三位一体字标注的汉语词法分析 *

于江德[1]，胡顺义[1]，余正涛[2]

（1.安阳师范学院计算机与信息工程学院，河南 安阳 455000；2.昆明理工大学信息工程与自动化学院，云南 昆明 650051）

摘要：针对汉语词法分析中分词、词性标注、命名实体识别三项子任务分步处理时多类信息难以整合利用，且错误向上传递放大的不足，提出一种三位一体字标注的汉语词法分析方法，该方法将汉语词法分析过程看作字序列的标注过程，将每个字的词位、词性、命名实体三类信息融合到该字的标记中，采用最大熵模型经过一次标注实现汉语词法分析的三项任务。并在 Bakeoff2007 的 PKU 语料上进行了封闭测试，通过对该方法和传统分步处理的分词、词性标注、命名实体识别的性能进行大量对比实验，结果表明，三位一体字标注方法的分词、词性标注、命名实体识别的性能都有不同程度的提升，汉语分词的 F 值达到了 96.4%，词性标注的标注精度达到了 95.3%，命名实体识别的 F 值达到了 90.3%，这说明三位一体字标注的汉语词法分析性能更优。

关键词：汉语词法分析；最大熵模型；三位一体；字标注

中图分类号：TP391　　**文献标识码：**A

Trinity Character-Based Tagging for Chinese Lexical Analysis

YU Jiang-de[1], HU Shun-yi[1], YU Zheng-tao[2]

(1.School of Computer and Information Engineering, Anyang Normal University, Anyang Henan 455000, China; 2.School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming Yunnan 650051)

Abstract: According to the difficult to integrate multi-information and the insufficiency of error amplification while transferred upwards, a trinity character-based tagging for Chinese lexical analysis was presented both in processing word segmentation, part-of-speech tagging and named entity recognition. Treating Chinese lexical analysis as a character sequence tagging problem, each character tagging could be integrated with three kinds of information that is word-position, part-of-speech and named entity. After the tagging process was done, adopting maximum entropy model three subtasks of Chinese lexical analysis could be completed at once. Based on this closed evaluations is performed on PKU corpus from Bakeoff2007 and a large number of comparative experiments to the traditional step-by-step processing method have been done. The results show that the trinity approach achieved an F-score of 96.4% on word segmentation, POS tagging precision with 95.3% and 90.3% on named entity recognition. The raising performance indicates that the trinity character-based tagging method is much better than traditional method.

Key words: Chinese lexical analysis; maximum entropy model; trinity; character-based tagging

1 引言

在中文信息处理领域，汉语词法分析是其中一项重要的基础性研究课题。它不仅是句法分析、语义分析、篇章理解等深层中文信息处理的基础，也是机器翻译、问答系统、信息检索和信息抽取等应用的关键环节[1-3]。汉语词法分析包括汉语分词、词性标注与命名实体识别三项子任务，在国内外一些相关的评测中，常常将它们作为三个独立的子任务进行评测。在已有的研究中，大部分研究者也习惯将三项子任务独立起来加以考虑，习惯于将汉语分词和词性标注依次处理，分词之后再在词序列基础上考虑词性标注问题。这种将汉语词法分析的三项子任务独立处理的方法，虽然符合人们对汉语词法分析的认知规律，但容易造成错误向上传递放大累加，且多类信息难以整合利用的不足。针对这一问题，本文提出一种三位一体字标注的汉语词法分析方法，该方法将汉语词法分析过程看作字序列的标注过程，在每个字的标记中融合了词位、词性、命名实体三类信息，采用序列数据标注模型之一的最大熵模型实现了汉语分词、词性标注、命名实体识别三位一体的汉语词法分析。并在 Bakeoff2007 语料上进行了实验，分析了最大熵模型迭代次数对

基金项目：由国家自然科学基金项目（60863011），河南省基础与前沿技术研究计划项目（112300410182）和河南省教育厅科学技术研究重点项目（14A520077）支持。

标注性能的影响，将字标注汉语分词，词语序列基础上汉语词性标注，字标注命名实体识别作为Baseline，通过大量实验比较了它们和三位一体字标注汉语词法分析方法性能的差异。

2 相关研究和三位一体字标注思路

对于汉语词法分析这一问题，国内外已经进行了大量研究，在已有的研究中，多数研究将汉语词法分析的三项子任务独立起来进行，也有一些学者对汉语词法分析的分词、词性标注、命名实体识别三项任务的一体化进行了探索。白栓虎在 1996 年就提出了基于统计的汉语词语切分和词性标注一体化模型，在词语切分中充分利用词性标注的信息，来消除切分歧义[4]。刘群，张华平等提出了基于层叠隐马模型的汉语词法分析方法，将汉语分词、词性标注、切分排歧和未登录词识别集成到一个完整的理论框架中[2]。文献[5]深入比较了分词、词性标注两步走和一体化的优劣，认为基于字标注的一体化分词和词性标注方法是最佳方案，其分词系统获得了 SIGHAN2003 四个测试语料中三项封闭测试第一，同时又肯定了两步走方案在训练和测试时间上的优势。石民等探索了古代汉语，特别是先秦文献中的词语切分和词性标注一体化的方法[6]。文献[7-9]也都研究了汉语分词和词性标注的一体化问题。

本文在前人研究的基础上提出一种三位一体字标注的汉语词法分析方法，该方法将汉语词法分析三个子任务全部统一到字标注的框架中，在每个字的标记中包含了词位、词性、命名实体三类信息，形式为“**词位_词性或命名实体类别**”，字标记由两部分组成，中间用下划线隔开，下划线之前是词位信息，之后是词性或命名实体类别信息。其中，词位是指该字在所构成的特定词语中所占据的构词位置，本研究中规定字只有四种词位：B（词首）、M（词中）、E（词尾）和 S（单字成词）。根据字序列标记中的词位信息就可以实现汉语分词。词性是该字所在的特定词语所属词语类别。本文所用词性标注集为北京大学计算语言学研究所的词性标注集。如果该字所在的词语为命名实体，则标记中下划线后为相应命名实体类别。本文研究的命名实体包括人名、地名、组织机构名三类，分别用 PER、LOC、ORG 标识。根据字序列标记中的词性和命名实体类别部分可以分别实现汉语词性标注和命名实体识别。三位一体字标注汉语词法分析就是把词法分析过程看作是一个字序列的标注过程。如果一个汉语字串中每个字的标记都确定了，那么该汉语字串的分词、词性标注、命名实体识别也就完成了。例如：要对字串序列“中国政府顺利恢复对香港行使主权，”进行词法分析，只要得到该字串的标注结果（如图 1 所示），然后再根据三位一体字标注汉语词法分析的思想，由标注结果中的词位部分可以得到分词结果，由词性或命名实体类别部分可以得到词性标注和命名实体识别结果，综合这些结果就得到相应的词法分析结果。该字串的汉语词法分析结果为“中国政府/ORG　顺利/ad　恢复/v　对/p　香港/LOC　行使/v　主权/n　，/wd”。

中　B_ORG
国　M_ORG
政　M_ORG
府　E_ORG
顺　B_ad
利　E_ad
恢　B_v
复　E_v
对　S_p
香　B_LOC
港　E_LOC
行　B_v
使　E_v
主　B_n
权　E_n
，　S_wd

图 1 三位一体字标注示意图
Fig. 1 Trinity character-based tagging

另外，三位一体字标注的汉语词法分析中还有几个问题需要注意。（1）对于汉语真实文本中包含的标点符号、西文字母、数字等少量非汉字字符和汉字同等对待。（2）标注结果中多字词的多个字的标记中，每个字的词性或命名实体类别标记部分未必一致，这时该如何确定该词的词性或命名实体类别呢？是取词首字的，还是词尾字的或词中字的标记作为整个词的词性或命名实体类别的呢？例如，字标注结果“希 B_v 望 M_v 工 M_n 程 E_n”使得词语“希望工程”可以选取词性“动词 v”，也可以选取“名词 n”。本文根据实验对比选取词尾字的标记作为整个词语的词性或命名实体类别。

3 基于最大熵模型的三位一体字标注

由于最大熵模型可以有效地把各种约束条件整合在一起，近年来在自然语言处理领域被广泛应用[10-14]。本文采用最大熵模型实现三位一体字标注，本小节重点解释最大熵模型如何对三位一体字标注建模。

3.1 最大熵模型简介

最大熵模型是建立在最大熵理论基础之上的。最大熵理论反映了自然界的一条基本原则：事物是约束和自由的统一体，并且在约束下事物总是争取最大自由度，即最大熵。因此，在已知条件下，熵最大的事物，最可能接近它的真实状态。基于最大熵理论对一个事物建模时，往往只掌握该事物的部分情况，对其他情况一无所知。建模时，对于已知的部分要尽量地拟合，使模型符合已知的情况。对于未知情况，让可能出现的每种结果保持平均分布，使该事物的熵最大，这样构建的模型就是最大熵模型。

对于三位一体字标注汉语词法分析问题，给定一些训练样本(x, y)，其中 x 表示上下文，即字序列，y 表示字的标注序列，可根据这些已知的样本构建一个能够对实际问题进行准确描述的概率统计模型 $p(y \mid x)$ 用来预测未知的标记。该模型的概率分布与训练语料中的经验概率分布应该相符。最大熵原理表明，x, y 的正确分布应该是在满足训练语料中已知条件（约束）的情况下熵最大的分布，这样构建的模型是最大熵模型，其一般形式为：

$$p(y \mid x)=\frac{1}{Z(x)}\exp\left[\sum_{i=1}^{k}\lambda_i f_i(x,y)\right], \quad (1)$$

其中，

$$Z(x)=\sum_{y}\exp\left[\sum_{i=1}^{k}\lambda_i f_i(x,y)\right], \quad (2)$$

$Z(x)$ 为归一化因子，保证对所有可能的上下文 x 及其标注 y，$\sum p(y \mid x)=1$。$f_i(x,y)$ 是特征函数，k 为特征函数的数目，参数 λ_i 是反映特征函数 f_i 对于模型重要程度的权重。这些特征函数用来描述已知的约束条件，一般情况下特征函数是一个二值函数，形式如下：

$$f(x,y)=\begin{cases}1 & \text{如果}(x,y)\text{满足某种约束}\\ 0 & \text{否则}\end{cases} \quad (3)$$

3.2 最大熵模型对三位一体字标注的建模

基于最大熵模型进行三位一体字标注首先要建立模型，其中的关键问题是针对三位一体字标注这个特定任务为模型选择合适的上下文特征，即筛选出对最大熵模型有表征意义的上下文特征，包括选取适当的上下文范围和设定特征模板，即样本窗口的大小设定和特征模板集的构建。

3.2.1 样本窗口的大小设定

采用最大熵模型进行三位一体字标注汉语词法分析时，上下文将为正确的标注提供所需的语言知识和相关资源。通常情况下，上下文的选取是基于当前字左右一定范围进行的，这个固定的范围被称为“窗口”。窗口中的上下文实质是一个特定样本，所以将该窗口称为“样本窗口”。进行词法分析时所需的语言知识将从该窗口产生的大量样本中进行统计学习。建模时首先要考虑上下文范围，即样本窗口开设大小问题，这需要通过对比实验看看多大的样本窗口使得汉语词法分析的性能最好。图 2 是可能的样本窗口的图示，显然可以根据需要来选取上下文的范围，即样本窗口的大小。可以限定样本窗口是“5 字窗口”，即使用当前字前后各两个字作为上下文。也可以限定样本窗口是“3 字窗口”，即使用当前字前后各一个字作为上下文。

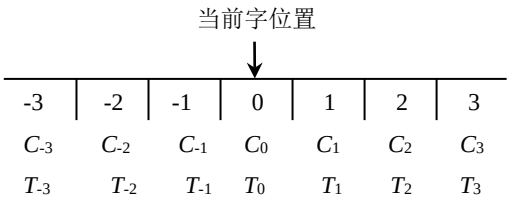

图 2 可能的样本窗口

Fig. 2 The possible sample window

3.2.2 特征模板集的构建

统计语言建模中上下文特征的刻画是通过特征模板实现的。特征模板的主要功能是定义上下文中某些特定位置的语言成分或信息与某类待预测事件的关联情况。由于本文是根据一个字串序列中的当前字及其上下文来确定该字的标记信息，因此就由该字前后出现的字、字的组合、词位、词性、命名实体类别等信息及这些信息出现的位置来确定上下文特征。习惯上，特征模板可以看作是对一组上下文特征按照共同的属性进行的抽象。在最大熵模型的训练学习中，每个特征都对应了一组特征函数，这些特征函数对最大熵模型的训练至关重要。而这些特征又是通过特征模板扩展来的，所以设定合适的特征模板集就显得尤为重要。

在使用 MaxEnt 工具包进行三位一体字标注汉语词法分析时，在图 2 所示的可能样本窗口下，

可以将上下文特征按照特征模板中出现的字与当前字的字距属性进行抽象。表 1 给出了“5 字窗口”下常用的 10 个特征模板及其表征的意义。表中的 C_n代表当前字和当前字相距若干字位的字。例如，C_0表示当前字，C_1表示当前字的后一个字，C_{-1}表示当前字的前一个字，依此类推。从表 1 可以看到，最后一行的特征模板是：$T_{-1}T_0$，该模板用于表征上下文中相邻两个字标记的转移特征 $T_{i-1} \to T_i$。

表 1 特征模板列表

Table 1 List of feature templates

模板	特征模板表征的意义
C_{-2}	当前字的前面第二个字
C_{-1}	当前字的前一个字
C_0	当前字
C_1	当前字的后一个字
C_2	当前字的后面第二个字
$C_{-2}C_{-1}$	当前字的前面两个字组成的字串
$C_{-1}C_0$	当前字前一个字和当前字组成的字串
C_0C_1	当前字及其后一个字组成的字串
C_1C_2	当前字的后面两个字组成的字串
$C_{-1}C_1$	当前字的前一个字和后一个字组成的组合
$T_{-1}T_0$	相邻两个字标记的转移特征

根据图 2 给出的可能样本窗口，如果限定样本窗口是“5 字窗口”，则这一具体任务的上下文特征是指当前字本身、以及当前字前后各两个字及其字标记所组成的特征。针对三位一体字标注汉语词法分析这一具体任务，“5 字窗口”下常见上下文特征抽象为表 1 中列出的 10 类，记这种模板集为 TMPT-10，该特征模板集包含的模板如表 2 所示。如果限定样本窗口是“3 字窗口”，即使用当前字前后各一个字作为样本窗口，则这一具体任务的字特征是指当前字本身、以及当前字前后各一个字及其标记所组成的特征，此时的特征模板集为六特征模板集：TMPT-6。

为了对汉语词法分析中的特征模板有个“量”的认识，我们从多个角度进行定量分析并设计了相关实验。表 2 列出了实验中用到的几组特征模板集。其中，序号 1-3 的特征模板集是“5 字窗口”的模板集，4-6 的特征模板集是“3 字窗口”的模板集。另外，后缀“Single”和“Double”分别表示相应特征模板集中的只有单个字的特征模板集和有双字组合构成的特征模板集。例如，T10-Single 是指 TMPT-10 中单个字特征模板。另外，这六组特征模板集中都包含字标记转移特征。

表 2 特征模板集列表

Table 2 Sets of feature templates

序号	特征模板集	包含的特征模板
1	TMPT-10	C_{-2}，C_{-1}，C_0，C_1，C_2，$C_{-2}C_{-1}$，$C_{-1}C_0$，C_0C_1，C_1C_2，$C_{-1}C_1$，$T_{-1}T_0$
2	T10-Single	C_{-2}，C_{-1}，C_0，C_1，C_2，$T_{-1}T_0$
3	T10-Double	$C_{-2}C_{-1}$，$C_{-1}C_0$，C_0C_1，C_1C_2，$C_{-1}C_1$，$T_{-1}T_0$
4	TMPT-6	C_{-1}，C_0，C_1，$C_{-1}C_0$，C_0C_1，$C_{-1}C_1$，$T_{-1}T_0$
5	T6-Single	C_{-1}，C_0，C_1，$T_{-1}T_0$
6	T6-Double	$C_{-1}C_0$，C_0C_1，$C_{-1}C_1$，$T_{-1}T_0$

4 实验及其分析

4.1 实验环境及实验数据集

本文所有实验的软硬件环境为：实验所用计算机型号为 DELL Optiplex 760 台式机，其主要参数为：中央处理器：Intel(R) Core(TM)2 Quad CPU Q8200 2.33GHZ；内存：4GB；操作系统：Microsoft Windows XP。

本文采用的训练语料和测试语料是 SIGHAN 举办的第四届国际汉语语言处理评测 Bakeoff2007 所使用的语料，是由北京大学(PKU)提供的汉语词性标注语料和命名实体语料，其中汉语词性标注语料大小为 8.42MB，词数为 1116574 个。命名实体语料大小为 11.2MB。这两种语料所标注的文本内容完全相同，进行三位一体字标注汉语词法分析训练或测试时需要将这两种语料进行处理后融合到一起，图 3 是语料处理过程的示意图。首先是将原词性标注语料拆分为一字一标记的格式，此时的标记形式为“**词位_词性类别**”，然后再根据命名实体语料将所有命名实体的那部分字的标记修改为“**词位_命名实体类别**”，融合后的语料大小为 15.0MB。然后将十分之九作为训练语料，十分之一作为测试语料。统计发现，这些语料中共有字标记 257 种，由于标记较多，所以本文的实验采用最大熵模型实现，而没有采用序列数据标注模型条件随机场（conditional random fields，CRFs）实现，因为采用 CRF++工具包训练时，在标记类别多和语料较大的情况下不能正常进行训练，而最大熵模型则没有此类问题。采用 MaxEnt 工具包进行模型训练时，还需要对融合后的语料进行预处理，按照设定的样本窗口和特征模板集将语料处理为一行

一个事件的语料，也就是对每一个样本按照特征模板扩展出相应的上下文特征作为一个事件。

词性标注语料中原始语料如下：		
中国/ns 政府/n 顺利/ad 恢复/v 对/p 香港/ns 行使/v 主权/n ，/wd		
词性标注拆分后语料如下：	命名实体原始语料如下：	融合后的语料如下：
中 B_ns 国 E_ns 政 B_n 府 E_n 顺 B_ad 利 E_ad 恢 B_v 复 E_v 对 S_p 香 B_ns 港 E_ns 行 B_v 使 E_v 主 B_n 权 E_n ， S_wd	中 B_ORG 国 I_ORG 政 I_ORG 府 I_ORG 顺 N 利 N 恢 N 复 N 对 N 香 B_LOC 港 I_LOC 行 N 使 N 主 N 权 N ， N	中 B_ORG 国 M_ORG 政 M_ORG 府 E_ORG 顺 B_ad 利 E_ad 恢 B_v 复 E_v 对 S_p 香 B_LOC 港 E_LOC 行 B_v 使 E_v 主 B_n 权 E_n ， S_wd

图 3 语料处理过程示意图
Fig. 3 Schematic diagram of corpus processing

4.2 性能评估

在对三位一体字标注汉语词法分析进行性能评估时，本文采用两类评估方法。一类是根据设定的特征模板集进行整体评价，采用的评价指标是字标注准确率。该准确率表示在测试语料全部字标注中，正确的所占的比值。另一类是该方法和传统分步处理的分词、词性标注、命名实体识别的性能进行对比，采用的评估指标如下所述。

在对汉语分词性能进行评估时，采用了常用的 5 个评测指标：准确率（*P*）、召回率（*R*）、综合指标 *F* 值（*F*）、未登录词召回率（*OOV RR*）、词表词召回率（*IV RR*）。准确率表示在切分的全部词语中，正确的所占的比值。召回率指正确切分的词语占标准答案中词语的比值。综合指标 *F* 值是综合准确率和召回率两个值进行评价的一种办法。*OOV RR* 和 *IV RR* 分别指测试中未登录词和词表词的召回率。

在对汉语词性标注性能进行评估时，采用了常用的评测指标：标注精度。标注精度表示在对全部词语标注的词性中，正确标注词性的词语所占的比值。

在对汉语命名实体识别进行评估时，采用了常用的 3 个评测指标：准确率（*P*）、召回率（*R*）、综合指标 *F* 值（*F*）。准确率表示在识别的全部命名实体中，正确的所占的比值。召回率指正确识别的命名实体占标准答案中的比值。*F* 值是综合准确率和召回率两个值进行评价的一种办法。

4.3 实验及其结果分析

4.3.1 实验设计

本文设计了两个阶段的实验，分别配合两类评估方法对三位一体字标注的汉语词法分析性能进行评估。第一个阶段是在测试语料的字标注结果上进行的，采用字标注的准确率进行评估。在第一阶段结果的基础上第二个阶段分别就汉语分词、词性标注、命名实体识别三项子任务的性能进行三组对比实验：（1）三位一体字标注汉语词法分析的分词性能和基于字标注的汉语分词性能对比实验。（2）三位一体字标注汉语词法分析的词性标注性能和词序列基础上的汉语词性标注性能对比实验。（3）三位一体字标注汉语词法分析的命名实体识别性能和基于字标注的命名实体识别性能对比实验。

4.3.2 三位一体字标注的汉语词法分析性能

我们首先分别使用表 2 中序号为 1~6 的六组特征模板集，在预处理后的训练语料上进行了三位一体字标注汉语词法分析的训练，训练时采用不同迭代次数，最大熵模型迭代次数从 50 增加到 400，间隔 50。表 3 给出了使用这六组特征模板集在部分迭代次数下的训练过程记录数据。综合分析表 3 中的数据可以得出如下结论：（1）同等条件下，训练出的模型大小与扩展出的特征数成正比，训练出的模型大小随迭代次数的变化很小。（2）模型训练的时间长短和扩展出的特征数并没有必然联系，和训练的迭代次数成正比。

然后分别采用训练出的模型，对测试语料进行三位一体字标注测试，测试的字标注准确率如表 4 所示。从表 4 中的数据可以得出如下结论：（1）迭代次数到一定值时标注准确率不再提升，甚至有少许下降。例如，对于 TMPT-10 特征模板集来说，迭代次数从 50 增加到 100，标注准确率增加最多，之后趋于平缓，迭代次数为 200 时，标注准确率达到了最高，之后有少许下降。所以第二阶段的对比实验都是在迭代次数为 200 下进行的。（2）从样本窗口大小的角度来分析，对比序号 1-3 和 4-6 的特征模板集下的标注性能，可见“5 字窗口”下的标注性能比“3 字窗口”的好。所以第二阶段的对比实验中，三位一体字标注汉语词法分析都是在“5 字窗口”下进行的。

表 3 PKU 语料上不同迭代次数的训练过程记录数据

Table 3 Record data of training process on PKU corpus with different number of iterations

模板集序号	特征数	迭代次数为 50		迭代次数为 100		迭代次数为 200		迭代次数为 300		迭代次数为 400	
		训练时间/s	模型大小/Mb	训练时间/s	模型大小/Mb	训练时间/s	模型大小/Mb	训练时间/s	模型大小/Mb	训练时间/s	模型大小/Mb
1	1 518 892	2 641.30	96.7	5 218.64	95.6	10 615.20	95.3	16 165.80	95.1	21 409.70	94.9
2	23 309	2 459.92	12.3	4 856.86	12.1	9 775.34	11.9	14 793.25	11.9	19 778.59	11.9
3	1 495 841	1 576.08	84.0	3 202.50	82.4	6 492.45	81.4	9 863.49	81.4	13 167.28	81.4
4	983 324	1 298.97	54.6	2 562.98	53.9	5 212.39	53.7	7 820.64	53.6	10 405.36	53.6
5	14 090	1 211.98	5.7	2 388.00	5.6	4 790.66	5.5	7 201.20	5.5	1 211.98	5.5
6	969 493	909.47	48.5	1 857.45	48.0	3 725.41	47.4	5 569.80	47.4	7 445.13	47.4

表 4 不同迭代次数的三位一体汉语词法分析标注准确率

Table 4 Tagging precision of Chinese lexical analysis via trinity character-based tagging with different number of iterations

迭代次数及性能 / 模板集序号	50	100	150	200	250	300	350	400
1	94.07%	95.26%	95.39%	95.39%	95.36%	95.28%	95.19%	95.15%
2	91.30%	92.63%	92.75%	92.84%	92.79%	92.65%	92.59%	92.47%
3	80.27%	92.10%	93.00%	93.17%	93.23%	93.22%	93.23%	93.19%
4	92.21%	93.85%	94.10%	94.15%	94.20%	94.21%	94.22%	94.22%
5	89.60%	91.20%	91.60%	91.76%	91.88%	91.86%	91.86%	91.88%
6	84.56%	90.52%	92.13%	92.62%	92.65%	92.59%	92.60%	92.58%

4.3.3 三位一体字标注词法分析与其他方法比较

在三位一体字标注的基础上第二个阶段分别就汉语分词、词性标注、命名实体识别三项任务的性能进行对比实验。首先是对本文三位一体字标注汉语词法分析中的分词性能和基于单一字标注的汉语分词性能进行对比。其中，单一字标注汉语分词采用条件随机场模型实现，设定的样本窗口大小和特征模板集和三位一体字标注方法相同，都是“5 字窗口”和 TMPT-10。表 5 给出了本文方法和字标注方法汉语分词性能对比。从表 5 的数据中可以看到，三位一体字标注的汉语词法分析中的汉语分词性能比单一字标注的汉语分词方法的性能的综合指标 *F*值提高了 2.3 个百分点，这说明在字的标记中融入词性和命名实体的信息对汉语分词性能有一定的提高。

表 5 不同方法的汉语分词结果

Table 5 Experimental results of Chinese word segmentation

不同方法	*P*	*R*	*F*	*OOV RR*	*IV RR*
三位一体方法	0.964	0.963	0.964	0.949	0.963
单一字标注	0.945	0.937	0.941	0.629	0.945

然后对三位一体字标注汉语词法分析的词性标注性能和词序列基础上的汉语词性标注性能进行了对比实验。其中，词序列基础上的方法也采用最大熵模型实现，设定的样本窗口为“3 词语窗口”，特征模板集为“W_{-1}，W_0，W_1，$T_{-1}T_0$”。表 6 给出了本文方法和词序列基础上的汉语词性标注性能对比情况，其中对于多字词的词性选取的是词尾字的词性标记。从表 6 的数据可以看到，三位一体字标注中的汉语词性标注性能比基于词序列的汉语词性标注性能提高了 0.7 个百分点。

表 6 不同方法的汉语词性标注结果

Table 6 Experimental results of Chinese POS tagging

不同方法	标注精度
三位一体方法	95.3%
词序列基础上的方法	94.6%

最后对三位一体字标注汉语词法分析的命名实体识别性能和基于单一字标注的命名实体识别性能进行对比实验。其中，单一字标注的命名实体识别采用条件随机场模型实现，设定的样本窗口大小和特征模板集分别为“5 字窗口”和 TMPT-10。表 7 给出了实验结果。从表 7 中的数据可见，本文的方法比单一字标注的方法提高了 2 个百分点多。

表 7 不同方法的中文命名实体识别结果

Table 7 Experimental results of Chinese NER

不同方法	*P*	*R*	*F*
三位一体方法	0.9282	0.8785	0.9026
单一字标注	0.9137	0.8523	0.8819

5 结语

在中文信息处理领域，汉语词法分析是其中一项重要的基础性研究课题。针对汉语词法分析中分词、词性标注、命名实体识别三项子任务分步处理时多类信息难以整合利用，且错误向上传递放大的不足，本文提出一种三位一体字标注的汉语词法分析方法，该方法将汉语词法分析过程看作字序列的标注过程，将每个字的词位、词性、命名实体三类信息融合到该字的标记中，采用最大熵模型经过一次标注实现汉语词法分析的三项任务。实验结果表明，三位一体字标注方法的分词、词性标注、命名实体识别的性能都有不同程度的提升。今后将进一步完善该方法，力争能在中文信息处理的实际任务推广应用。

参考文献：

[1] 姜维，王晓龙，关毅，等. 基于多知识源的中文词法分析系统[J]. 计算机学报， 2007，30(1):137-145.
JIANG Wei, WANG Xiaolong, GUAN Yi, et al. Research on Chinese lexical analysis system by fusing multiple knowledge sources [J]. Chinese Journal of Computers, 2007，30(1):137-145.

[2] 刘群，张华平，俞鸿魁，等. 基于层叠隐马模型的汉语词法分析[J]. 计算机研究与发展，2004, 41(8):1421-1429.
LIU Qun, ZHANG Huaping, YU Hongkui, et al. Chinese lexical analysis using cascaded hidden Markov model [J]. Journal of Computer Research and Development, 2004, 41(8):1421-1429.

[3] 孙晓，黄德根. 基于最长次长匹配分词的一体化中文词法分析[J].大连理工大学学报, 2010,50(6):1028-1034.
SUN Xiao, HUANG Degen. Chinese integrative lexical analysis based on maximum matching and second-maximum matching segmentation [J]. Journal of Dalian University of Technology, 2010,50(6):1028-1034.

[4] 白栓虎．汉语词切分及词性自动标注一体化方法[J]．中文信息，1996，（2）：46-48.
BAI Shuanhu. A unified approach to Chinese word segmentation and POS tagging [J]. Chinese Information Processing, 1996，（2）：46-48.

[5] Hwee Tou Ng, Jin Kiat Low. Chinese part-of-speech tagging: One-at-a-time or all-at-once? Word-based or character-based? [C]. // Proceedings of the Conference on Empirical Methods in Natural Language Processing, Barcelona: ACL Press, 2004: 277-284.

[6] 石民，李斌，陈小荷．基于 CRF 的先秦汉语分词标注一体化研究[J]. 中文信息学报，2010，24（2）：39-45.
SHI Min, LI Bin, CHEN Xiaohe. CRF based research on a unified approach to word segmentation and POS tagging for Pre-Qin Chinese[J]. Journal of Chinese Information Processing, 2010，24（2）：39-45.

[7] LUO Xiaoqiang. A maximum entropy Chinese character-based parser [C]. // Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Sapporo, Japan: ACL Press, 2003: 192-199.

[8] JIANG Wenbin, HUANG Liang, LIU Qun, et al. A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging [C]. // Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, Columbus: ACL Press, 2008: 897-904.

[9] 朱聪慧，赵铁军，郑德权. 基于无向图序列标注模型的中文分词词性标注一体化系统[J].电子与信息学报，2010,32(3):700-704.
ZHU Conghui, ZHAO Tiejun, ZHENG Dequan. Joint Chinese word segmentation and POS tagging system with undirected graphical models [J]. Journal of Electronics & Information Technology, 2010, 32(3):700-704.

[10] BERGER A L, DELLA-PIETRA S A, DELLA-PIETRA V J. A maximum entropy approach to natural language processing [J]. Computational Linguistics,1996,22(1):39-71.

[11]刘挺，车万翔，李生. 基于最大熵分类器的语义角色标注[J]. 软件学报，2007, 18(3):565-573.
LIU Ting, CHE Wanxiang, LI Sheng. Semantic role labeling with maximum entropy classifier [J]. Journal of Software, 2007, 18(3):565-573.

[12]何径舟，王厚峰. 基于特征选择和最大熵模型的汉语词义消歧[J]. 软件学报，2010, 21(6):1287-1295.
HE Jingzhou, WANG Houfeng. Chinese word sense disambiguation based on maximum entropy model with feature selection [J]. Journal of Software, 2010, 21(6):1287-1295.

[13]赵岩，王晓龙，刘秉权，等. 融合聚类触发对特征的最大熵词性标注模型[J]. 计算机研究与发展，2006,43(2):268-274.
Zhao Yan, Wang Xiaolong, Liu Bingquan, et al. Fusion

of clustering trigger-pair features for POS tagging based on maximum entropy model [J]. Journal of Computer Research and Development, 2006,43(2):268-274. (in Chinese)

[14] 张贯虹，斯•劳格劳，乌达巴拉. 融合形态特征的最大熵模型蒙古文词性标注模型[J]. 计算机研究与发展，2011,48(12):2385-2390.
ZHANG Guanhong, S. Loglo, Odbal. Fusion of morphological features for Mongolian part of speech based on maximum entropy model [J]. Journal of Computer Research and Development, 2011,48(12):2385-2390.

于江德（1971-），博士，教授，主要研究领域为自然语言处理和机器学习。E-mail：jiangde_yu@163.com

胡顺义（1981-），硕士，讲师，主要研究领域为自然语言处理。E-mail：neilh@163.com

余正涛（1970-），博士，教授，主要研究领域为自然语言处理和信息检索。E-mail：ztyu@hotmail.com