一种新的错误驱动学习方法在中文分词中的应用

夏新松　肖建国
（北京大学计算机科学技术研究所　北京 100084）

摘　要　中文分词应用中一个很重要的问题就是缺乏词的统一性定义。不同的分词标准会导致不同的分词结果，不同的应用也需要不同的分词结果。而针对不同的分词标准开发多个中文分词系统是不现实的，因此针对多种不同的分词标准，如何利用现有的分词系统进行灵活有效的输出就显得非常重要。本文提出了一种新的基于转换的学习方法，对分词结果进行后处理，可以针对不同的分词标准进行灵活有效的输出。不同于以往的用于分词的转换学习方法，该方法有效利用了一些语言学信息，把词类和词内结构信息引入规则模板和转换规则中。为了验证该方法，我们在4个标准测试集上进行了分词评测，取得了令人满意的效果。

关键词　中文分词，规则模板，词类，词内结构，基于转换的学习(TBL)

A New Error-driven Learning Approach for Chinese Word Segmentation

XIA Xin-Song　XIAO Jian-Guo
(Institute of Computer Science and Technology of Peking University, Beijing 100084)

Abstract　A well known problem for Chinese word segmentation(CWS) is that we can not have a unique definition of words. Different standards may result in different word segmentation outputs. It is unrealizable to develop different CWS systems according to different applications or standards, so it is significantly important to flexibly adapt segmentation outputs towards different standards or applications using existing CWS system. The paper presents a linguistically enriched transformation-based learning approach for performing CWS adaptation as a postprocessor. Different from other transform-based learning used in CWS, the approach utilizes some linguistics information, and introduces word class and word internal structure to rule templates and transformations. The performance of the approach is evaluated on four different test sets, which represent four different standards. It turns out to be comparable to several state-of-the-art approaches which perform Chinese word segmentation based on single standard.

Keywords　Chinese word segmentation, Rule template, Word class, Word internal structure, Transformation-based Learning(TBL)

1　引言

众所周知，英文是以词为单位的，词和词之间是靠空格隔开。而中文是以字为单位，句子中所有的字连起来才能描述一个意思。例如，英文句子 I am a student，用中文表述则为：我是一个学生。计算机通过空格可以很容易理解 student 是一个词，但是不能很容易理解学、生两个字合起来才表示一个词。把中文的汉字序列切分成有意义的分隔单元，让计算机更容易理解，这就是中文分词的主要目的。

中文分词是中文信息处理的基础，也是很多中文应用首先要面对的问题。如语音合成、机器翻译、信息检索、自动分类、自动摘要、自动校对等等，都需要用到中文分词。由于中文分词没有统一的定义，新词的不断涌现，分词中存在交集型歧义和组合型歧义，这些都导致了中文分词是一个富有挑战性的问题。从众多的关于分词的研究中我们可以看到，对于中文分词的最大障碍就是如何正确识别在词典之外的词(OOV)。

此外，分词的最终切分结果取决于不同的实际应用。在不同的分词标准或不同的应用中，汉字最终的切分结果可能会有比较大的差异。在信息检索中，人们会希望分隔单元尽可能大一些；而在机器翻译中，人们则希望分隔单元尽可能小一些[1]。

不同的分词标准会导致不同的分词结果，而且不同的应用也需要不同的分词结果。而针对不同的分词标准开放多个中文分词系统是不现实的，因此针对多种不同的分词标准，如何利用现有的分词系统进行灵活有效的输出就显得非常重要。

基于上述考虑，我们提出了一种新的基于转换的错误驱动学习方法(TBL)来解决中文分词问题。不同于以往的用于中文分词的 TBL 方法，该方法能有效利用一些语言学信息，把词类和词内结构信息引入规则模板和转换规则中，使所学的规则更有效，也更具代表性。我们在 SIGHAN 举办的第一届中文分词大赛[1]中提供的 4 个语料集上都进行了开放测试，结果表明，我们分别在 PK、CTB、HK 和 AS 语料集上排在 2、3、2、1 位。不同的语料集代表着不同的分词标准，尽管不同标准之间词的定义差别可能很大，我们的系统仍然可以与那些基于某个特定标准而设计的优秀的分词系统相媲美。这充分证明了我们的学习方法是有效的。

夏新松　博士研究生，主要研究方向是中文分词处理、文本检索和知识问答；**肖建国**　教授，博士生导师，研究领域是网络与数据库技术应用、彩色图像处理、中文信息处理、数字通信。

2 相关工作

2.1 中文分词方法

关于中文分词处理已经提出了很多方法[2,5]，这些方法可以大致分为基于词典的方法和基于统计的方法。目前许多优秀的分词系统都是二者的有机结合。

对于基于词典的分词方法，它是按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行匹配。若在词典中找到某个字符串，则匹配成功(识别出一个词)。比较常见的有前向最大匹配算法(FMM)、后向最大匹配算法(BMM)等。通常这些方法还须配合启发式规则来处理分词歧义。这些方法的效果在很大程度上依赖于所使用的词典的覆盖率。

对于基于统计的分词方法，从统计学的角度判断相邻的字符是否可以合成词。从形式上看，词是稳定的字的组合，因此在上下文中，相邻的字同时出现的次数越多，就越有可能构成一个词。字与字相邻共现的频率或概率能够较好地反映成词的可信度。但这种方法也有一定的局限性，会经常抽出一些共现频度高、但并不是词的常用字组，例如“这一”、“之一”、“有的”、“我的”、“许多的”等，并且对常用词的识别精度差，时空开销大。

近年来，人们开始把一些机器学习的方法用在中文分词中，文[6，7]把基于转换的错误驱动算法(TBL)用于中文分词中，并取得了一定程度的成功。文[8]把分词问题看作是基于字符的标注问题，采用最大熵模型来进行中文分词，也取得了比较好的效果，而且该方法对识别新词比较有效。

此外，文[9,10]把中文分词视为句子分析的副产品，所产生的句法树的叶节点就是最后的分词结果，这个方法也取得不错的结果。

为了更好地处理不同标准或应用造成的分词输出结果的差异，在文[11]中，Wu 认为每个词都可以被看作是一棵树，而分词的最终输出只是对该树进行特定的切割，而切割方式是与分词标准相对应的。该方法在 SIGHAN 的语料中进行了评测，取得了非常好的成绩，但它需要耗费大量的时间来了解不同标准之间的差异。在我们的方法中将借鉴这种树结构的思想，并用自动学习的方式来实现树的切割。

2.2 基于转换的学习方法

Brill 提出的基于转换的错误驱动学习方法(TBL)是一种符号式的机器学习方法，在很多自然语言应用中都得到了成功的应用，如词性标注、短语识别等[12]。Palmer 把它扩展到了中文分词的应用中，实验表明 TBL 既可以作用于单独的中文分词系统，也可以作为后处理引擎对中文分词进行后续处理。尽管 Palmer 在实验中选择的规则模板是非常简单的词例化模板，不必花费大量的语言工程，但仍然非常有助于提高其他分词算法的性能[6]。Hockenmaier 又进一步扩展了 Palmer 的工作，采用了一些与文[6]不同的转换规则模板，分词性能略有提高。他同时还验证了如果在模板中加入一些简单的语言学信息，会获得更好的性能[7]。

TBL 学习方法需要 3 个重要组成部分：未切分的训练语料、参照语料、规则模板。其中参照语料作为标准答案，是按照特定分词标准切分的语料。而规则模板则表示一个规则集，它限定了学习中可能会用到的转换规则的选择空间。学习过程如图 1 所示。

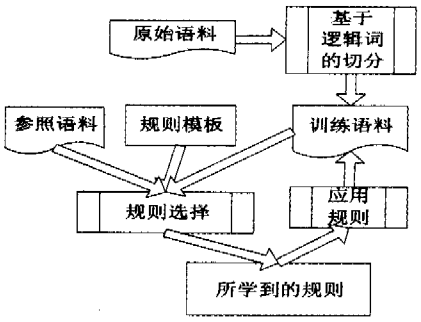

图 1 基于转换的学习过程

1. 首先，利用初始分词系统对原始语料进行切分，形成训练语料；
2. 根据规则模板构造所有可能的转换规则，作用于训练语料，产生新的标注语料。通过比较参照语料和标注语料，选择能最大减少分词错误数的那条转换规则；
3. 一旦选取某条转换规则，就把它作用于当前的训练语料中，进行重新标注，形成新的训练语料；
4. 重复步骤 2,3，直到分词错误数不再减少；
5. 输出最终的转换规则序列。

3 基于词类的 TBL 方法(CTBL)

由于规则模板限定了 TBL 方法所要搜索的转换规则的空间，因此选择合适的规则模板至关重要。通过分析 Palmer、Hockenmaier 的实验，并在做了一些尝试性实验的基础上，我们认为，由于未知词和分词歧义等原因，仅使用词例化的规则模板，采用 TBL 方法的分词系统，分词精度通常不会超过 89%，这与目前比较好的分词系统的分词精度(95%左右)是有一定差距的。另外，同其他机器学习算法一样，TBL 同样面临由于数据稀疏而造成的过学习问题。此外，Hockenmaier 的实验也表明在 TBL 中加入适当的词法和语法信息有助于提高分词性能。

基于上述考虑，我们提出了基于词类的 TBL 方法：CTBL 方法。该方法通过引入一些词法和语法信息，使所学的规则更有效、更具代表性。

3.1 词类和词内结构

在很多关于中文分词的文章中都提及：精确分词的最大障碍在于不能有效识别词典外的词(OOV)。它们通常具有动态特性，并按照一定的构词法则合成。在本文中，我们借用文[3]中关于词的分类，把词分为 5 大类，①词典词，即收录在词典中的词；②命名实体词(NE)，如人名、地名和机构名等词；③实体名词(Factoids)，如时间、数字、数量等词；④语法派生词(MDW)，如重叠词(高高兴兴)、合并词(上下班)等词；⑤新词，如三个代表、非典等词。由于 NE、Factoids 和 MDW 这三类词都具有一定的构词规律，是最常见也最具代表性的词典外词，而且在不同的分词标准中的处理方式差别也比较大，我们主要专注于处理这三类词。

针对这三种类型的词，我们根据构词法则，定义了 45 个词类。所谓词类，就是具有相同或相似的构词方式的一类词。如对于实体名词，我们就从语言学的角度，定义了日期、货币、数量词等 10 个词类。词 1999 年 4 月和词 5 月 5 日就属于同一词类日期。

从某种意义上说，这类词都可以看作是词组，是由一些基本的构词单元组成的。而且，如文[11]所述，考虑词的内部结

构，每个词都可以被解析成一棵树。如果我们对这些词进行抽象，如图 2.1 所示：树的根节点是词类名，树的叶节点是最基本的构词单元，而非叶节点是一些结构信息，是由基本构词单元按照特定的语法规则构成的。我们称这些结构信息为词内结构。例如，词内结构 year 就是由基本构成单元 digit 和字符年所组成。

为了避免混淆，也为了准确起见，我们把分词后的基本单位不再称为词，而称其为分隔单元，而在本文中我们把满足词类定义的词称为逻辑词。可以看出，每个逻辑词可以是一个分隔单元（如词典词），也可以是几个关系很密切的分隔单元按照一定的构词规则合成的，如实体名词中的数量词就是由数词和量词合成的。

一旦逻辑词用树状结构来表示，那么在不同分词标准下对该词的不同的切分结果就可以通过对该树的不同剪切方式来获得[11]。也就是说，分词问题变成了一个树状结构的显示问题。由于篇幅所限，具体论述可以参考文[11]。如图 2.2 所示，对于逻辑词 1995 年 4 月 14 日，它属于词类日期。当对该树采用切割方式 1 时，分词结果为 1995 年 4 月 14 日，由 3 个分隔单元组成；而采用切割方式 2 时，分词结果为 1995 年 4 月 14 日，由 6 个分隔单元组成。而我们所要做的就是使 CTBL 能有效学得这些构词规则，对词的树状结构实现有效切割。

由于词类是从语言学的角度定义的，尽管最终的分词结果可能随着分词标准的不同而不同，但基于词类的切分结果是与分词标准无关的。因此，在我们的分词系统中，是分几步来完成最后分词的。首先，我们利用现有的分词系统，对句子进行初步切分。该系统可以非常简单，如采用 FMM 算法进行分词。然后，用相关工具对句子中出现的词类进行标注。最后，采用错误驱动的学习方法 CTBL，自动学习构词规则，输出最终的分词结果。

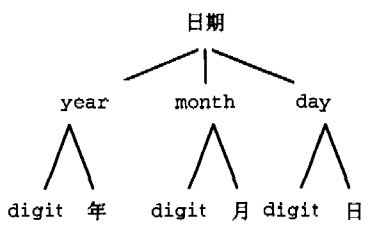

图 2　词的树状结构的抽象表示

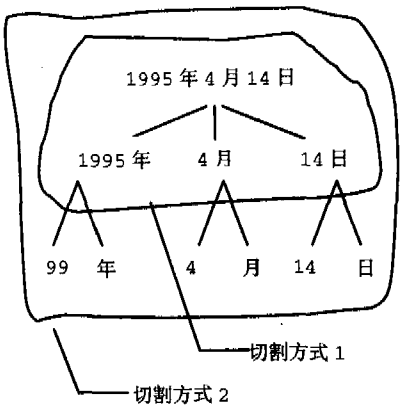

图 3　词的树状结构的具体表示

3.2　转换规则

在 CTBL 方法中，具有两种规则模板。对于基于词类的转换规则模板，我们称它为词类模板，它作用于特定词类，并包含相应的词内结构信息。模板形式如下：

条件：特定词类

动作：

- 插入-在词内结构之间插入分词标记
- 删除-在词内结构之间删除分词标记
- 插入-在词条和词内结构之间插入分词标记
- 删除-在词条和词内结构之间删除分词标记
- ……

此外，我们还具有词例化模板，形式如下：

- 插入-在两个词条之间插入分词标记
- 删除-在两个词条之间删除分词标记
- ……

在本文中，词条表示基本的词典词，不必通过任何产生式的词形变换过程来合成词。它们通常是一些单字符词、双字符词或四字符成语。

根据不同的模板，CTBL 会学得不同类型的转换规则；我们定义两种类型的转换规则：第一种是基于结构的转换规则，用于处理词内结构信息；第二种是基于词条的转换规则，用于处理词例化信息。使用词例化规则模板，只能获得基于词条的转换规则。而使用词类规则模板，就可以同时获得基于结构的转换规则和基于词条的转换规则。关于规则模板和转换规则如表 1 所示。

表 1　模板和转换规则

模板类型	规则模板	转换规则类型	转换规则示例
词类模板	如果当前词类是某个特定词类，在特定的词内结构和词条之间插入分隔标记	基于结构的转换规则	如果当前词类是 year，在数字和词条‘年’之间插入分隔标记 1999 年→1999 年
词类模板	如果当前词类是某个特定词类，删除某词条之前的分隔标记	基于结构的转换规则	如果当前词类是 year，删除词条‘年内’之前的分隔标记 1999 年内→1999 年内
词例化模板	在某词条之前插入分隔标记	基于词汇的转换规则	在词条‘总统’之前插入分隔标记 副总统→副总统
词例化模板	在特定词条之间删除分隔标记	基于词汇的转换规则	删除词条‘老人’和词条‘家’之间的分隔标记 老人家→老人家

3.3　学习和分词过程

在学习和分词过程之前，我们要对训练数据和测试数据进行预处理。首先对未切分的原始数据采用现有的分词系统进行初始分词，并使用工具识别 NE、Factoids、MDW 等词类，产生逻辑分词序列。尽管 NE、Factoids 和 MDW 的识别不太准确，但这已经非常有助于提高最终分词系统的性能。

数据预处理完成后，CTBL 通过在训练数据上的学习，会学得一系列转换规则，其中包括基于结构的转换规则和基于词条的转换规则。然后把所学到的转换规则序列应用到测试数据上，删除多余的分词标记，插入新的分词标记，实现中文分词。

例如，在 CTBL 的学习过程中，获得了两个转换规则：

1. 如果某个逻辑词属于 PN 词类，那么在词内结构姓和词内结构名之间插入分词标记。
2. 如果某个逻辑词以词条硕士开头，那么在词条硕士之后插入分词标记。

表 2 描述了如何应用规则序列进行分词。

表 2 使用 CTBL 方法的中文分词过程

分词步骤	分词结果
原始汉字序列	王小明是硕士生
逻辑切分序列	[P 王小明] 是 硕士生
应用转换规则 1	[P 王 小明] 是 硕士生
删除词类符号标记	王 小明 是 硕士生
应用转换规则 1	王 小明 是 硕士 生
最终输出	王 小明 是 硕士 生

4 评测和错误分析

4.1 数据集和评测指标

在中文分词领域，人们辛勤探索了 20 多年，提出并尝试过数十种方法，开发了上百个分词系统。为了比较和评价不同方法和系统的性能，第 41 届国际计算语言联合会（41st Annual Meeting of the Association for Computational Linguistics，41th ACL）下设的汉语特别兴趣研究组（the ACL Special Interest Group on Chinese Language Processing，SIGHAN）于 2003 年 4 月 22 日至 25 日举办了第一届国际汉语分词评测大赛。大赛采取大规模语料库测试，进行综合打分的方法，语料库和标准分别来自北京大学（简体版）、宾西法尼亚大学（简体版）、香港城市大学（繁体版）、台湾“中央研究院”（繁体版）。语料库的基本信息如表 3 所示[1]。

表 3 评测语料集

语料库	简称	训练集的大小	训练集的词数	测试集的大小
Beijing University	PK	3.53M	1.1M	17K
U. Penn Chinese Treebank	CTB	0.84M	250K	40K
Hong Kong City U.	HK	0.77M	240K	35K
Academia Sinica	AS	17.8M	5.8M	12K

为了验证 CTBL 方法的有效性，我们把它集成在现有的分词系统中，作为一个后处理工具，对初始分词系统输出的结果做进一步的处理。在我们的实验中，初始分词系统采用的算法有 CLSP 算法和 FMM 算法。其中 CLSP 是 Gao 提出的基于统计的中文分词方法，它可以在分词的同时识别 NE 等信息[4]，该分词工具可以从站点 http://research.microsoft.com/~jfgao 下载。

为了具有可比性，我们在这 4 个语料库上都做了开放式测试（Open Test）。分词系统的性能评价指标包括：精度（P）、召回率（R）、F-Score 指标，在此 F-Score 被定义为 2PR/(P+R)。

4.2 评测结果

在 4 个标准测试集上的开放测试结果如表 4 所示。对于每个测试集，我们报告了在不同的初始分词系统下（CLSP 和 FMM），引入 CTBL 进行后处理前后的分词成绩（行 1～4）。为了对比，我们还在第 5、6 行列出了在第一届中文分词大赛中相应语料集上排名前 2 位的成绩。

表 4 4 个标准测试集上的评测结果

	PK			CTB			HK			AS		
	P	R	F	P	R	F	P	R	F	P	R	F
CLSP	.852	.814	832	.829	.825	.827	.827	.820	.823	.836	.840	.838
CLSP+TBL	.950	.964	.957	877	.911	.893	.943	.959	.951	.950	.959	.955
FMM	.817	.827	.822	.773	.800	.786	.759	.790	.774	.798	.822	.810
FMM+TBL	.923	.947	.935	.831	.886	.858	.918	.943	.930	.936	.950	.943
Rank1in Bakeoff	.956	.963	.959	.907	.916	.912	.954	.958	.956	.894	.915	.904
Rank2in Bakeoff	.943	.963	.953	.891	.911	.901	.863	.909	.886	.853	.892	.872

根据上述实验结果可以看出，当采用 CLSP 作为我们的初始分词系统时，在不同的测试集上我们的分词系统都基本达到或超过了当前最好的分词系统的性能，在测试集 PK、CTB、HK 和 AS 上分别排在 2、3、2、1 位。甚至在采用简单的 FMM 算法进行初始分词时，也达到了非常好的性能。在参加 SIGHAN 评测的所有系统中，还没有哪一个系统能在 4 个开放性测试中始终保持前 3 位的位置。另外，尽管这 4 个数据集代表的是 4 个不同的分词标准，但我们没有花费时间去研究不同标准之间的差异，也没有手工引入专家知识，所有的转换规则都是自动学习获得的。这些都表明，我们的 CTBL 方法是非常有效的。

需要注意的是，在初始分词为 FMM 算法的情况下，我们无法识别 NE 信息。因此在经过 CTBL 后，性能比初始分词 CLSP 的性能要稍差一些。

此外，我们还针对 CTBL 方法的鲁棒性进行相关的实验。对于 PK 数据集，即使只采用 1/5 的训练数据，在开放测试中 F-Score 就可以达到 95.0%。这说明，CTBL 具有不错的鲁棒性，较少的训练数据就可以达到很好的分词结果。

4.3 词类信息在 CTBL 的作用

CTBL 方法优于其他 TBL 方法的重要一点就是引入了有效的语言学信息：词类和词内结构。我们将在下面的实验中对词类的作用做进一步的分析。

我们的词类都属于 3 个大的类别：NE、Factoids、MDW。我们在以下的实验中分别研究了每个大类下所属词类的作用，结果如表 5 所示。表中基线值（行 1、5）是指在没有引入词类信息的情况下使用 CTBL 方法所得的结果，也就是仅使

用词例化模板，这类似于 Palmer 实验中采用的规则模板。而表中其他行中的数据是在识别了相应类别下的词类信息后用 CTBL 方法后处理所得的结果。可以看出，词类的作用非常明显，尤其是 NE 和 Factoids 所起的作用。

表 5 词类信息在 CTBL 中的作用

	PK			CTB		
	F-Score	Improved F-Score	Error Reduction	F-Score	Improved F-Score	Error Reduction
CLSP+TBL(Baseline)	.881	——	——	.844	——	——
CLSP+NE+TBL	.921	4%	33.6%	.874	3.0%	19.2%
CLSP+NE+Factoid+TBL	.951	7%	58.8%	882	3.8%	24.4%
CLSP+NE+Factoid+MDW+TBL	.957	7.6%	63.9%	.893	4.9%	31.4%
FMM+ TBL(Baseline)	.898	——	——	.840	——	——
FMM+Factoid +TBL	.929	3.1%	30.4%	.845	0.5%	3.1%
FMM+Factoid+MDW+TBL	.935	3.7%	36.3%	.858	1.8%	11.3%

4.4 分词错误分析

通过认真分析初始分词系统造成的分词错误和加入 CTBL 方法后造成的分词错误，我们大致可以把分词错误分为 4 大类。

①不能准确识别词类的边界。如对于字串李光真图，本来李光真是个人名，而在我们的词类识别时，把李光识别为人名，真图作为一个词典词，从而导致最终的分词错误。

②由于我们所采用的是预先定义好的词类，词类的粒度比较粗，词类的划分也不够准确，有时会产生歧义。如对于实体名词 1999 年内和 1999 年 5 月，词类识别时系统认为它们属于同一词类日期，因此把同样的规则序列作用于这两个词，造成分词错误。

③不能正确识别新词，经常把新词切分为多个单字符。如星巴克被错误切分为星巴克。

④参照语料的不一致性造成分词错误。我们发现在参照语料中有些相同的字串(具有相同含义)在不同的地方却以不同的切分结果出现，这就有可能学得错误的规则。

当然，训练数据比较小也是产生错误的一个不可忽略的原因。

结束语 综上所述，我们提出的 CTBL 方法是非常有效的。当采用 CLSP 系统作为初始分词系统时，经过 CTBL 方法进行后处理后，分词性能在 SIGHAN 举办的第一届中文分词大赛的 4 个标准数据集 PK、CTB、HK 和 AS 上分别排在 2、3、2、1 位。而且该方法还具有一定的鲁棒性。

本文的贡献主要体现在 3 方面：①我们提出一种新的 TBL 方法来解决中文分词问题。通过建立新的词类规则模板，引入有效的语言学信息，我们取得了很好的效果。②我们验证了一些语言学语法信息，如词类、词内结构，对中文分词是非常有用的。③该方法对处理不同分词标准之间的差异是非常有效的。

致谢 在此，我们感谢对本文的工作给予支持和建议的朋友和同学，尤其对微软亚洲研究院自然语言处理组的高剑锋和李沐等人的热情帮助表示衷心的感谢。

参 考 文 献

1 Richard S, Emerson T. The first international Chinese word segmentation bakeoff. SIGHAN 2003
2 Richard S, Shih C. Corpus-based methods in Chinese morphology and phonology. In: COOLING 2002
3 Gao Jianfeng, Li Mu, Huang Chang-Ning. Improved source-channel model for Chinese word segmentation. ACL2003
4 Gao Jianfeng, Wu Andi, Li Mu, et al. Adaptive Chinese word segmentation. ACL2004
5 Wu Zimin, Tseng Gwyneth. Chinese text segmentation for text retrieval achievements and problems. JASIS, 1993, 44(9): 532～542
6 Palmer D. A trainable rule-based algorithm for word segmentation. ACL '97
7 Hockenmaier J, Brew C. Error-driven learning of Chinese word segmentation. In: the 12th Pacific Conference on Language and Information, Singapore, Chinese and Oriental Languages Processing Society, 1998. 218～229
8 Xue Nianwen. Chinese word segmentation as character tagging. Computational Linguistics and Chinese Language Processing, 2003
9 Wu Andi, Jiang Z. Word Segmentation in Sentence Analysis. In: Proceedings of the 1998 International Conference on Chinese Information Processing, Beijing, China, 198. 169～180
10 Wu Andi. Customizable segmentation of morphologically derived words in Chinese. International Journal of Computational Linguistics and Chinese Language Processing, 2003, 8(1): 1～27
11 Wu Andi. Chinese Word Segmentation in MSR-NLP. The first international Chinese word segmentation bakeoff, SIGHAN 2003
12 Brill E. Transformation-based error-driven learning and natural language processing: a case study in Part-of-Speech tagging. Computational Linguistics, 1995, 21(4)

(上接第 74 页)

UDP 数据包的处理程序被写为 API 方式，以在 udp_input()和 udp_creat_udp()实现时调用，其主要的函数为：

```
WORD check_udp(UDPKT * udp, LWORD sip, LWORD dip,
    int ulen);UDP 的校验和
unsigned char udpRead(unsigned char address);UDP 输入
void udpWrite(unsigned char address, unsigned char value);UDP
    输出.
```

结束语 作为一种实时性的嵌入式操作系统，uC/OS-II 内核具有代码公开、可移植、可裁剪的特点，而被广泛地运用到各个领域。但 uC/OS-II 只支持串口通信的局限，限制了 uC/OS-II 支持下的设备的 Internet 的接入。在 uIP 和 uC/OS-II 相结合后，将整个系统移植到一个由 TI 公司的 MSP430F449 和 Cirrus Logic 的 CS8900A 组建的嵌入式 websever 后，系统运行稳定，具有较高的实时性。

参 考 文 献

1 古天龙，健雄. 嵌入式实时系统及其相关问题[J]. 电子商务，1999(12)：12～15
2 Labrosse J J. uc/os-II 源码公开的实时嵌入式操作系统[M]. 邵贝贝译. 北京：中国电力出版社，2003
3 Thomas H. An Introduction to TCP/IP for Embedded Engineers [A]. In: Proc. of the embedded Systems Conf. [C], 2000
4 Dunkels A. uIP-A Free Small TCP/IP Stack[Z]. 2004
5 Postel J. Internet Protocol(IP)[Z]. RFC, 791
6 张懿慧，陈泉林. 源码公开的 TCP/IP 协议栈在远程监测中的应用[J]. 单片机及嵌入式系统应用，2003(8)：18～21