利用扩展标记集的词结构分析*

孙静，方艳，丁彬，周国栋

（苏州大学计算机科学与技术学院，江苏 苏州 215006）

摘要：本文给出了一种与传统分词不同的词法分析选择，即分析词的内部结构，提出了一种利用扩展标记集来实现词结构分析的方法。首先阐述了词的内部结构特点，然后把结构中的前后缀视为特殊的词，通过识别出每一个词的前后缀来识别词的内部结构。把词内部结构识别问题转换成序列标注问题，通过扩展标记集，采用 CRF 模型来实现词的内部结构分析。最终实验表明，无论是在总体性能上，还是在各层结构的识别上都取得了较高的准确度。

关键词：扩展标记集；词结构分析；前后缀；序列标注问题

中图分类号：TP391　　**文献标识码：**A

Words Structures Analysis Method by Extending the Word Tag Set

SUN Jing, FANG Yan, DING Bin, ZHOU Guodong

(School of Computer Sciences and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: This paper proposes a different choice of lexical analysis, which analyzes the internal structures of words, then presents a words structures analysis method by extending the word tag set. First, we describe the characteristics of the internal structures of words, and then regard the prefixes and suffixes within words structures as special words, finally, we identify the internal structures of words through the identifying of prefixes and suffixes. We convert the issue of identifying the internal structures of words to the sequence tagging problem, then adopt the CRF model to realize the words structures analysis using extending the word tag set. The experiment shows that they achieve a higher accuracy both on overall performance and on the identification of each layer structure.

Key words: Extend the word tag set；Words structures analysis；Prefixes and suffixes；Sequence tagging problem

1 引言

在中文信息处理领域中，词法分析的一般做法是通过分词来划定词和短语的边界，从而使汉语的后续处理过程，如语义分析、句法分析等，能够跟英语等西方语言基本一致。然而，汉语中单字词与语素之间、词与短语之间的界限比较模糊。许多情况下甚至连语言学家也难以确定某些语言单位是语素、词还是短语，这就导致实践中人工标注的分词语料存在严重的不一致性，这种不一致性无疑会制约汉语的后续处理工作。

分词语料的不一致性不仅体现在不同语料库间分词标准不同，而且同一语料库中的分词标准也不一致。例如，在 PKU 语料库中，“总教练”被切分为“总”和“教练”两个词，而“总书记”却是一个词。但是，“总教练”和“总书记”都有相同的结构，即前缀“总”加名词构成，它们可以表示成具有内部结构的标注形式：“[总 书记]”和“[总 教练]”。另一方面，不同的自然语言处理应用对词的粒度大小也有不同的需求，单一的分词标准难以满足各种要求。

由此可见，要解决分词标准的不一致性以及应用的不同需求，一个有效的方法就是分析

*收稿日期：2014 - 06 - 15　　定稿日期：2014 - 07 - 28

基金项目：自然科学基金青年项目（61202162）；教育部博士点基金新教师类课题（20123201120011）；国家 863 计划前沿技术研究类项目（2012AA011102）

作者简介：孙静（1986—），女，博士研究生，主要研究领域为自然语言处理。Email:sj44581@163.com；方艳（1989—），女，硕士研究生，主要研究领域为自然语言处理。Email: 20114227021@suda.edu.cn；丁彬（1991—），女，硕士研究生，主要研究领域为自然语言处理。Email: 20124227006@suda.edu.cn；周国栋（1967—），男，教授，博士生导师，主要研究领域为自然语言处理。通讯作者，Email: gdzhou@suda.edu.cn。

词的内部结构。对词进行结构分析，是为了提供一种与传统分词不一样的词法分析选择，它更加符合汉语词法及句法边界模糊的事实，有利于发挥词法分析在实际应用中的作用。

汉语词语结构的自动分析方面，Wu[1]提出了一种基于规则的词语内部结构分析方法。赵海[2]对目前分词范式所存在的问题进行了分析，并提出了利用汉字之间的依存关系对词进行表述。Zhang 等[3]提出汉语中绝大多数的词都有语法结构，分析词的结构对分词、词性标注及句法分析精确度的提高都有帮助。针对目前的分词规范在理论上和实际运用中的不足，Li[4]从理论和应用两个角度论证了分析词语结构的必要性，并提出了汉语词法与句法统一分析的方法。方艳等[5]提出了一种基于层叠 CRF 模型的词结构分析方法，对于无结构词的输出等同于传统的分词输出，对于有结构词的输出结果中不仅有词的边界信息，还包含词的内部结构信息。该方法包括底层模型和高层模型两部分。底层模型主要实现汉语字串的细粒度分词，该任务与传统分词相同，仅在词的粒度上有所区别。高层模型是对经细粒度分词后的词序列使用 CRF 模型来识别词的内部结构。实验结果表明，该方法对词结构的识别取得了较高的准确率，总体性能达到了实用水平。

虽然方艳等[5]方法的总体性能达到了实用水平，但此方法需要先进行细粒度分词，然后在分词基础上进行词结构分析，因此最终结果受分词结果好坏的影响。通过统计方艳等[5]所用语料库，发现语料中 99%含结构的词都是结构简单的词（小于等于两层结构），并且有 97%的词结构可以表示成 “词根+后缀”或者 “前缀+词根”的形式。针对此特点，可以把前后缀视作特殊的词，通过识别出每一个词的前后缀来识别词的内部结构。这样可以把词内部结构的识别问题转换成序列标注问题。因此本文提出了一种一步到位实现词结构自动分析的方法，即通过扩展标记集来实现词结构的自动分析，避免细粒度分词所带来的错误传递。

本文第二部分简要介绍词结构分析任务定义；第三部分介绍扩展标注集设置；第四部分介绍实验设置；第五部分对实验结果进行分析与比较；最后总结全文。

2 词结构分析任务定义

汉语的词可以分为单纯词和离合词，单纯词仅有一个语素，谈不上内部结构。而复合词顾名思义是由一些词复合形成的，它含有多个语素，都有内部结构。本文分析的含结构的词并非所有的复合词，因为从自然语言处理角度来看，有些复合词的结构并不需要分析，例如“研究”虽为复合词，但自然语言处理应用系统一般不需要对其内部结构进行分析。本文所指的有结构的词界定如下：

1．词中包含中心成分，并且该结构具有能产性[1]。例如“工程师”，其中的“师”为中心成分，并且“师”字是能产的。

2．具有中心成分但不满足能产性的情形，如果该中心成分对应的所有词构成平行的语义类别，则也作为有结构的词。例如，“大学、中学、小学”，虽然表示“学校”的“学”字不具有能产性，但标注这组词语结构以后，类似“大中小学”的结构分析就能与前面三个词语的结构分析统一起来。

3．不具有中心成分的“离心结构”，如果具有能产性并且产生的词句法功能一致，也是本文所指的有结构的词。例如“反革命”中的两个成分“反”和“革命”都不是整个词的中心成分，但由于“反+名词”这种结构具有能产性（反贪，反帝，反华，反浪费，反盗版等），本文仍将其作为有结构的词。

4．汉语的人名是一类特殊的含结构的词，每个人名都包含姓与名，故本文对汉语人名的结构也作了分析。

本文将词结构（除人名外）中能产的部分（以及第二种情况下的中心成分）称为前缀或后缀（不同于语言学上的前后缀）。例如，在词结构“[总 经理]”中，“总”称为前缀，“经理”称为词根。本文的前后缀不仅限于单个汉字，也可能是多个汉字，如：主义，阶级。词的结

[1]能产性指由某种规则能产生大量新词，也可指某语言单位能产生大量更大的单位，但本文中的能产性更偏向于后者。

构可能是一层，也有可能是多层的，如 “总工程师”具有两层结构，如图 1 所示，本文用方括号表明了词的内部结构，一层括号表示一层词的结构，即图 1 的结构表示为“[总 [工程 师]]”。表 1 列出了具有不同形式的词结构，从中可以看出，汉语中词的内部结构纷繁复杂。

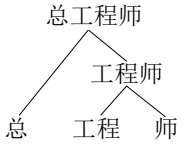

图 1 “总工程师”的结构

表 1 词的内部结构举例

人名		[江 泽民]，[兰 红光]，[王 元]
前缀	序数词	[第 一]，[第 三十五]，[第 108]
	代词	[本 镇]，[各 族]，[该 校]
	名词	[副 教授]，[软 组织]，[代 总理]
	动词	[超 额]， [抗 旱]，[反 党]
	形容词	[不 景气]，[最 高]，[易 燃]
后缀	地名	[黑龙江 省]，[镇江 市]，[加利福尼亚 州]
	时间	[一九一四 年]，[十一 月]，[清 朝]
	名词	[支持 者]，[实习 生]，[收割 机]
	动词	[信息 化]，[冲 走]，[闯 入]
	处所词	[湖 边]，[门 前]，[校 内]
多层结构		[总 [工程 师]]， [[古 [人 类]] 学]， [[北京 市] 人]，[[[无 政府] 主义] 者]

本文中词结构自动分析的任务不仅（以空格）分隔出一个句子中的词，而且给出词的内部结构，并且这种结构可能是嵌套的。如在下列句子中：

1．林志浩是总工程师

2．林志浩 是 总工程师

3．林 志浩 是 总 工程师

4．[林 志浩] 是 [总 [工程 师]]

其中句 1 是未经分词的原始句子，句 2 和句 3 是两种不同的分词结果。显然，就分词的颗粒度而言，句 2 和句 3 是不同的。句 4 是本文的词结构分析所要输出的结果，它不仅包含了各种可能的分词情况，而且用方括号表明了词的内部结构，由此可见，词结构分析符合了汉语中词与短语界限不清的特点，并且该种词法分析可以很好地兼容不同的分词标准，不同的应用可以根据需要提取不同粒度的词，克服了目前分词中所存在的问题。

3 扩展标记集设置

运用序列标注法来分析词的内部结构，就是把识别词内部结构的过程视为字（这里的“字”不一定只表示汉字，还可以是外文字母，阿拉伯数字，标点符号等字符）在字符串中的标注问题。由于在分析词内部结构这个问题上，词有词根与词缀之分，故需要将词根与词缀分别设立构词位置才能识别词的内部结构。

词根是无内部结构的词，它的识别等同于传统的分词，故本文借鉴基于字的序列标记法使用的标记集来识别词根。用序列标注法来分词所使用的标记方法有多种：如基于词边界的 0/1 标记法[6]、2 词位标记法[7]、4 词位标记法[8]、6 词位标记法等[9]。不同的标记方法都有各自的优缺点，标记个数越多的，如 6 词位标记，对于识别长度较大的词效果越好，但是所需时间复杂度和空间复杂度越高；标记个数越少的，时间复杂度和空间复杂度较低，但需

要设计更为复杂的特征模板来提高分词的准确率。综合考虑，本文对词根的识别选择了 4 词位标记集，即 B（词首）、M（词中）、E（词尾）和 S（单独成词）。

由于词结构中词缀有前缀和后缀之分，故本文首先增加三个构词位置：P（前缀）、F（后缀词首）、G（后缀词尾）。之所以有后缀词首和后缀词尾之分，是因为在所有的后缀中，有些后缀不是单字，而是双字（无三字和三字以上的后缀，无多于一字的前缀），如主义、阶级。故分别用 F 和 G 表示二字后缀的首字和尾字。若后缀是单字，则直接用 F 表示。汉语中的人名是一类特殊的词，它包括姓和名两部分。本文对人名作特殊处理，把人名中的姓专门设定一个词位 N，因此，共增加了四个标记，如表 2 所示。根据以上分析，句子（a）可以直接表示成如（b）所示的逐字标注形式：

表 2 增加的标记及意义

标记	N	P	F	G
意义	人名中的姓	前缀	后缀首字	后缀尾字

（a）**[游泳 队] [总 教练] [陈 运鹏] 赴 珀斯 观赛**

（b）**游/B 泳/E 队/F 总/P 教/B 练/E 陈/N 运/B 鹏/E 赴/S 珀/B 斯/E 观/B 赛/E**

对（b）式还原的时候需分两步进行，第一步首先还原所有词根，以及二字后缀，即将所有是 B、M、E、S 和 G 的标记还原。如（b）句经第一步还原后的结果如（b’）所示。

（b’）**游泳 队/F 总/P 教练 陈/N 运鹏 赴/S 珀斯 观赛**

第二步对词缀还原。还原词缀标记时，若标记为 P，则将该字与其后一个词结合成形如**“词→前缀+词”**的结构。若标记为 F，则将该字与其前一个词结合成形如**“词→词+后缀”**的结构（人名的还原如同前缀还原）。例如在（b’）中，“队”的标记是 F，因此将“队”与其前一个词“游泳”合并成“[游泳 队]”。还原所有的标记后便可得到（a）式结构。但是这样的标记设置也会出现问题，比如句子（c）：

（c）**[林 志浩] 是 [总 [工程 师]]**

（d）**林/N 志/B 浩/E 是/S 总/P 工/B 程/E 师/F**

根据新增的标记集，（c）式的标记序列如（d）所示，但还原（d）式时，在进行第二步词缀还原时，对于词“总工程师”，在前缀“总”和后缀“师”哪个优先与“工程”合并的问题上将产生歧义。即（d）式还原的结果可能是（c）式，也有可能是（g）式。

（g）**[林 志浩] 是 [[总 工程] 师]**

对于这种问题，本文将其定义为前后缀优先合并歧义，即当词根的左右两边同时出现前缀和后缀时，词根首先应该跟前缀结合还是应该先跟后缀结合就会产生歧义。为了解决这个问题，本文又增加了两个标记：T 和 H。T 也表示前缀标记，但此前缀与 P 的不同点在于 T 一般与 F 同时出现在有前后缀优先合并歧义的词中，并且表示后缀优先合并。例如增加标记“T”后，“总工程师”的标记将是“总/T 工/B 程/E 师/F”，此时由于 T 的优先级低于 F，“工程”先与“师”合并成“[工程 师]”，再与“总”合并成“[总 [工程 师]]”。H 也表示后缀标记，但 H 与 F 的不同在于 H 一般与 P 同时出现在有前后缀优先合并歧义的词中，并表示前缀优先合并。例如增加“H”后，“低收入者”的标记将是“低/P 收/B 入/E 者/H”，这样便很容易得到正确的结构“[[低 收入] 者]”。

因此，本文所设定的标记集中，词根部分采用的是 4 词位标记，即 B、M、E、S；词缀部分增加了 6 个标记，如表 3 所示。之所将前缀和后缀分别设立两种不同的标记，是为了解决前后缀优先合并歧义问题。本文规定 P 一般与 H 同时出现在有前后缀优先合并歧义的词中，并且表示前缀优先合并；而 T 一般与 F 同时出现在有前后缀优先合并歧义的词中，并且表示后缀优先合并。例如改用此种标记后，对于有前后缀优先合并歧义的词，如图 2 中的两种结构，若前缀优先合并，则前后缀的标记如图 2（a）所示；若后缀优先结合，则前后缀标记如图 2（b）所示。

表 3 增加的标记及意义

标记	N	P	T	F	H	G
意义	人名中的姓	优先结合的前缀	后结合的前缀	优先结合的后缀首字	后结合的后缀首字	后缀尾字

（a）前缀优先合并　　（b）后缀优先合并

图 2 结构标记示例

4 实验设置

当标记集确定了之后，可以很方便的运用序列标注法来实现词的内部结构分析。考虑到 CRF 在序列标注任务中表现出的优越性，故本文采用的分类器为 CRF 分类器。

4.1 语料介绍

本文使用文献[5]中标注的语料，该语料严格按照本文第二节中有结构词的范围界定，对 PKU1998 年 1 月的《人民日报》语料中所有含结构的词进行二次标注，如图 3 所示。

```
中共中央 [总 书记] 、 国家 主席 [江 泽民]
（ [一九九七 年] [十二 月] [三十一 日] ）
[１２ 月] [３１ 日] ， 中共中央 [总 书记] 、
国家 主席 [江 泽民] 发表 [１９９８ 年] [新 年]
讲话 《 [迈 向] 充满 希望 的 [新 世纪] 》 。 （
[新华 社] [记 者] [兰 红光] 摄 ）
[同胞 们] 、 [朋友 们] 、 [女士 们] 、 [先生
们] ：
```

图 3 人工标注的语料

在实验中将已标注的 PKU 语料分成两部分，分别作为训练语料和测试语料，其规模如表 4 所示。

表 4 训练语料及测试语料的规模

类型	句子数	词次	词数
训练语料	前 15,588 句（80%）	898,620	49,315
测试语料	后 3,896 句（20%）	211,327	21,897

4.2 预处理

与分词的过程一样，在进行 CRF 模型训练之前，需对语料中的一些特殊的字符作特殊处理。因为在语料中，时间词、字母串、数词往往都是未登录词，而对未登录词的识别一直是中文分词中的难点。若不对语料作任何处理，这些特殊的词就很容易识别错误。为了提高识别准确率，我们引用“分类泛化”思想来对语料作适当的处理，具体方法是将训练语料和测试语料中的四类不同字符分别替换成四个特殊字符：

1. 所有的英文字母，如：a、A、b、B 等，都替换成字母‘A’；
2. 所有阿拉伯数字，如：1、2、3 等，都替换成字母‘B’；
3. 所有中文数字，如：一、二、三、百、千、万等，都替换成字母‘C’；
4. 所有标点符号，如：。、!、？等，都替换成字母‘D’。

保留测试语料中所有被替换的字符，然后用替换后的训练语料来训练 CRF 模型。当测试结束后，再把测试语料中被替换掉的字符还原成原来的字符。

4.3 特征设置

在使用 CRF 模型时，特征模板将对 CRF 模型具有表征意义的上下文特征按照共同属

性分类，它的选择至关重要。本文在实验过程中经过多次实验对比，发现传统的 5 字窗口宽度的特征效果并不理想，采用 4 字窗口宽度的特征效果较好。因此，本文最终选择了 4 字窗口宽度，采用两组字特征，即当前字本身及其前后两个字构成的位置特征，如表 5 所示。其中，Unigram 是单字符特征，Bigram 双字符结合作为二元组合特征。在特征模板中，C_0 指当前字符，C_{-1} 指当前字符的前一个字符，C_1 指当前字符的下一个字符，以此类推。

表 5 CRF 词结构分析特征模板

特征类型	特征模板
Unigram	C_{-1}，C_0，C_1
Bigram	$C_{-1}C_0$，C_0C_1，C_1C_2

4.4 评测标准

本文使用文献[5]中提出的正确率，召回率和 F-值作为评测的三个指标，计算公式分别如下：

$$\text{正确率（Precision）}=\frac{\text{分析得到的正确词语结构的个数}}{\text{分析得到的词语结构的总数}}\times 100\% \quad (1)$$

$$\text{召回率（Recall）}=\frac{\text{分析得到的正确词语结构的个数}}{\text{标准结果中的词语结构个数}}\times 100\% \quad (2)$$

$$F-\text{值}=\frac{2\times \text{Precision}\times \text{Recall}}{\text{Precision}+\text{Recall}}\times 100\% \quad (3)$$

5 实验结果与分析

5.1 实验结果

利用扩展标记集来实现词内部结构的自动分析，就是把前后缀视作特殊的词，通过识别出每一个词的前后缀来识别词的内部结构。该方法所得到的词结构分析的最终结果如表 6 所示。其中，零层结构考察模型识别无结构的词和词结构中的最底层词的性能；一层结构考察模型识别含有一层结构词的性能；两层及两层以上结构考察模型识别大于等于两层结构词的性能。

从表 6 中可以看出，采用扩展标记集实现词结构分析方法无论是在总体性能上，还是在各层结构的识别上都取得了较高的准确度。其中总体性能达到了 95.1%，零层结构的性能达到了 95.1%。虽然随着结构层数的增加，性能有所下降，但是下降的幅度不是很大。两层及两层以上的结构在语料中只占不到 11.5%的比例，性能却达到了 82.2%。高层结构性能下降的主要原因有两个，一是训练数据较少所导致的稀疏性问题；二是存在某些特殊短语结构，系统不能正确的识别，该部分将在后续 5.4 节作详细的讨论。

表 6 实验结果

结构类型	准确率	召回率	F 值
总体性能	95.4	94.9	95.1
零层结构	95.2	95.0	95.1
一层结构	92.6	90.3	91.4
两层及两层以上结构	86.6	78.3	82.2

分析实验结果后发现，采用扩展标记集实现的词结构分析能克服文献[5]中基于层叠 CRF 模型的词结构分析结果中的许多错误。例如，本方法的实验结果中能正确识别“新华文摘”的正确结果是“新华 文摘”，以及在“一名热爱海的看海者”中也不会错误的把“[热爱 海]”作为词语结构。但是，对于一些多产的前后缀也有许多无法避免的错误。例如在“一副冷眼歧视的面孔”中，系统错误的把量词“副”识别为前缀，即将“[副 冷眼]”作为一个结构。此外，某些短语结构无法被本方法识别。例如，“[[书法 爱好] 者]”被错误的识别为“书法 [爱好 者]”，“[人民 [日 报]]”被错误的识别为“人民 [日 报]”。

本文没有进一步给出和 sighan 比较的实验结果，是因为目前的 sighan 没有词结构的分析这一项，若与分词这一任务相比，两者虽都属于词法分析，但是任务定义不同，所以可比性有限。且我们的词结构分析结果跟目前最好的分词结果相比，虽有所差距，但在同一水平上。我们的下一步工作就是寻找更进一步提高词结构性能的方法。

5.2 考察测试语料对实验结果的影响

为了考查语料对实验结果的影响，本文设置了五种不同的语料来考察利用扩展标记集实现词结构分析系统的性能。由于目前对词结构的相关研究并不多，没有更多可供使用的公测语料，因此我们的实验都是在人工标注的 PKU 语料上进行。我们把人工标注的 PKU 语料平均分成五份，分别将其中的每一份作为测试语料，剩下的四份作为训练语料，共产生五种不同的测试语料和训练语料。

图 4、5、6、7 分别是系统在这五种不同的语料下获得的词结构分析的总体性能、零层结构的性能、一层结构性能、两层及以上结构的性能。每张表的横纵轴表示的意义相同，其中横轴表示语料种类，纵轴表示性能值。并且本文从准确率，召回率和 F 值三方面来考察系统的性能。

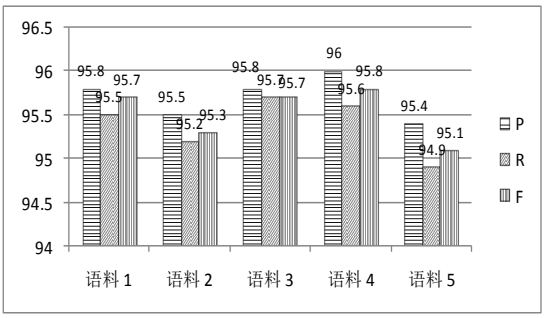

图 4 五种语料下词结构总体性能

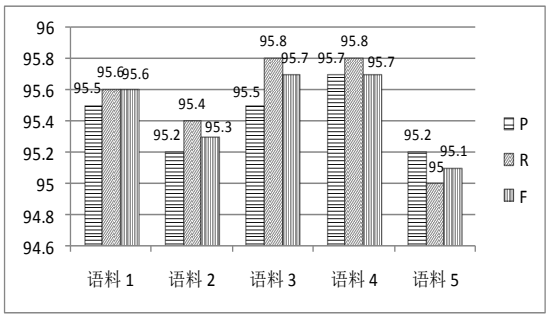

图 5 五种语料下零层结构性能

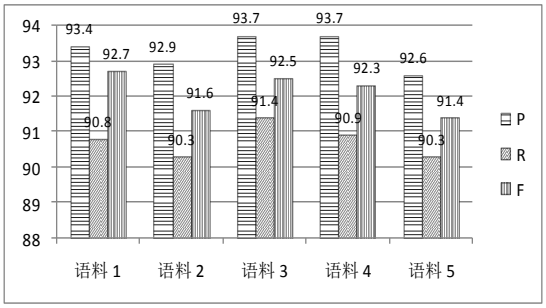

图 6 五种语料下一层结构性能

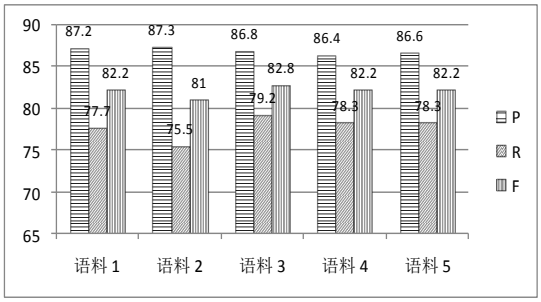

图 7 五种语料下两层及以上结构性能

从这四张表中可以看出，虽然不同的语料产生的最终结果有所差别，但差别较小。词结构分析的总体性能 F 值最高是 95.8%，最低是 95.1%，五个结果的平均值是 95.5%。零层结构的性能 F 值最高是 95.7%，最低是 95.1%，五个结果的平均值是 95.5%。一层结构的性能 F 值最高是 92.7%，最低是 91.4%，五个结果的平均值是 92.0%。两层及两层以上结构的性能 F 值最高是 82.8%，最低是 81.0%，五个结果的平均值是 82.1%。从结果中可看出，系统对于不同的语料，性能差别较小，总体性能较稳定。并且对于两层及两层以上的结构，系统的性能平均 F 值能达到 82.1%。这说明随着词语层次数的增加，虽然系统识别能力在不断的降低，但降幅相对较小，系统对高层次的结构具有较高的识别能力。

5.3 与其他系统的比较

本文将文献[5]中基于层叠 CRF 模型的词结构分析系统与本文的词结构分析系统作了比较。所采用的对比数据是在五种不同语料下两个系统所取得的性能平均值，并分别从总体性能 F 值，零层结构性能 F 值，一层结构性能 F 值，两层及两层以上结构性能 F 值四方面

作了对比。

表 7 是两个系统在五种语料下测试的性能平均值。从表中可看出，利用扩展标记集实现的词结构分析系统无论是在总体性能，还是在各层次结构的识别性能上，都比基于层叠 CRF 模型的词结构分析系统更优秀。其中，总体性能提高了 0.8%，零层结构的性能提高 0.7%，一层结构的性能提高了 1.4%，两层及两层以上结构的性能提高了 18.8%。由于零层结构相当于细粒度分词的结果，因此从零层结构的对比中可以看出，增加了标记集后的序列标注法实现的细粒度分词，相比于四标记的细粒度分词，性能有所提高，并且基于扩展标记集实现的词结构分析对于两层及两层以上词结构的识别性能提升幅度较大。

表 7 两种模型在不同结构层次上的性能对比

	基于层叠 CRF 模型的词结构分析系统	利用扩展标记集实现的词结构分析系统
总体性能	94.7	95.5
零层结构	94.6	95.5
一层结构	90.6	92.0
两层及两层以上结构	63.3	82.1

5.4 不能处理的结构

虽然运用扩展标记集实现的词结构分析系统相比于基于层叠 CRF 模型的词结构分析系统，性能有很大的提高，特别是在识别两层及两层以上词的结构，性能提高了 18.8%。但是利用扩展标记集实现词结构分析时，在理论上仍有两种结构是不能被系统正确识别的。

第一种是短语结构。虽然本文所分析的是词的内部结构，但是一些特殊的短语结构也在本文的分析范围之内。由于这些短语结构与其他词具有相同的句法功能和语义类别，为了达到语料标注的一致性，本文对这类特殊的短语也作了结构分析。这种短语结构不能表示成**“词→词根、词→词+后缀、词→前缀+词”**中的任何一个，或者是不能完全表示成这些结构，即不满足上下文无关文法。例如“身体健康者”的结构“[[身体 健康] 者]”。虽然该结构中的外层结构相当于**“词→词+后缀”**结构，即由词“身体健康”加后缀“者”构成，但是其内层的结构却不同于**“词→词根、词→词+后缀、词→前缀+词”**中的任何一个。因此在本文的识别结果中，能识别出“者”是一个后缀，“身体”和“健康”是两个词，但却不能正确的识别整个结构。因为在标记还原的时候，系统将“者”与其前一个词构成一层结构，所得到的最终结构为“身体 [健康 者]”。类似的情况如“大中小学”的结构“[[大 中 小] 学]”，也不能被正确的识别。虽然这类特殊的短语结构在本文中不能被正确的识别，但是由于这类结构的数量较少（在语料中共有 830 个，占所有有结构词总数的 2.6%），故对实验结果产生的影响不大。

第二种不能被识别的结构有两类，如图 8 所示。其中图（a）是双前缀加一个后缀的情况，之所以不能识别，是因为词根首先与前缀 1 结合，再与后缀结合。根据我们设定的标记集含义，此时前缀 1 的标记必为 P，后缀的标记为 H；但此时前缀 2 无论标记是 P 或是 T，都会与后缀产生前后缀优先合并歧义。图（b）表示的是一个前缀加双后缀的情况。该情况不能识别的原因是如果词根首先与后缀 1 结合再与前缀结合，那后缀 1 的标记必为 F，且前缀的标记是 T。但此时后缀 2 无论标记是 F 还是 H，都会与前缀产生前后缀优先合并歧义。如“古人类学”、“副部长级”。虽然在理论上这两种结构的是无法被正确识别的，但是由于在本文标注的语料中这两种结构数量较少，其中，图（a）所示的情况在语料中并未出现，图（b）所示的情况在所有标注的有结构的词中共有 16 个，因此对最终的性能影响较小。

（a）双前缀　　（b）双后缀

图 8 不能处理的结构举例

6 小结

本文研究了利用扩展标记集实现词内部结构的分析。由于本文所分析的词有 97%是结构简单的词，并且结构能用上下文无关文法表示，因此可以把前后缀视作特殊的词。通过识别出每一个词的前后缀来识别词的内部结构，并且可以很方便的运用序列标注法来实现。实验结果表明，该方法进行的词结构分析在总体性能上相比于基于层叠 CRF 模型的词结构分析有显著的提高。经过比较零层结构的准确率可知，增加了标记集后的序列标注法对实现细粒度分词也有很大程度的性能提高。虽然本系统在理论上有两类特殊的结构不能被识别，但由于在语料中这两种结构所占的比例极小，故对系统的最终效果影响较小。

参考文献

[1] Wu A D. Customizable segmentation of morphologically derived words in Chinese[J]. International Journal of Computational Linguistics and Chinese Language Processing, 2003, 8(1): 1-27.

[2] Zhao H. Character-level dependencies in Chinese: Usefulness and learning[C]//Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2009: 879-887.

[3] Zhang M S, Zhang Y, Che W, et al. Chinese parsing exploiting characters[C]//51st Annual Meeting of the Association for Computational Linguistics. 2013.

[4] Li Z G. Parsing the internal structure of words: a new paradigm for Chinese word segmentation[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 1405-1414.

[5] 方艳, 孙静, 丁彬，等. 基于层叠CRF模型的词结构分析[J]. 中文信息学报. 已录用.

[6] Li S, Huang C. Word Boundary Decision with CRF for Chinese Word Segmentation[C]. Proceedings of PACLIC-2009. 2009:726-732.

[7] Peng F, Feng F, McCallum A. Chinese segmentation and new word detection using conditional random fields[J]. 2004.

[8] Xue N W, Converse S P. Combining classifiers for Chinese word segmentation[C]//Proceedings of the first SIGHAN workshop on Chinese language processing-Volume 18. Association for Computational Linguistics, 2002: 1-7.

[9] Zhao H, Huang C N, Li M. An improved Chinese word segmentation system with conditional random field[C]//Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney: July, 2006, 1082117.