文章编号：1000-8608(2010)06-1028-07

基于最长次长匹配分词的一体化中文词法分析

孙 晓[*1,2]， 黄德根[1]

（1.大连理工大学 计算机科学与技术学院，辽宁 大连 116024；
2.大连民族学院 计算机科学与工程学院，辽宁 大连 116600）

摘要：针对当前大多数词法分析系统“流水线”式处理方式存在的不足，提出一种一体化同步词法分析机制.在最长次长匹配分词的基础上，在切分有向图中增加词性信息和候选未登录词节点，并拓展隐马尔可夫模型，在切分有向图内同步完成分词、歧义消解、未登录词识别和词性标注等词法分析任务.实现了分词与词性标注的一体化、未登录词识别与分词的一体化以及不确定词性未登录词处理的一体化.一体化机制使词法分析中各步骤实现真正意义上的同步完成，充分利用上下文词法信息提高整体精度并保证了系统的高效性，避免了各步骤间的冲突.开放测试表明，系统综合测试的 F 值为 98.03%.

关键词：中文词法分析；一体化模型；最长次长匹配；未登录词；切分有向图
中图分类号：TP391 **文献标志码**：A

0 引 言

词是自然语言中有意义并且可以独立使用的最小单位，词法分析是句法标注、语义标注等语料库深层加工的基础.大多数自然语言处理系统如机器翻译、语音合成、信息抽取、文献检索等都以词为基本的处理单位，正确的词法分析具有十分重要的意义.汉语中词与词之间没有空格等明显的分割标记，因此分词和词性标注是汉语词法分析中必须首先解决的两个重要问题，其中分词又包括歧义消解、未登录词识别等.

在汉语词法分析领域，研究者们经过了 20 多年的努力开发了众多词法分析系统，提出了多种理论、模型和技术，并在分词算法、歧义消解、词性标注和未登录词识别等方面取得了很大的进展.现有的词法分析所用到的方法大体可以分为三大类：第一类是基于规则的方法，例如各种专家系统，通过构建语言学知识规则库进行词法分析；第二类是基于统计的方法，通过对大规模语料库的机器学习，利用各种数学统计模型进行词法分析，常用的统计模型有隐马尔可夫模型[1,2]、n 元语法模型、信道-噪声模型、最大熵模型、支持向量机(SVM)模型、条件随机场(CRF)模型；第三类是规则与统计结合的方法，在实际应用系统中，往往采用统计与规则等多种方法相结合，在以规则为主的系统中采用统计方法来训练规则模型，而在以统计为主的模型中采用一定的规则来消除歧义、识别未登录词等.

尽管汉语词法分析的研究已经取得了显著成果，但是其中存在的问题仍然是制约汉语句法分析精度提高的主要因素.黄昌宁等对一个实用化句法分析器(正确率约 73%)错误分布进行了统计，结果如下(按句子)：

分词错误 40% 词性错误 24% 组块 12% 中心词 24%

并指出 NLP 底层工作不扎实造成了 MT 性能低下.只有实现高精度的词法分析，构建于词平面之上的各种后续语言分析工作才有实际意义.近来，研究者对一体化词法分析的研究进一步深入，提出了多种中文一体化词法分析的方法和系

收稿日期：2008-04-22； **修回日期**：2010-07-02.
基金项目：中央高校基本科研业务费专项资金资助项目(DUT10RW202).
作者简介：孙 晓*(1980-)，男，博士，E-mail：sunxiao@dlnu.edu.cn；黄德根(1965-)，男，博士，教授，博士生导师，E-mail：huangdg@dlut.edu.cn.

统[2~8]，构建了诸如基于 HHMM 的一体化分词方法[2]，其一体化处理流程一般为：首先对句子进行初步切分和未登录词识别；然后将未登录词视为普通词，进行进一步分词和词性标注；另外研究者还提出了考虑词长作为特征，扩展隐马尔可夫的一体化方法[3]；引入词和词性的三元 HMM 模型的一体化方法[4]，基于分治算法的一体化词法[5]以及利用感知器，通过训练区别统计模型进行一体化词法分析的方法[6]. 在当前这些一体化词法分析系统中，对中文词法分析和词性标注的一体化处理提出了高效的处理方法，并取得了较高的精度，但是，在这些一体化方法和系统中，仍然存在一些需要解决的问题，例如候选路径搜索空间膨胀问题，未登录词的融合和一体化问题以及未登录词的词性标注问题.

针对现有词法分析系统存在的问题，本文基于最长次长匹配分词，提出了一种真正意义上的一体化同步处理机制，即基于最长次长匹配构建切分有向图[1]，并在切分有向图中加入词性信息，利用最长次长匹配方法生成候选分词和词性标注路径时，只保留从某位置开始的最长和次长切分结果及其相应的词性来构建切分有向图，可以解决候选搜索路径空间膨胀带来的系统整体效率降低的问题. 为了将未登录词融入一体化处理中，本文构造未登录词识别自动机，在切分有向图中识别出所有候选未登录词，加入到切分有向图中，赋予切分有向图各节点和边相应的代价，最后在包含未登录词的切分有向图中进行路径选择，得到切分和标注结果.

1　分词与词性标注一体化统计模型

在文献[4]中，通过引入三元模型并扩展 HMM 模型实现了一体化的词法分析. 为了一体化模型能在最长次长匹配建立的切分有向图中实现，将扩展的 HMM 模型与有向图整合，并融入未登录词处理，首先对 HMM 模型进行类似的二元扩展. 设句子 S 的某个可能分词结果为 $W=\{w_1,w_2,\cdots,w_n\}$，词 w_i 的词性为 t_i，则对 W 的词性标注序列记为 $T=\{t_1,t_2,\cdots,t_n\}$. 取概率最大的 W^* 作为最终的分词标注结果，即

$$W^*=\arg\max_{W,T}\{P(W,T)\} \tag{1}$$

根据贝叶斯(Bayes) 概率论规则，$P(W,T)$ 可以分解为一单词序列和一标注序列的概率，需要将 $P(W,T)$ 分别按照单词和词性展开，将展开后的两式整合，形成分词与词性标注一体化的统计模型. 首先利用贝叶斯公式对 $P(W,T)$ 按词性展开，得到

$$W_{\text{tag}}=\arg\max_{W}\{P(W\mid T)P(T)\} \tag{2}$$

假设某单词的概率仅受到其词性的约束，而且该词性的出现概率仅与其前面出现的词性有关系. 利用隐马尔可夫进一步展开式(2)，词性概率 $P(T)$ 采用二元(bi-gram) 模型并取负对数，得到

$$W_{\text{tag}}=\arg\min_{W}\left\{\sum_{i=1}^{n}(-\log_2 p(w_i\mid t_i)-\log_2 p(t_i\mid t_{i-1}))\right\} \tag{3}$$

式中：$p(w_i\mid t_i)$ 表示在某词性 t_i 条件下单词 w_i 出现的概率；$p(t_i\mid t_{i-1})$ 表示在上一个词性 t_{i-1} 出现的前提下，当前词性 t_i 出现的概率，可以通过对大规模语料库的统计学习得到.

类似式(2)，利用贝叶斯公式对 $P(W,T)$ 按词展开，得到 $P(W,T)$ 的另一个变形：

$$W_{\text{word}}=\arg\max_{W}\{P(T\mid W)P(W)\}$$

词形概率 $P(W)$ 采用二元(bi-gram) 模型，并取负对数得到：

$$W_{\text{word}}=\arg\min_{W}\left\{\sum_{i=1}^{n}(-\log_2 p(t_i\mid w_i)-\log_2 p(w_i\mid w_{i-1}))\right\} \tag{4}$$

其中 $p(t_i\mid w_i)$ 表示某单词 w_i 被标注成词性 t_i 的概率. $p(w_i\mid w_{i-1})$ 表示前一个单词为 w_{i-1} 的前提下当前单词 w_i 出现的概率. 式(3) 决定了汉字序列的分词及标注，而在此后增加 $p(t_i\mid w_i)$ 会对先前的词性标注增加偏差. 所以将式(4) 中的 $p(t_i\mid w_i)$ 去掉，得到

$$W_{\text{word}}=\arg\min_{W}\sum_{i=1}^{n}\left\{(-\log_2 p(w_i\mid w_{i-1}))\right\} \tag{5}$$

结合式(3) 和(5) 得到概率最大的分词结果

$$W^*=\arg\min_{W}\left\{\sum_{i=1}^{n}(-\log_2 p(w_i\mid t_i)-\log_2 p(t_i\mid t_{i-1}))+\alpha\arg\min_{W}\sum_{i=1}^{n}(-\log_2 p(w_i\mid w_{i-1}))\right\} \tag{6}$$

式中 α 为词典中词的总数及词典中词性总数间的比值. 式(6)为扩展后的 HMM 模型,与传统意义上的 HMM 模型相比,扩展后的 HMM 模型引入了词的上下文信息,为了验证增加词上下文信息的效果,对 10 000 个句子进行了开放测试,实验结果见表 1.

表 1 式(3)和(6)计算结果的对比

Tab. 1 Comparison between results of Eqs. (3) and (6)

计算公式	分词准确率	词性标注准确率	
		一级	二级
式(3)	0.964 2	0.963 9	0.941 2
式(6)	0.971 5	0.964 0	0.943 2

从表 1 中可以看出,在式(3)中加入词的上下文信息后,分词的准确率明显提高,标注准确率也有相应的提高.

2 未登录词的识别与处理

2.1 未登录词识别与分词一体化

为了解决分词与未登录词识别的矛盾,引入分词与未登录词识别的一体化同步处理机制. 首先对原始字串进行初步切分,形成初始切分有向图,根据未登录词识别自动机在初始切分有向图中识别出所有的候选未登录词及其词性,加入到初始切分有向图中,让候选未登录词与候选普通词一起参与有向图中的路径选择,未登录词及其词性与普通词及词性在路径选择中同步确定. 在初始切分有向图中进行未登录词识别的自动机包括启动规则、归并规则和终止规则. 以地名为例,构造地名识别下推自动机,设 S 为起始符号,LF 为地名前词(性)[9] 集,LU 为地名用词(性)[9] 集,LS 为地名特征词[9] 集,UN 为未登录单字集(切分碎片),构造下推自动机:

$$S \to aA, A \to bA, A \to bd, A \to cd,$$

$$A \to c_1 c_2$$

其中 $a \in LF, b \in LU \cup UN, c \in LS, d \notin LU \cup UN \cup LS, c_1 \neq c_2$.

实际上,某种类型的识别自动机是将构造该类未登录词的人类知识形式化. 以“出生在聊城镇”为例,原始的切分有向图如图 1 所示.

其中“在”是地名前词(LF),“聊城”和“聊”是地名用字(LW),“城”和“镇”是地名特征词(LS),在对原始切分有向图进行未登录词识别之后有向图如图 2 所示.

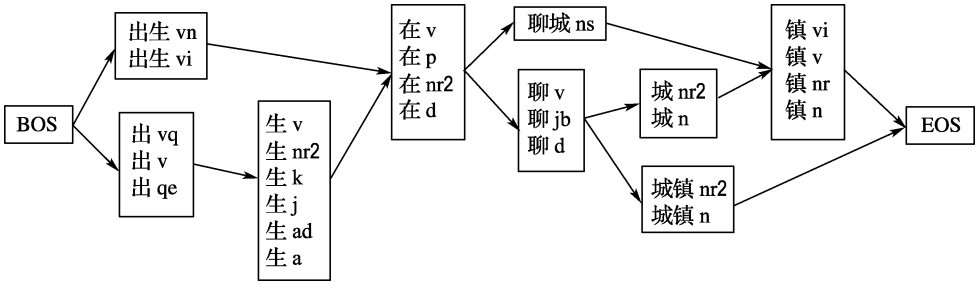

图 1 原始切分有向图

Fig. 1 Original segmentation directed graph

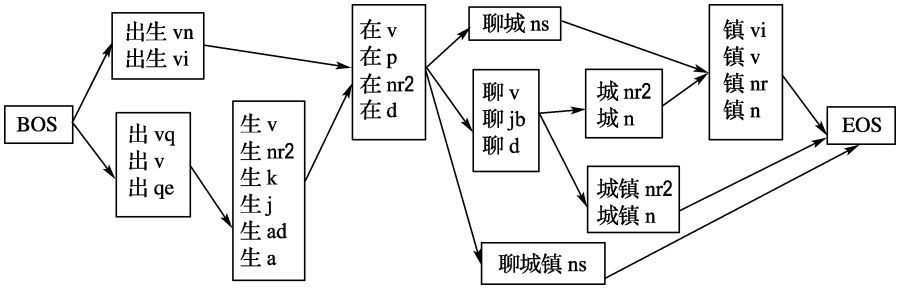

图 2 地名识别后的切分有向图

Fig. 2 Segmentation directed graph after location name recognition

遍历初始切分有向图，将识别出的候选未登录词(聊城镇/ns)加入到初始切分有向图中，将候选未登录词与候选普通词同等对待，共同参与候选路径的竞争.

2.2　不确定词性未登录词的一体化识别

传统的词法分析系统一般是对不同词性的未登录词(人名、地名、机构名等)分别进行识别处理[10~13]，无法解决不同未登录词之间可能出现的冲突现象. 本文引入对不确定词性未登录词进行同步识别的一体化机制，将不确定词性的未登录词作为多个候选，均加入到切分有向图中. 在识别候选未登录词的过程中，相同的上下文可能激活不同的未登录词识别过程，例如“在李华的帮助下”和“在泰山的东南麓”，其中人名“李华”和地名“泰山”具有相同的上下文“在…的”，“在”会激活人名和地名的识别过程. 另外，未登录词可能完全由未登录单字构成，并且前后文没有明显的起指示作用的特征词，此时未登录词词性无法确定，即未登录词的词性有多个候选. 在上述两种情况下，均激活不同词性未登录词的一体化识别过程. 不确定词性未登录词的一体化识别过程中的识别下推自动机描述如下：

设 S 为起始符号，UF 为未登录词前词[9](所有类别未登录词的前词)集，UN 为未登录单字集，则构造下推自动机：

$$S \to aA, A \to bA, A \to bc$$

其中 $a \in UF, b \in UN, c \notin UF \cup UN$.

3　算法实现

3.1　有向图中的节点、边及路径代价

为便于说明，定义有向图 $G = \langle V, E \rangle$ 为切分所构成的有向图，其中 V 是有向图中节点的集合，E 是有向图中边的集合，G 为一个有向无环图. 设 G 头节点(起始节点)为 v_s，尾节点(终止节点)为 v_e，头节点 v_s 和尾节点 v_e 被作为特殊的词处理，其中头节点 v_s 对应的词与词性分别为 SOS 和 ss，尾节点 v_e 对应的词与词性分别为 EOS 和 es. 设节点 v_i 对应的词为 w_i，对应的词性为 t_i，则有向图中的节点 v_i 的代价为

$$Cost(v_i) = -\log_2 p(w_i \mid t_i) \tag{7}$$

在有向图中，如果边 $e \in E$ 对应节点对 (v_i, v_j)，并且节点 v_i 对应的词与词性分别为 w_i、t_i，节点 v_j 对应的词与词性分别为 w_j、t_j，则有向图中边 e 的代价为

$$Cost(e) = Cost((v_i, v_j)) = -\log_2 p(t_j \mid t_i) - \alpha \log_2 p(w_j \mid w_i) \tag{8}$$

在有向图中的节点代价和边代价都确定后，可利用 Dijkstra 最小代价路径算法在有向图 G 中求出从头节点 v_s 到尾节点 v_e 的最小代价路径. 由前面所得到的节点及边代价公式可以得到有向图 G 的最小代价路径 $\min Cost((v_s, v_e))$ 的计算公式为

$$\min Cost((v_s, v_e)) = \arg\min_G \left\{ \sum_{i=1}^{n} (-\log_2 p(t_i \mid t_{i-1}) - \alpha \log_2 p(w_i \mid w_{i-1})) + \sum_{i=1}^{n} (-\log_2 p(w_i \mid t_i)) \right\} \tag{9}$$

其中 $\sum_{i=1}^{n}$ 表示的是对某一条路径中所有的节点或边求和.

3.2　有向图中未登录词节点的代价计算

在切分开放语料时，有向图 G 中可能会出现不同词性的未登录词节点，如人名、地名或机构名，未登录词的 $P(W \mid T)$ 与 $P(W)$ 无法从训练语料库中直接获得，相应的有向图 G 中的未登录词节点和边的代价也无法直接求出，为解决这个问题，引入未登录词的“修正条件概率”的概念来计算未登录词的 $P(W \mid T)$ 和 $P(W)$，用地名库中统计出来的词频 $P_1(W)$ 代替未登录词的词性条件概率 $P(W \mid T)$ 和 $P(W)$. 同样以地名为例，来说明 $P(W \mid T)$ 和 $P(W)$ 的计算. 从训练语料库中抽取出所有地名得到地名库，用不含地名的分词字典切分得到地名用字(词)库 D_1，统计地名切分结果得到地名用词的单词频度 P_d 和双词接续频度 P_s，在切分开放语料时，如果遇到的未登录词 W，$W = \{w_1, w_2, \cdots, w_n\}(w_i \in D_1)$，则采用 $P_1(W)$(W 在 D_1 中的概率)来代替 $P(W \mid T)$ 和 $P(W)$，替换公式为

$$P(W \mid T) = \lambda P_1(W) = \lambda(\phi P_d(W) + \varphi P_s(W)) \tag{10}$$

$$P(W) = \theta P_1(W) = \theta(\phi P_d(W) + \varphi P_s(W)) \tag{11}$$

其中系数 $\phi+\varphi=1$，λ、θ 为比例调整系数，$\theta/\lambda=P(T)$，在系统中取 $\phi=0.8$，$\varphi=0.2$，$\theta=0.9$. 式(10)和(11)中的 P_d 和 P_s 计算如下：

$$P_d(W)=\sqrt[n]{\prod_{i=1}^{n}P_d(w_i)}$$

$$P_s(W)=\sqrt[n-1]{\prod_{i=1}^{n-1}P_s(w_i,w_{i+1})};\ w_i\in D_l$$

由未登录词的修正条件概率式(10)、(11)可以得到未登录词节点在有向图 G 中的节点代价及包含未登录词边的代价

$$Cost(v_{un})=-\log_2\lambda(\phi P_d(w_{un})+\varphi P_s(w_{un})) \tag{12}$$

$$Cost(e_{un})=-\log_2 p(t_j\mid t_i)-\alpha\log_2[\theta(\phi P_d(w_{un})+\varphi P_s(w_{un}))] \tag{13}$$

式(12)、(13)中 v_{un} 是有向图中的未登录词节点，w_{un} 是节点 v_{un} 对应的未登录词.

4 实验结果与分析

在实验中采用北大《人民日报》语料库 2000 年半年语料，共 56.9 MB，1 000 余万字，75×10^4 词次，将该语料库分成训练语料库(80%)和测试语料库(20%)；并对其中的标注集进行了调整，去掉一些二级标注，如 wd、wf、wj 和 wky 等，增加一系列标注，如 ss(句首标点符号)、es(句末标点符号)、md(数量单位)等，最终生成包含 19 个大类(一级标注)、99 个小类(二级标注)的词性标注集. 将测试语料库平均分为 5 组，对以下 5 种模型分别进行测试：M 为未加入未登录词识别模块的模型；M_l 为 M 中加入地名识别模块的模型；M_n 为 M 中加入人名识别模块的模型；M_o 为 M 中加入机构名识别模块的模型；M_{lno} 为 M 中加入地名识别、机构名识别和人名识别模块的模型. 得到测试结果见表 2.

为了测试未登录词识别模块性能，在实验中分别利用 SIGHAN 2005[14,15] 中的北大开式分词语料与 SIGHAN 2008[16] 中 3 个简体中文分词开式语料(SXU、CTB、NCC)，对本文中的系统进行了测试，并将分词结果与用同一语料做开放测试的系统(选取其中 F 值最高者)作了比较，如表 3、4 所示，表中 R_{oov} 是对系统词表以外的词的召回率，R_{iv} 是对系统词表内的词的召回率.

表 2 测试结果

Tab. 2 Testing result %

模型		封闭测试			开放测试		
		P	R	F	P	R	F
M	分词	99.52	99.84	99.68	95.83	96.01	95.91
	词性	96.32	—	—	95.45	—	—
M_l	分词	99.34	99.44	99.39	96.14	96.51	96.32
	词性	97.02	—	—	96.02	—	—
M_n	分词	99.23	99.51	99.37	96.74	97.03	96.88
	词性	96.52	—	—	96.12	—	—
M_o	分词	99.04	99.18	99.11	96.02	96.71	96.36
	词性	96.84	—	—	96.03	—	—
M_{lno}	分词	99.14	99.23	99.19	96.94	97.59	97.26
	词性	97.96	—	—	96.63	—	—
	综合	97.39	98.01	97.70	96.54	97.10	96.82

注：最后一行综合测试为 M_{lno} 模型下的综合测试，综合测试时词性只计算一级标注，其他测试中词性标注为二级

表 3 本系统与 S19 在 SIGHAN 2005 PKU 开放测试语料上的比较

Tab. 3 Comparison between system in this paper and S19 in SIGHAN 2005 PKU open test %

PKU	R_{oov}	R_{iv}	R	P	F
S19	83.8	97.6	96.8	96.9	96.9
本系统	84.3	98.0	97.3	96.4	96.8

表 4 SIGHAN 2007 语料上的测试比较

Tab. 4 Comparison on corpus in SIGHAN 2007 %

		R_{oov}	R_{iv}	R	P	F
CTB	No31	90.89	98.05	97.66	97.21	97.43
	第一名	96.85	99.28	99.14	99.26	99.20
	本系统	87.31	98.65	96.89	97.42	97.15
SXU	No31	78.25	98.72	97.68	97.03	97.35
	第一名	78.25	98.72	97.68	97.03	97.35
	本系统	80.13	97.64	97.01	98.22	97.61
NCC	No31	63.37	97.83	96.20	94.96	95.57
	第一名	88.93	97.77	97.35	97.79	97.57
	本系统	85.43	97.11	97.62	96.48	97.05

从测试结果中可以看出，封闭测试时，加入未登录词识别模块前后，精确率及召回率变化不大，说明一体化模型较为稳定. 开放测试时，在未加入未登录词识别模块前，由于未登录词碎片的存在，精确率较低，并且召回率比精确率高. 在加入一种未登录词识别模块后，精确率和召回率都相应提

高，其中召回率提高很明显，而精确率比召回率稍低，主要是因为一种类型的未登录词识别自动机在识别产生候选未登录词时，将部分并不属于该类的未登录词也召回，这样就使得精确率受到影响. 加入全部未登录词识别模块后，模型 M_{lno} 的测试结果达到了实用要求. 在对未登录词识别的结果进行错误分析时，发现大多数错误集中在地名以及机构名的简称上，如“中油”、“上证”等，这是由于对地名和机构名的简称没有构建专门的识别自动机，有一部分地名和机构名简称被识别成人名，下一步将对这些地名和机构名简称进行学习，构建相应的识别自动机，进一步提高对未登录词识别的精确率.

5　结　语

本文针对现有词法分析系统所存在的问题，将分词、歧义消除、词性标注、未登录词识别等词法分析步骤整合到同一理论框架下，基于最长次长匹配，并拓展了传统隐马尔可夫模型，实现了同步完成词法分析的各步骤的一体化词法分析过程，减少了词法分析各步骤之间可能会出现的冲突现象. 扩展后的隐马尔可夫模型使各种词法信息在一体化模型中得到充分利用，从整体上提高了分词和词性标注以及未登录词识别的精度. 实验结果表明，这种一体化模型及其算法是有效的. 在处理未登录词时采用自动机识别候选与隐马尔可夫模型相结合，保证了系统处理效率. 今后将结合使用其他统计模型，如支持向量机模型、最大熵模型以及条件随机场模型，以进一步提高整个系统的词法分析精确率.

参考文献：

[1] 黄德根，朱和合，王昆仑，等. 基于最长次长匹配的汉语自动分词[J]. 大连理工大学学报，1999，**39**(6)：831-835
(HUANG De-gen, ZHU He-he, WANG Kun-lun, *et al*. Chinese automatic words segmentation based on maximum matching and second-maximum matching [J]. **Journal of Dalian University of Technology**, 1999, **39**(6):831-835)

[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

[3] JIANG F, LIU H, CHEN Y Q, *et al*. An enhanced model for Chinese word segmentation and part-of-speech tagging [C] // **ACL SIGHAN Workshop 2004**. Barcelona: Association for Computational Linguistics, 2004:28-32

[4] 高　山，张　艳，徐　波，等. 基于三元统计模型的汉语分词及标注一体化研究[C] // 自然语言理解与机器翻译——全国第六届计算语言学联合学术会议论文集. 北京：清华大学出版社，2001:116-122

[5] SUN M S, XU D L, BENJAMIN K T. Integrated Chinese word segmentation and part-of-speech tagging based on the divide-and-conquer strategy [C] // **Proceedings of 2003 International Conference on Natural Language Processing and Knowledge Engineering**. Beijing: IEEE Computer Society, 2003: 610-615

[6] ZHANG Y, CLARK S. Joint word segmentation and POS tagging using a single perceptron [C] // **Proceedings of ACL2008**. Columbus: Association for Computational Linguistics, 2008:888-896

[7] GAO J F, LI M, HUANG C N. Improved source-channel models for Chinese word segmentation [C] // **Proceedings of ACL2003**. Sapporo: Association for Computational Linguistics, 2003:272-279

[8] GAO J F, WU A D, LI M, *et al*. Adaptive Chinese word segmentation [C] // **Proceedings of ACL2004**. Morristown: Association for Computational Linguistics, 2004:462-469

[9] 黄德根，岳广玲，杨元生. 基于统计的中文地名识别[J]. 中文信息学报，2003，**17**(2)：36-41

[10] 黄德根，朱和合，杨元生. 基于单词与双词可信度的汉语自动分词[J]. 计算机研究与发展，2001(7)：132-135

[11] 张华平，刘　群. 基于角色标注的中国人名自动识别研究[J]. 计算机学报，2004，**27**(1)：85-91

[12] ZHANG H P, LIU Q, ZHANG H, *et al*. Automatic recognition of Chinese unknown words

recognition [C] // **First SIGHAN Workshop Attached with the 19th COLING.** Taipei: Association for Computational Linguistics, 2002:71-77

[13] KIT C Y, XU Z M, WEBSTER J J. Integrating n-gram model and case-based learning for Chinese word segmentation [C] // **Proceedings of SIGHAN-2.** Sapporo: Association for Computational Linguistics, 2003:160-163

[14] RICHARD S, THOMAS E. The first international Chinese word segmentation bakeoff [C] // **First SIGHAN Workshop Attached with the ACL2003.** Taipei: Association for Computational Linguistics, 2003:133-143

[15] THOMAS E. The second international Chinese word segmentation bakeoff 2005 [C] // **The 4th SIGHAN Workshop.** Korea: Association for Computational Linguistics, 2005:123-133

[16] JIN G J, CHEN X. The fourth international Chinese language processing bakeoff: Chinese word segmentation, named entity recognition and Chinese POS tagging [C] // **Sixth SIGHAN Workshop on Chinese Language Processing.** Hyderabad: Association for Computational Linguistics, 2008:69-81

Chinese integrative lexical analysis based on maximum matching and second-maximum matching segmentation

SUN Xiao[*1,2], HUANG De-gen[1]

(1. School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China;
2. School of Computer Science and Engineering, Dalian Nationalities University, Dalian 116600, China)

Abstract: An integrative lexical analysis mechanism is proposed in order to solve the limitation of mostly existing lexical analysis system with "pipelining" mechanism. Based on maximum matching and second-maximum matching (MMSM) model, in the directed graph built by MMSM model, candidate words, parts-of-speech (POS) tags and all the candidate unknown words are added and considered, hidden Markov model (HMM) is extended, so Chinese word segmentation, ambiguity resolution, unknown word recognition and POS tagging are solved synchronously. The integrations of word segmentation and POS tagging, unknown words recognition and known word segmentation, uncertain unknown words recognition are realized. All the tasks of lexical analysis are accomplished synchronously, the conflicts between all the tasks in the Chinese lexical analysis are avoided, and high precision can be gained. The open test indicates that the F-score of the system is 98.03%.

Key words: Chinese lexical analysis; integrative model; maximum matching and second-maximum matching; unknown word; segmentation directed graph