基于多知识源的中文词法分析系统

姜 维 王晓龙 关 毅 赵 健
(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 汉语词法分析是中文自然语言处理的首要任务.文中深入研究中文分词、词性标注、命名实体识别所面临的问题及相互之间的协作关系,并阐述了一个基于混合语言模型构建的实用汉语词法分析系统.该系统采用了多种语言模型,有针对性地处理词法分析所面临的各个问题.其中分词系统参加了 2005 年第二届国际汉语分词评测,在微软亚洲研究院、北京大学语料库开放测试中,分别获得 F 量度为 97.2%与 96.7%.而在北京大学标注的《人民日报》语料库的开放评测中,词性标注获得 96.1%的精确率,命名实体识别获得的 F 量度值为 88.6%.

关键词 词法分析;汉语分词;词性标注;命名实体识别;语言模型
中图法分类号 TP391

Research on Chinese Lexical Analysis System by Fusing Multiple Knowledge Sources

JIANG Wei WANG Xiao-Long GUAN Yi ZHAO Jian
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

Abstract Chinese lexical analysis is the foundation task for most Chinese natural language processing. In this paper, word segmentation, POS tagging, named entity recognition and their relation are well discussed. Moreover, a pragmatic lexical analysis system based on mixed language models is presented, which adopts many models, such as n-gram, hidden Markov model, maximum entropy model, support vector machine and conditional random fields, they have good performance in the special sub-tasks. The Word Segmenter participated in the Second International Chinese Word Segmentation Bakeoff in 2005, and achieved 97.2% and 96.7% in terms of F-measure in MSR and PKU open test respectively. While the POS tagging and named entity recognition modules achieved 96.1% in precision and 88.6% in F-measure respectively in open test with the corpus that came from six-month corpora of Chinese Peoples' Daily.

Keywords lexical analysis; Chinese word segmentation; part-of-speech tagging; named entity recognition; language model

1 引 言

词法分析主要包括分词、词性标注与命名实体识别三项子任务,它是句法分析与语义分析的基础,其性能将直接影响到后续应用,如机器翻译、信息抽取、问答系统的性能.本文以国家自然科学基金重点项目"问答式信息检索的理论与方法"为背景,全面

收稿日期:2005-11-15;修改稿收到日期:2006-06-06.本课题得到国家自然科学基金重点项目"问答式信息检索的理论与方法"(60435020)及国家自然科学基金(60504021)资助.**姜 维**,男,1978 年生,博士研究生,研究方向为自然语言处理、词法分析、信息抽取.E-mail: jwSeaBreeze@hit.edu.cn.**王晓龙**,男,1955 年生,教授,博士生导师,主要研究领域为人工智能、自然语言处理.**关 毅**,男,1970 年生,博士,副教授,研究方向为问答系统、Web 挖掘.**赵 健**,男,1975 年生,博士,研究方向为中文命名实体识别、信息抽取.

阐述一个实用的词法分析系统的构成、实现以及探讨词法分析的未来研究方向.

1.1 词法分析研究的现状

词法分析的研究主要可被划分为三类:一是基于规则的方法,如最大匹配分词、基于错误驱动的词性标注、基于规则的命名实体识别等方法;二是基于统计的方法,如 n-gram 模型分词、隐马尔科夫(HMM)词性标注与最大熵(ME)命名实体识别等;三是统计与规则相结合的混合方法,该方法可综合利用语言统计信息与语言本身的知识,往往具有更好的性能,如 Zhang[1] 采用的层次隐马尔科夫模型、Gao[2] 采用的基于类的语言模型.

分词问题上,还有学者提出基于字的 ME 模型[3] 与 CRF 分词模型[4]. 虽然该两类模型在利用众多特征后,比 Trigram 模型要好一些,但从运行效率来看,前者基于字处理,需要计算大量特征. 而 Trigram 模型采用基于词的处理方式,具有较高的运行效率. 在应用平滑算法后,Trigram 完成词典词的切分通常具有约 98%~99%的精度[5].

词性标注除基于统计的方法(如 HMM,ME)外,还存在基于规则的方法,如 Transformation Based Learner(TBL)以及基于统计决策树(SDT)方法. 基于规则的方法适应性较差,并且非统计模型的本质使得它不能给出每种可能分类结果的概率值,因此通常其作为一个独立分类器,而很难被用作一个更大概率模型的组件部分;SDT 虽然可以加入丰富的特征,但是在处理基于词的自然语言问题时,需要数据划分,因而存在严重的数据稀疏问题,应用时需要复杂的平滑技术.

在中文命名实体任务上,HMM 方法[1-2]、ME 方法[6] 以及 SVM 方法是目前比较常见的统计模型. 文献[1]采用基于角色的策略可以显式地描述先验知识(即由人来指明特征). 文献[2]采用了基于类的处理方法. 现有的识别模型常视其为序列标注问题,所以采用的模型通常与词性标注模型相同.

1.2 本文词法分析的方法

本文基于以下观点建立词法分析系统:(1)分词、词性标注、命名实体识别之间的协调处理能够有助于改善整个词法分析系统的性能;(2)采用易于融合更多统计特征与语言知识的模型有助于改善词法分析系统的性能;(3)恰当的特征集(如增加远距离特征)将有助于改善词法分析系统的性能. 因此,只有当系统能够较好地描述词法知识,才能获得好的词法分析性能.

机器学习中的"没有免费的午餐"定理已指出,必须设法寻找更适合当前任务的语言模型,而"丑小鸭"定理指出我们必须去寻找适合当前任务的有效特征,二者都恰好强调了先验知识的重要性. 从信息增益角度来看,多种知识源分析的方法也正是设法充分地利用先验知识,来提高词法分析中的各个子任务的性能. 在目前无法基于单一模型来完美地、一体化处理全部任务的情况下,本文协调地处理各个任务之间的依赖关系会是有益的.

基于以上的三个观点,本文所做的工作如下:首先,鉴于三项子任务相互交织,从信息增益角度我们采用协调利用彼此处理过程的知识;其次,在考虑运行效率的情况下,采用较优越的语言模型,并提出应用粗糙集理论获取远距离特征与复杂特征,将其融入语言模型中;最后,我们提出双层混合命名实体识别体系结构,试图克服变化实体切分等因素带来的影响,并有效地融合多种识别模型.

2 词法分析系统结构

从计算语言角度来看,分词、词性标注、命名实体识别面临着不同的任务. 分词可被看作序列切分的过程;词性标注则是序列标注的过程;而命名实体识别的过程则不仅需要识别实体的边界,还需要识别出实体的类型. 正是由于各自的任务不同,因此较难采用单一模型处理全部问题. 从已知的技术来看,若强制采用单一模型统一来处理,会造成模型过于复杂、模型规模大、运行效率慢、难于训练或者难于进一步改善模型性能等缺点. 另一方面,任务分解并不意味着各个子任务之间彼此独立. 实质上,这三项任务相互交织、相互促进:命名实体不仅是单独的任务,也是分词与词性标注中未知词标注的重要组成部分;相比分词与词性标注任务,命名实体识别的任务更为困难. 鉴于以上问题,本文从已有的技术角度出发,倾向于运用更有效的模型解决特定的任务,再有机地结合各项处理结果.

图 1 中,基本分词模块完成切分词典词、仿词与派生词以及识别新词的任务,同时还识别出仿词与派生词的类型. 评测实验表明基本分词模块的 R_{iv} 指标性能约为 98%①,而基本词性标注在不考虑未知词与复杂虚词时,可获得约 97%的标注精度. 前

① 本文采用北京大学的分词、词性标注、命名实体的定义标准. 该值为 Sighan2005 的评测结果[5].

两步的处理结果为命名实体识别提供较为准确的词信息与词性信息. 反过来，在分词与词性标注过程中出现的未知词中，命名实体占有主要部分，相比来说它更难于处理. 而词特征与词性特征会有助于命名实体的识别. 尽管如此，我们不能忽略前续操作中的错误切分带来的影响，例如："张/华/平等"中错误切分"平等"，从而易使识别过程无法复原实体[1]，这将在"歧义边界判别"模块中得以修正.

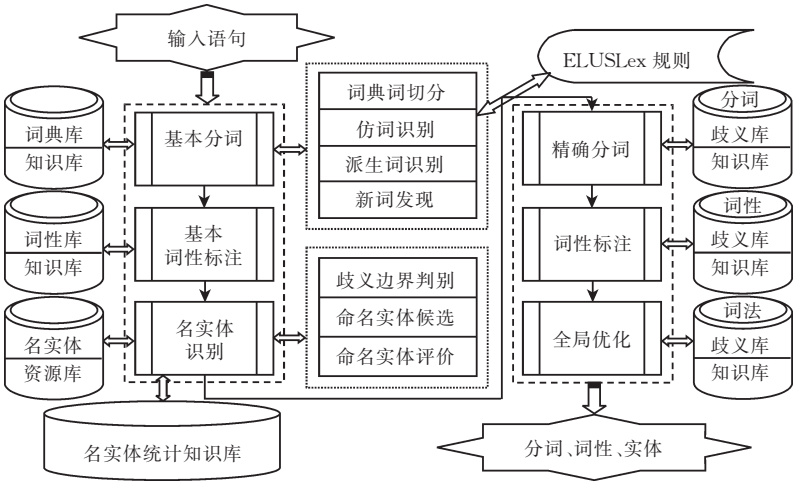

图 1　ELUS 词法分析系统的体系结构

对于复杂歧义可在前续处理后利用更加高级的特征进行消解[7]，如远距离约束"只有→才能"用于消歧"才能"或"才/能"，所以在"精确分词"部分主要针对这类复杂歧义部分进行消歧处理. 在完善分词之后，词性标注过程需对重切分句子重新标注词性；用名实体识别结果标注词性；同时利用消歧模型对复杂兼类词进行消歧.

3　词法分析子系统构成

3.1　仿词与派生词

仿词(factoid word)属于一种未知词，其类型如表 1. 仿词变化较多，但是同类型仿词的作用相似，根据仿词类型还能确定其词性，并为分词、词性标注、命名实体识别模型提供稳定的特征.

表 1　仿词的类型

类型	包含的仿词	示例
Number	Integer, real, etc.	2910, 46.12, 二十九
Date	Date	5 月 12 日，2004 年
Time	Time	8∶00, 十点二十分
English	English word,	How, are, you
www	Website, IP address	http://www.hit.edu.cn
Email	Email	jwSeaBreeze@ hit.edu.cn
phone	Phone, fax	0451-86413322-85

有限自动机(FSA)是识别仿词的有力方法，而人们更习惯于书写产生式规则，如表 2. 为此，我们制作三个工具：LexEdit, LexCompiler, LexRunner 用于编辑、编译、运行 ELUSLex 语法规则.

表 2　ELUSLex 规则举例

```
〈digit〉—>[0..9]|[0..9];          //定义英文数字
〈integer〉::={〈digit〉+};          //定义英文整数
〈real〉::=〈integer〉(.|.|•|点)〈integer〉; //小数
〈day〉—>〈integer〉日;              //定义"日"
〈month〉—>〈integer〉月;            //定义"月"
〈year〉—>〈digit〉〈integer〉年;     //定义"年"
〈date〉::=〈year〉〈month〉〈day〉;   //定义日期
```

对于派生词(morphological derived word)的识别，我们采用规则与词典相结合的方法. 首先利用规则搜集一些可能的派生词，例如："一排排、一座座"可以采用规则"一 AA"；"高高兴兴、快快乐乐"可采用规则"AABB"；"进出口"可被判别为两个词"进口"与"出口"的派生词. 最后通过对派生词候选进行人工校对来获得派生词词典.

3.2　汉语分词

分词作为许多任务的最基本过程，需要具备较高的处理效率. 本文采用 Trigram 作为分词的基本处理模型，Trigram 模型基于马尔科夫条件假设，以词作为基本特征，因而具有较高的切分效率.

如果以词的类别代替词本身，即词典词类别是其本身，仿词类别为仿词的类型，词形词类别为词形词的根，那么在 trigram 中，设 $w=w_1w_2\cdots w_n$ 代表词类序列，s 为输入语句，那么最佳序列 w^* 为

$$w^*=\arg\max_w P(w)P(s|w)$$

$$=\arg\max\prod_{i=1}^{n}P(w_i|w_{i-2}w_{i-1}) \qquad (1)$$

其中，设 w_{-2}，w_{-1} 分别代表句子额外附加的两个头

节点.于是可以利用 Viterbi 算法搜索词网格(如图 3),获得式(1)中具有概率最大的最优切词序列.

但直接应用 Trigram 会存在数据稀疏问题.例如若语料库未出现"[Date]纪念[Number]",则按照极大似然估计(MLE),P([Number],[Date],纪念)=0,此时无法计算路径概率.为此基于高阶模型通常具有更好的描述精度但是更加稀疏,而低阶模型具有较差的描述精度却存在较少的数据稀疏问题的假设,我们可以运用插值或者回退平滑算法,结合高低阶规则的优点.本文运用绝对平滑性算法克服数据稀疏问题,算法描述如下:

$$N_{1+}(w_{i-n+1}^{i-1}\bullet)=|\{w_i:c(w_{i-n+1}^{i-1}w_i)>0\}|,$$

其中·代表任意变量,$c()$代表计数函数.此时转移概率为

$$p(w_i|w_{i-n+1}^{i-1})=\frac{\max\{c(w_{i-n+1}^{i})-D,0\}}{\sum_{w_i}c(w_{i-n+1}^{i})}+\frac{D}{\sum_{w_i}c(w_{i-n+1}^{i})}N_{1+}(w_{i-n+1}^{i-1}\bullet)p(w_i|w_{i-n+2}^{i-1}) \quad (2)$$

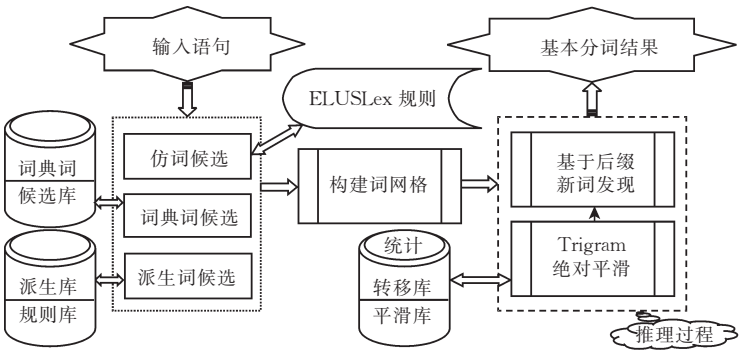

图 2 基本分词过程

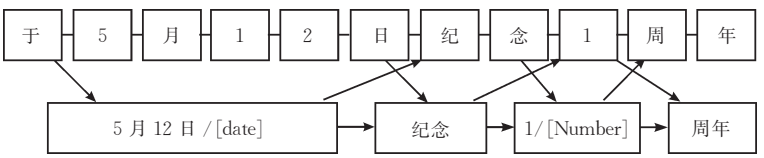

图 3 基于类的词网格构建过程

由于我们采用 Trigram 模型,n 的最大取值为 3.参数 D 是折扣(discount)参数,用于对每个非零计数进行折扣,D 可通过训练语料上的删除算法来估计:$D=\frac{n_1}{n_1+2n_2}$,其中 n_1 与 n_2 分别代表语料库中出现次数为 1 与 2 的对象个数.

正确识别一些基于后缀的新词是有意义的,如"景观+灯"、"海风+牌"等.我们通过方差方法收集构词力较强的后缀字,通过最大熵模型识别[5].

表 3 基于方差获取新词后缀

	S_1	S_2	…	S_m
W_1	c_{11}	c_{21}	…	c_{m1}
W_2	c_{12}	c_{22}	…	c_{m2}
…	…	…	…	…
W_n	c_{1n}	c_{2n}	…	c_{mn}

表 3 中 $S_1\cdots S_m$ 为 m 个候选后缀,$W_1\cdots W_n$ 为 n 个位于词典中各个去除后缀字的词前缀.例如:S_1 代表"灯",W_1 代表"景观",W_1S_1 构成"景观灯".设 $C_{xy}=Count(S_x,W_y)$,$CV(S_x)=Count(c_{x2}>0)$,则 $Sum(S_x)=\sum_{i=1}^{m}c_{xi}$, $avg(S_x)=Sum(S_x)/CV(S_x)$, $p_{xi}=C_{xi}/Sum(S_x)$, $V_{xi}=p_{xi}\times(C_{xi}-avg(S_x))\times(C_{xi}-avg(S_x))$.于是方差 $V(Sx)=\sum_{i=1}^{m}V_{xi}$.此外还存在两个因素:(1)该后缀字语料库出现的次数;(2)该后缀字所构成词的种类数.通过上述三种因素限定,我们截取前 25 个后缀字作为有意义的新词后缀,如制、牌、型、式[5].

3.3 汉语词性标注

词性标注是为词序列语句中的各个词标记相应的词性.在图 4 中,输入的词序列是经过分词后的语句,如:"于/5 月 12 日[Date]/纪念/1[Number]/周年".由于同类型的仿词具有相似的语法功能,所以按其类别,得到"于 [Date] 纪念 [Number] 周年".

HMM 标注模型所用的上下文特征为转移概率与发射概率,其分类器决策为

$$T^{\#}=\arg\max_{T_1T_2\cdots T_n}\prod_{i=1}^{n}P(W_i|T_i)P(T_i|T_{i-n}^{i-1}) \quad (3)$$

其中，T_i代表当前词的词性标记(当 i 小于 0 时代表所附加的语句头节点的词性标记).

相比 HMM 模型，MEMM 分类器决策规则为

$$T^{\#}=\arg\max_{T_1T_2\cdots T_n}P(T_1\cdots T_n\mid w_1\cdots w_n)$$

$$=\arg\max_{T_1T_2\cdots T_n}\prod_{i=1}^{n}P(T_i\mid h_i) \tag{4}$$

其中 $w_1\cdots w_n$代表输入语句的序列. $P(T_i\mid h_i)$是为每个词的词性分配的条件概率，h_i为当前上下文.

CRF 是一种无向图模型. 在给定一系列输入随机变量值 X 的情况下，一系列输出随机变量值 Y 的条件概率 $P(Y\mid X)$定义为与无向图中各个团(cliques)的势函数(potential function)的乘积成正比. 而在常用的链状 CRF 模型中，势函数一般被定义为团的所有特征的带权和的指数形式[8]. 此时，对于输入句子 x，最佳词性标注序列 y^* 可由下式表示[9]：

$$y^*=\arg\max_y p_{\boldsymbol{\lambda}}(y\mid x)=\arg\max_y\left(\frac{\exp(\boldsymbol{\lambda}\cdot F(y,x))}{Z_{\boldsymbol{\lambda}}(x)}\right) \tag{5}$$

其中，$Z_{\boldsymbol{\lambda}}(x)$为归一化因子，权重向量 $\boldsymbol{\lambda}$ 值可采用 L-BFGS 算法训练获得[9]. 相比上述 HMM 模型公式，CRF 模型不需独立性假设，可以方便融入更多上下文特征；而相比 MEMM 模型，二者的不同是：MEMM 定义在单个观察对象上，而 CRF 条件概率定义在整个序列上，这是由于全局特征 $F(y,x)=\sum_i f(y,x,i)$是通过若干个局部特征获得，所以 CRF 可以有效解决标记偏置问题.

与 MEMM 相同，CRF 通过特征模板抽取上下文特征，我们的系统采用的特征模板如表 4 所示.

表 4　词性标注中的特征模板

特征类型	特征模板
近距离特征	w_{i-2}，w_{i-1}，w_i，w_{i+1}，w_{i+2}，$w_{i-1,i}$，$w_{i,i+1}$

例如，对关键词“纪念”通过表 4 所示的模板来抽取，此时图 4 中的特征抽取模块抽取的特征为：“$w-2$：于 $w-1$：[Date] $w0$：纪念 $w1$：[Number] $w2$：周年 $w-1,0$：[Date]：纪念 $w0,1$：纪念：[Number]”. 由此可知，相比 HMM，CRF 可以更充分地利用上下文特征.

若从提高标注性能的角度出发，可以考虑进行多模型组合，例如后面实验中通过投票法组合多种分类结果可以获得更好的标注性能. 但实验表明其性能提高不大，考虑到运行效率，往往不被采用.

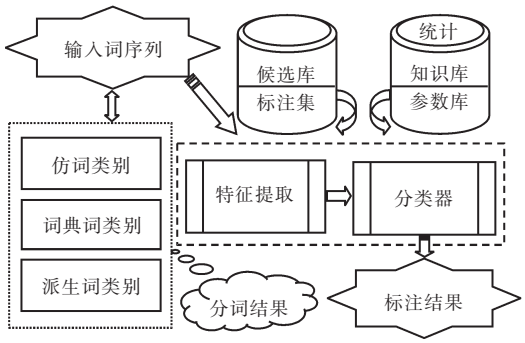

图 4　基本词性标注过程

3.4　汉语命名实体识别

命名实体主要分为三类：实体类(人名、地名、机构名)、时间类(日期、时间)、数量类(金钱值、百分比). 其中，时间类与数量类可通过仿词识别就可获得较好的识别效果；而实体类中的人名、地名、机构名的识别却较为困难，故将其作为本节的重点.

表 5　名实体分布情况统计

命名实体	占名实体百分比/%	占语料库百分比/%
人名	15.59	1.69
地名	23.69	2.57
机构	10.07	1.09
时间	16.33	1.77
数量	34.43	3.72

不同于英文，中文命名实体识别任务不仅需识别实体类型还需判别实体边界. 这种不确定的实体边界通常会影响实体的识别效果. 例如“对{张/红}说”、“对{孙/桂/平}说”，其中“对[Person]说”是较好的语法规则，却因二者切分不同，不易获得一致特征. 类似的特征如组块特征等也受到该问题的严重影响. 此外，一些切分错误，如将“张华平/等”切分为“张/华/平等”，也会影响实体识别. 既然稳定的上下文环境能为识别具有不同表现形式的同类型实体提供重要信息，于是我们提出采用双层混合模型(如图 5)，克服上述问题.

从归纳偏置角度看，图 5 的混合模型并未比单一模型做更多的假设，相反，它恰是设法利用单一模型的优点：n 个序列标注器从多种角度来提高边界识别的召回率，并利用一定的规则融合这些序列标注器的标注结果生成实体候选，再通过实体判别器提高实体识别的精度. 一般来说，单一模型实体标注方法可作为序列标注器，如 HMM、ME 等；而实体判别器可在相对稳定的上下文环境中判别实体，从而易于使用更多的先验知识来提高识别效果.

本文以 HMM、ME 作为序列标注器，处理中文人名(PER)、中文地名(LOC)和中文机构名

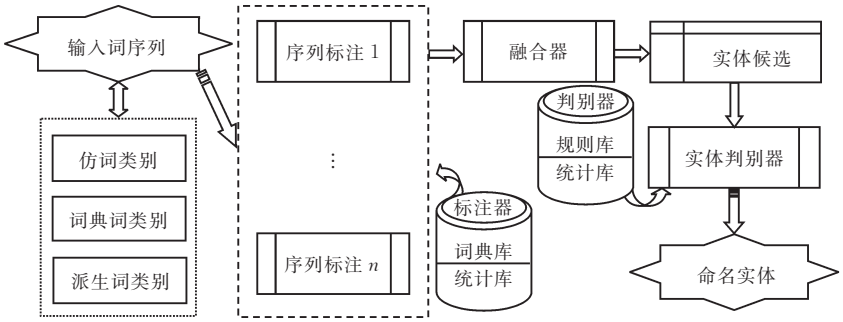

图 5 双层混合模型命名实体识别体系结构

(ORG). 其中,为了便于采集特征和识别,中文人名被分成了中国人名(CPN)和外国人名的中文译名(FPN). 采用 B-X,I-X 结构标记每一类实体,"O"为非名实体的类别,所以总共有 $4\times2+1=9$ 种标记. 此时,实体识别过程类似于词性标注的过程,如"我/O 是/O 黄/B-CPN 中/I-CPN 敏/I-CPN 的/O 朋友/O". 一些资源词典相当于先验知识,可以提高模型识别性能,如地名"北京,纽约,马家沟",姓"张,王,孙",人名前缀"老,阿,小",地名后缀:"山,湖,海",机构名后缀"会,组织,局"等.

由于融合有效先验知识会有助于提高分类器的性能,我们显式地指明实体的前缀、后缀以及连接两个同类型实体之间的中缀词[1]. 如图 6 所示.

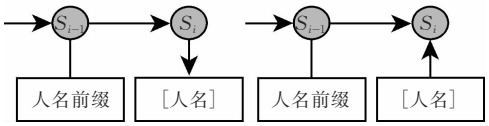

图 6 HMM 模型(左)与 ME 模型(右)

图 6 的 HMM 模型中,在计算路径概率时,从公式 $P(S_i|S_{i-1})\times P([\text{人名}]|S_i)$ 中的转移概率 $P(S_i|S_{i-1})$ 可以看出,丰富、恰当的状态 S_{i-1} 将有助于精细地描述 $P(S_i|S_{i-1})$. 同理,这种先验知识完全可以应用于 ME 标注中. 设 p-X,s-X 分别代表实体类型 X 的前缀词、后缀词. 而实体中缀不区分类型,标记为 M.

ME 模型的特征模板如表 6(为简化,此处只用基本特征,可参考文献[6]).

表 6 最大熵命名实体识别基本特征模板

特征类型	特征模板(w 代表词,t 代表标记)
一阶特征	w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, t_{i-1}
二阶特征	$w_{i-1,i}$, $w_{i,i+1}$

经过 HMM 与 ME 标注生成的实体候选需要融合生成实体候选,例如:

HMM:位于 98 号 的/p-ORG 杭州/B-ORG 四季青/I-ORG 乳品厂/I-ORG.

ME:位于 98 号 的/p-LOC 杭州/B-LOC 四季青/I-LOC 乳品厂/I-ORG.

上面的例子中,"杭州 四季青"被 ME 错标记为地名. 而汉语中以地名开头的机构名是很多的,例如"哈尔滨秋林公司". 因此上例中,可采用 HMM 标记结果修正 ME 的识别. 融合器的规则主要来源于统计分析 ME 与 HMM 自身分类器的错误,寻找能够互补的有益规则. 显然,这种方法可以方便扩展到多个序列标注器的问题上.

目前,实体判别器采用人工规则与统计结合的方法,其主要的三个策略是:(1)如果两个序列具有相同识别的候选实体,则确认该实体;(2)如果两个实体具有不同的切分边界,则通过交叉熵来判别;(3)如果候选的实体存在两种可能的类别,例如地名或机构名,则还应判别词本身更符合哪个类别. 交叉熵由插值二阶与三阶概率获得:

$$H_c(W)=-\frac{1}{N_W}\sum_{i=1}^{N_W}(\lambda\log_2 P(w_i|w_{i-2}w_{i-1})+(1-\lambda)\log_2 P(w_i|w_{i-1})) \quad (6)$$

其中 λ 为经验参数,用于确定三阶概率所起到的作用. 该公式用于判别及选取相交叉的候选实体.

经验上,实体判别过程也应该可以基于机器学习算法完成,例如支持向量机、最大熵等模型. 但无论哪种方法,针对问题融入有效的先验知识是提高系统性能的重要因素.

3.5 精加工处理

经过基本分词、基本词性标注、命名实体识别后,仍会存在以下的歧义现象,例如:

仿词中时间"25 年"与时间段"25/年",分钟"30 分"与比赛得分"30/分";

复杂切分歧义,如代表一个人能力的"才能"与表示语句结构的"才/能";

复杂兼类词歧义,如"为"的介词与动词词性;

一些命名实体由于在某些上下文表现不明显而未被正确识别.

对于歧义只能从歧义自身信息与上下文信息入手.倘若自身能完全解决的歧义就是所谓的假歧义，而对于真歧义只能借助于上下文信息.因此，消歧过程正是设法充分利用上下文完成歧义消解的过程.笔者在文献[7]中给出基于最大熵模型的分词歧义消解算法.对于一些虚词，例如“为”，因为其词性往往受到句子句法结构的影响，近距离上下文不易判别其词性，我们称其为复杂兼类词.对于这类兼类词，挖掘有效的长距离特征是必要的，如：

这种/r 现象/n 为/v 建国/v 以来/f 所/u 罕见/a 。/w

深/d 为/p 老人/n 们/k 的/u 真情/n 所/u 感动/a 。/w

易看出，约束“(为/v)<—(所 罕见)”，“(为/p)<—(所 感动)有助于标记“为”的词性.为此，我们引入可变精度的粗糙集理论，可有效地提取这类复杂特征.关于粗糙集提取特征的具体方法以及详细的实验评价，笔者在文献[10]中给予了详细地阐述.

4　实验分析

4.1　分词评测

本系统参加了 2005 年第二届国际汉语分词评测(http://www.sighan.org/bakeoff2005/)，表 7、表 8 给出了系统的评测性能[5].为了适应 Sighan 评测规则，评测中未使用词性特征，并且在封闭测试中未使用命名实体识别模块.

表 7　Sighan2005 封闭测试结果

封闭	召回率/%	精确率/%	F 量度/%	*OOV*	R_{oov}/%	R_{iv}/%
PKU	95.4	92.7	94.1	5.8	51.8	98.1
MSRA	97.3	94.5	95.9	2.6	32.3	99.1
CITYU	93.4	86.5	89.8	7.4	24.8	98.9
AS	94.3	89.5	91.8	4.3	13.7	97.9

表 8　Sighan2005 开放测试结果

开放	召回率/%	精确率/%	F 量度/%	*OOV*	R_{oov}/%	R_{iv}/%
PKU	96.8	96.6	96.7	5.8	82.6	97.7
MSRA	98.0	96.5	97.2	2.6	59.0	99.0
CITYU	94.6	89.8	92.2	7.4	41.7	98.9
AS	95.2	92.0	93.6	4.3	35.4	97.9

在所有参赛系统中，本系统在 MSRA 与 PKU 语料库的开放测试分别获得第一名与第二名.此外，R_{iv} 获得 97.7%～99.1%的表现性能，说明本文的分词以及分词歧义消解模型已经具有非常好的表现能力.此外，简体中文的 R_{oov} 指标也排在前列，这说明命名实体识别模块也表现出良好的性能.

下面，我们单独测试图 1 中精确分词模块的消歧性能.在 2000 年《人民日报》中搜集带有复杂歧义切分的语句数据，利用基本分词模块进行切分，然后通过人工对歧义切分进行标注来获得训练与测试语料[7].实验对最大熵模型消歧与增加粗规则特征的最大熵模型消歧进行对比，结果如表 9 所示.

表 9　一些典型复杂歧义切分的消歧实验结果

歧义切分	切分类型	训练数	测试数	最大熵消歧/%	增加粗规则/%
才能	才能	704	190	90	93
	才/能	7612	300		
不要	不要	1421	150	91	95
	不/要	497	80		
从小学	从小/学	170	40	88	91
	从/小学	260	70		
将来	将来	1200	200	92	97
	将/来	35	10		
个人	个人	1016	150	89	94
	个/人	819	120		

表 9 中的实验结果表明，粗糙集提取远距离特征可以有效提高歧义词的切分精度.

4.2　词性标注评测

实验条件语料库为北京大学标注的 1998 年上半年《人民日报》.其中采用前 5 个月《人民日报》进行训练，第 6 个月语料进行测试.分别对 HMM(内嵌绝对平滑算法)、ME、MEMM 以及 CRF 方法进行词性标注实验，对比结果如表 10 所示.

表 10①　几种模型的词性标注性能对比

标注模型	训练时间	标注时间/s	标注精度/%
HMM-1	3min	15	94.47
HMM-2	7min	23	94.55
ME-1	1d	70	95.52
ME-2	1d	188	95.57
MEMM	1d	90	95.54
CRF	20d	637	96.10

注：(1) HMM-1 代表 1 阶 HMM；HMM-2 为 2 阶 HMM；
(2) ME-1 采用贪心算法进行标注；
(3) ME-2 采用 BeamSearch 算法，搜索宽度 5.

表 10 表明 CRF 模型可以获得更高的标注精度 96.10%，该精度相比 HMM 提高 1.55%，比 MEMM 提高 0.56%.相比 HMM，CRF 利用了更多的观察特征；相比 MEMM，CRF 克服了标注偏置问题.

下面考察多分类器组合的词性标注性能.表 11 中所有的单一模型采用第 1 个月的人民日报语料训

① 训练与标注时间包含操作文件时间.本实验中，ME 与 MEMM 的特征模板为 w_{i-2}，w_{i-1}，w_i，w_{i+1}，w_{i+2}，特征过滤阈值均为 5.本次对比实验中的所有算法均采用标准 C++实现，然而算法执行时间与实现有关，这里仅供参考.

练,测试采用第 6 个月的人民日报.为了更好地对比结果,我们去掉了命名实体识别等处理过程.

表 11 多模型融合的词性标注性能

语言模型	标注精度/%
HMM2 (2-order)	92.39
MEMM	92.50
SVM	92.89
CRF	93.59
Voting Combination	93.73

表 11 表明:相比其它单一模型,CRF 模型具有更好的性能;而基于投票法的组合模型获得最优的标注性能.但就实际应用而言,多模型会降低标注效率,因此需要根据实际情况考虑是否采用该方法.

4.3 命名实体识别评测

训练语料为北京大学标注的 1998 年 1 到 5 月的《人民日报》,测试语料为 6 月份《人民日报》语料.为了进行对比,分别进行了几个相关的实验,其中,基本 HMM 模型与基本 ME 模型是未使用前缀、中缀、后缀角色的最基本分类器;而 HMM 模型与 ME 模型使用了这三种标记信息;混合模型是以 HMM 模型与 ME 模型作为序列标注器.实验结果如表 12 所示.

表 12 命名实体识别实验对比

识别模型	召回率/%	精确率/%	F 量度/%
最大匹配方法 BaseLine	73.54	68.99	71.19
基本 HMM 模型	79.96	79.20	79.58
基本 ME 模型	83.23	84.77	83.99
HMM 模型	85.20	83.68	84.43
ME 模型	84.62	87.95	86.25
混合模型	87.81	89.32	88.56
上界估计	91.67	93.59	92.62

注:ME 模型采用 BeamSearch 搜索算法,宽度为 5.

为进一步衡量混合模型的性能,我们估计 HMM 与 ME 模型融合的下界与上界.其中,下界可认为是最优性能的单序列标注器的性能(ME 模型);而估计上界的方法是综合 HMM 模型与 ME 模型结果,判别二者中至少有一个标注是正确的.

表 12 的实验结果表明:相比单一模型的命名实体识别系统,混合模型具有更好的标注效果.考察 HMM 模型与 ME 模型在标注实体时呈现的互补关系,图 7 显示了两个模型对正确实体的召回情况.

图 7 表明两个模型存在一定的互补优势,而这种互补恰是模型融合的必要条件.同时,它也表明存在 91.26%−78.55%=12.71%的调整空间,对这部分实体的判别需要从实体自身特征、上下文环境甚至远距离上下文环境入手.

4.4 词法分析中主要问题的分析与探讨

我们统计词性标注与分词出错情况的特征使用

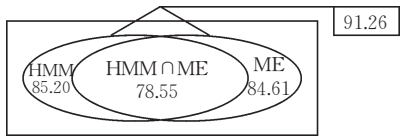

图 7 HMM 模型与 ME 模型互补关系情况/%

情况,如图 8.其中(a),(b)图为 CRF 词性标注时统计错误标注时的特征情况.

图 8(a)表明,出错时,似乎有很多特征应用于 CRF 标注过程,但(b)图指明在训练时真正相关的特征非常少,约 78%的情况下出现小于等于两个特征的情况.同样的数据稀疏问题也会出现在 ME 与 MEMM 模型中(笔者在文献[7]中指明,利用最大熵模型进行分词歧义消解时也呈现相同情况).上述表明,数据稀疏问题是影响词法分析系统性能提高的最主要因素.而数据稀疏的原因不能单纯归咎于训练样本稀少,还与我们所利用的特征是否有效、充分以及基于词的特征的表示方法是否适合有关.

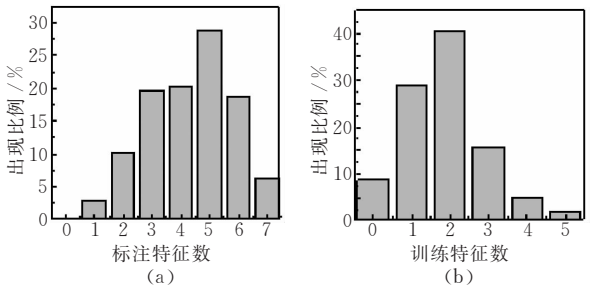

图 8 当词性标注出错时的特征统计情况

这引导我们可以从三方面思考进一步提高词法分析性能:(1)最直接的方法是提高训练语料规模和质量,然而 Zipf 定律指明该方法获得一定性能后,不易再大幅度提高性能了;(2)应该挖掘与选择更有效的特征用于描述词法知识,这也正是我们基于粗糙集等方法提取远距离特征的原因[10];(3)除了基于词的特征,还应该设法充分利用其他种类的特征,如颗粒度更大的句法特征等,从信息增益的角度看,这也是一种可尝试的方法.比如基于角色的命名实体标注方法也恰是有效地利用了粗颗粒度信息.除了上述三方面,语言模型的选择也是影响词法分析性能的重要原因之一.

5 结论与未来展望

本文以"问答式信息检索"项目为背景,全面阐述了中文词法分析的相关问题以及我们系统的构成.本文所做的工作如下:

(1)在词法分析的框架内,通过分解协作方式一体化地完成了分词、词性标注及命名实体识别的

加工过程，有效地利用了各模块带来的信息增益.

(2)详细地分析了词法分析所面临的主要困难，并且阐述了我们基于混合模型建立的词法分析系统，有针对性地克服了面临的主要问题.该系统具有较好的实用性.

(3)在模型研究上，基于双层混合模型构建命名实体识别系统以及引入粗糙集理论挖掘长距离特征与复杂特征的方法.

在今后的工作中，笔者拟在如下两个方向上继续探索：在理论方面，研究一体化词法分析的方法以及相应的机器学习方法；在实践方面，继续努力提高系统在分词、词性标注、名实体识别等环节的精确性，从而提高词法分析系统的整体性能.

参　考　文　献

[1] Zhang Hua-Ping, Liu Qun et al. Chinese lexical analysis using hierarchical hidden Markov model//Proceedings of the 2nd SIGHAN Workshop Affiliated with 4th ACL, Sapporo, Japan, 2003: 63-70

[2] Gao Jian-Feng, Li Mu et al. Chinese word segmentation: A pragmatic approach. MSR-TR-2004-123, 2004

[3] Xue N. Chinese word segmentation as character tagging. International Journal of Computational Linguistics and Chinese Language Processing, 2003, 8(1): 29-47

[4] Peng Fu-Chun, Feng Fang-Fang, McCallum Andrew. Chinese segmentation and new word detection using conditional random fields//Proceedings of the COLING, 2004: 562-568

[5] Jiang Wei, Zhao Jian et al. Chinese word segmentation based on mixing model//Proceedings of the 4th SIGHAN Workshop, Jeju Island, Korea, 2005: 180-182

[6] Zhao Jian, Wang Xiao-Long et al. Comparing features combination with features fusion in Chinese named entity recognition. Computer Application, 2005, 25(11): 2647-2649(in Chinese)
(赵　健，王晓龙等.中文名实体识别中的特征组合与特征融合的比较.计算机应用，2005，25(11)：2647-2649)

[7] Jiang Wei, Wang Xiao-Long et al. Applying rough sets in word segmentation disambiguation based on maximum entropy model. Journal of Harbin Institute of Technology(New Series), 2006, 13(1): 94-98

[8] Lafferty J, Freitag D, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data//Proceedings of the International Conference on Machine Learning, 2001: 282-289

[9] Sha F, Pereira F. Shallow parsing with conditional random fields//Proceedings of the Human Language Technology/North American Chapter of the Association for Computational Linguistics Annual Meeting, 2003: 213-220

[10] Jiang Wei, Wang Xiao-Long, Guan Yi et al. Applying rough sets to extract feature in POS tagging. Chinese High Technology Letters, 2006, 16(10): 996-1000(in Chinese)
(姜　维，王晓龙，关　毅等.基于粗糙集理论的词性标注模型.高技术通讯，2006，16(10)：996-1000)

[11] Jiang Wei, Wang Xiao-Long, Guan Yi. Improving sequence tagging using machine-learning techniques//Proceedings of the ICMLC2006, 2006

[12] Zhao Yan. Research on Chinese morpheme analysis based on statistic language model[Ph. D. dissertation]. Harbin: Harbin Institute of Technology, 2005(in Chinese)
(赵　岩. 基于统计语言模型的汉语词法分析研究[博士学位论文]. 哈尔滨：哈尔滨工业大学，2005)

JIANG Wei, born in 1978, Ph. D. candidate. His research interests include the theories and methods for machine learning, question answering, and information extraction.

WANG Xiao-Long, born in 1955, professor, Ph. D. supervisor. His research interests include artificial intelligence, machine learning, computational linguistics, and Chinese information processing.

GUAN Yi, born in 1970, associate professor. His research interests include question answering, web mining.

ZHAO Jian, born in 1975, Ph. D.. His research interests include conditional probabilistic model, named entity recognition.

Background

Chinese Lexical Analysis (CLA) includes word segmentation, Part-of-speech tagging and named entity recognition. CLA is a foundation and important task in most Chinese Natural Language Processing tasks. It affects not only the precision of successive processing, such as parsing, but also the performance of some applications, such as information extraction, question answer system etc. Although the lexical analysis is a basic task of NLP, it is still one of hot research fields. Lots of open evaluations have been held. International Chinese Word Segmentation Bakeoff (SIGHAN) greatly prompt word segmentation all over the world for it provides a platform to compare the performance of different Chinese word segmentation systems. And named entity recognition evaluation has been held in CoNLL-2002, CoNLL-2003 and ACE (Automatic Content Extraction). Though a lot of success has been achieved, more research work is needed in lexical analysis task. This paper presents a novel method for Chinese Lexical Analysis based on mixing language model, so as to take advantage of each model in solving the special subtasks. The system finished by the authors a key part of "the key project of National Natural Science Foundation of China" — "The Information Retrieval Based on Question Answer system". It can provide highly quality support for other natural language processing tasks.