	分词结果（F_1%）	联合分词和词性标注的结果（F_1 %）	
			不考虑词性标注	考虑词性标注
Baseline	CTB	97.30	97.77	93.10
引入平仄信息	CTB	98.22 ↑	98.30 ↑	93.53 ↑
只引入词典信息	CTB	97.51 ↑	96.61 ↓	91.76 ↓

由表 3 和表 4 可以看出，无论在 PD 上还是在 CTB 上，引入平仄信息后，单独分词和联合分词与词性标注的结果都大幅度提高；而只引入词典信息，单独分词的 F_1 值略有提高，但是联合分词与词性标注的 F_1 值大幅度下降。在分词方面，引入词典信息提高了 F_1 值，这是正常的，增加了词语覆盖率，增加了上下文约束，但是效果不是特别明显。在联合分词与词性标注方面，F_1 值大幅度下降，很大程度上是由于词典的噪声导致的，词性标注对语料的要求要高于单独分词。可见，单独引入词典信息，不加其他约束条件，不一定会改进词法分析的效果。由此，也反衬出确实是平仄信息使得词法分析精度提高，而不是词典信息起的作用。并且，PD 与 CBT 语料的分词和词性标注标准略有不同，但是这并没有改变平仄信息特征提高词法分析精度的事实。由此，更说明了引入平仄信息特征确实可以提高词法分析精度。

综上所述，通过一系列实验，证明了引入平仄信息确实可以大幅度提高中文词法分析的精度，单独引入词典不加其他约束条件的话不一定能提高词法分析精度。汉语词语内部发音的平仄信息确实有规律可循，如果充分、合理的利用这类规律，能够帮助改善词法分析的效果。

5 结语

本文提出了利用平仄信息改进中文词法分析的方法。我们用人民日报语料和宾州中文树库语料进行了实验，并且利用平均感知机算法，分别在人民日报语料、宾州中文树库语料上训练模型，对各个模型的分词以及联合分词与词性标注的效果进行了比较，实验结果表明，

本方法确实可以提高词法分析精度。

接下来，我们将继续研究平仄信息对于中文词法分析的影响，并着力从汉语的特色出发，挖掘汉语本身的特质，探究汉语拼音、韵律、节拍等语言学信息对词法分析的影响。

参考文献

Wenbin Jiang, Liang Huang, Yajuan Lv, and Qun Liu.2008. A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics. Pages 897-904.

Wenbin Jiang, Liang Huang, and Qun Liu.2009. Automatic Adaptation of Annotation Standards: Chinese Word Segmentation and POS Tagging – A Case Study. In Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics. Pages 522-530.

Wenbin Jiang, Haitao Mi and Qun Liu. 2008.Word Lattice Reranking for Chinese Word Segmentation and Part-of-Speech Tagging. In Proceedings of the 22nd International Conference on Computational Linguistics. Pages 385-392.

Zhongguo Li and Maosong Sun.2009. Punctuation as Implicit Annotations for Chinese Word Segmentation. In Proceedings of Computational Linguistics. Pages 505-512.

Zhongguo Li.2011. Parsing the Internal Structure of Words: A New Paradigm for ChineseWord Segmentation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Pages 1405-1414.

Hwee Tou Ng and Jin Kiat Low. 2004. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? In Proceedings of the Empirical Methods in Natural Language Processing Conference.

Kun Wang, Chengqing Zong and Keh-Yih Su.2010. A Character-Based Joint Model for Chinese Word Segmentation. In Proceedings of the 24th International Conference on Computational Linguistics.Pages 1173-1181.

Yue Zhang and Stephen Clark. 2007. Chinese segmentation with a word-based perceptron algorithm. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Pages 840-847.