一种面向移动终端的自然口语任务理解方法①

郭 群[1], 李剑锋[2], 陈小平[1], 胡国平[2]

[1](中国科学与技术大学 计算机科学与技术学院, 合肥 230027)
[2](安徽科大 讯飞信息科技股份有限公司研究院, 合肥 230088)

摘 要: 随着移动互联时代的到来和语音识别技术的日益成熟, 通过语音的交互方式来使用移动终端成为一种趋势. 如何理解用户自然状态下的口语输入, 传统的做法是手写上下文无关的文法规则, 但是文法规则的书写需耗费大量的人力和物力, 很难去维护和更新. 提出一种采用支持向量机和条件随机场串行结合的方法, 把口语任务理解分解为任务发现和信息抽取两个过程, 并最终将任务表达成语义向量的形式. 最终对“讯飞语点”语音助手用户返回的八个不同的任务种类的数据进行了测试, 在一比一的噪声中识别任务语义表达的准确率为 90.29%, 召回率为 88.87%.

关键词: 口语理解; 任务发现; 信息抽取; 支持向量机; 条件随机场

A Method to Understand Spontaneous Spoken Tasks for Mobile Terminals

GUO Qun[1], LI Jian-Feng[2], CHEN Xiao-Ping[1], HU Guo-Ping[2]

[1](School of Electronic Science and Technology, University of Science and Technology of China, Hefei 230027, China)
[2](Research Center, Anhui USTC iFLYTEK Co. Ltd, Hefei 230088, China)

Abstract: With the development of mobile Internet and automatic speech recognition (ASR), the mobile terminal through voice interaction has become a trend. The traditional method to understand user's spontaneous spoken language is to write context-free grammars(CFGs)manually. But it is laborious and expensive to construct a grammar with good coverage and optimized performance, and difficult to maintain and update. We proposed a new approach to spoken language understanding combining support vector machine(SVM)and conditional random fields(CRFs), which detect task and extract task semantic information from spontaneous speech input respectively. Tasks are represented as a vector of task name and semantic information. Eight different tasks from“iFLYTEK yudian”voice mobile assistant are tested, and the precision and recall of semantic representation of query are 90.29% and 88.87% respectively.

Key words: spoken language understanding; task detection; information extraction; support vector machine; conditional random fields

随着移动互联时代的到来, 移动终端的使用呈现了爆发式的增长, 用户对移动终端的使用已经不仅仅局限于通信. 个人电脑上的应用不断的出现在移动终端, 人们可以随时随地使用移动终端发送邮件、查询火车票、查询股票情况. 受限于传统的交互方式, 移动终端的输入设备很难满足不断膨胀、复杂多变的应用需求. 近些年语音识别技术日益成熟, 通过语音的交互来操作移动终端成为一种可能和趋势. 让移动终端成为我们身边的助手, 用户只需要将想法以自然状态下口语的方式告诉机器, 让机器去完成并给出反馈. 怎样理解用户自然状态下的口语成为了一个有待解决的非常关键的问题.

与书面语的工整表达和网络搜索用语的简练不同, 口语噪声更大, 表达复杂多变, 而且没有严格的句法格式. 相同的意思, 每个人的表达方式不同. 相同的人在不同的环境下、不同的上下文背景下表达方式也不一样.

① 收稿时间:2013-01-14;收到修改稿时间:2013-02-25

怎样理解用户的意图是一个非常棘手的问题. 传统的作法是通过手写文法规则的方式[2-4], 但是文法的标注需耗费大量的人力和物力, 并且文法之间的容易产生冲突, 很难去维护和更新. 1990 年由 DARPA 发起的面向航空信息的口语理解评价系统 ATIS[5](Air Travel Information System evaluations)出现之后, 口语理解得到了快速的发展, 众多基于统计模型的航空信息领域的口语理解技术和系统相续出现[6-10]. 其中文献[9]提出基于上下文无关规则和产生式模型相结合的方式, 在一定程度上缓解了统计学习方法中的数据稀疏问题, 减少了工作量和错误率; 文献[10] 基于上下文无关规则和判别式模型相结合的方式, 实验结果显示错误率比文献[9]的 HMM/CFG 模型降低了 20%. 相比基于规则的方法, 统计模型方法从大量的训练语料中自动学习, 训练数据的标注相对简单, 不需要技术背景. 目前这些研究大都是面向一个非常限定的领域, 预先定义该领域的语义框结构(semantic frame), 采用类似于语音识别的方式, 标注词语序列的对应语义序列, 匹配上正确的框架得到语义表达.

中文口语的研究起步较晚, 并且中文口语比英文口语更加的复杂, 表达方式更加丰富, 暂时还没有比较成熟的技术出现. 在面对移动终端的中文口语理解问题, 相比于英文研究文献[6-10], 领域更加开放, 使用传统方法需要巨大的工作量. 本文提出采用支持向量机模型和条件随机场模型串行结合的方式, 支持向量机模型实现对用户任务的发现, 条件随机场模型来抽取完成任务的必要语义信息.

1 背景知识

支持向量机[11,12](Support Vector Machine)和条件随机场[1](conditional random field)是统计自然语言处理中最常用监督学习方法. 支持向量机, 简称 SVM, 从线性最优分类面发展而来的, 广泛应用于统计分类和回归分析中. 条件随机场, 简称 CRF, 最常用的形式是线性链条件随机场, 它常用来实现从输入序列到隐含序列的标注. 相比于隐马尔科夫模型[13]和最大熵马尔科夫模型[14], 条件随机场模型的提出很好的解决了隐马尔科夫模型输出独立性假设和最大熵马尔科夫模型的标注偏见[1]问题, 可以充分考虑上下文特征, 且对训练语料中未出现的情况也能很好的预测. 线性链条件随机场模型常被用来实现词性标注、信息抽取和命名实体的识别[15].

1.1 支持向量机分类模型

支持向量机的基本思想是做出正确划分数据集并且保持几何间隔最大的分离超平面, 又称最大间隔分类器, 见图 1.

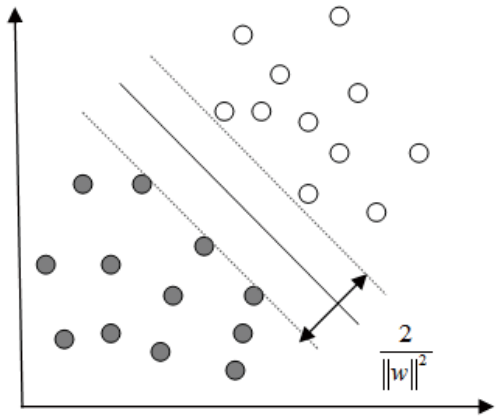

图 1 线性可分支持向量机

训练数据集 $\{(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\}$, 其中 $x_i \in R^n, y_i \in \{-1,+1\}, i=1,2,..,N$. 解决最大间隔超平面的问题为构造并求解约束最优化问题:

$$Minimize \frac{1}{2}\|w\|^2$$
$$s.t.\ y_i(w^T x_i + b) \geq 1 \quad i=1,2,...,N \tag{1}$$

对于非线性的问题, 采用惩罚参数 C, 确保几何间隔最大化的同时, 最小化经验误差. 对于非线性的分类问题, 采用核函数, 将非线性问题转化为线性的问题来求解, 本文采用的核函数是高斯径向基函数:

$$K(x,z) = \exp\left(-gram*\|u-v\|^2\right) \tag{2}$$

1.2 条件随机场模型

条件随机场最常用的形式是线性链条件随机场(见图 2), 即当观察序列和标记序列为线性链表示的一一对应的变量序列, 且给定随机变量序列的条件下, 随机变量序列的条件概率分布满足马尔科夫假设:

$$P(y_i \mid x, y_1, y_2, ..., y_n) = P(y_i \mid x, y_{i-1}, y_{i+1})$$

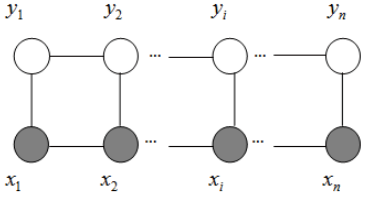

图 2 线性链条件随机场

线性链条件随机场的参数化形式为:

$$P(y \mid x,\lambda)=\frac{1}{Z(x)}\exp\left(\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1},y_i,x,i)\right) \quad (3)$$

$$Z(x)=\sum_{y}\exp\left(\sum_{i=1}^{n}\sum_{j}\lambda_j f_j(y_{i-1},y_i,x,i)\right) \quad (4)$$

其中 x 为观察变量 X 的实例，x_i 为 x 的第 i 个变量. y 为状态变量 Y 的实例，y_i 为 y 的第 i 个变量. $f_j(y_{i-1},y_i,x,i)$ 为状态特征函数 $s_l(y_i,x,i)$ 和转移特征函数 $t_k(y_{i-1},y_i,x,i)$ 的统一化形式，λ_j 为每个特征函数的权值. $Z(x)$ 为归一化因子.

和隐马尔科夫模型一样，条件随机场的序列预测，采用的是维特比算法，对于一个输入序列，标注对应最可能的标注序列 $\bar{y}$:

$$\bar{y}=\arg\max_{y} P_\lambda(y \mid x) \quad (5)$$

2　基于SVM和CRF的口语理解

口语理解的目的是将用户的语音输入转化成计算机可处理的表达. 结构如图 3 所示，用户的口语输入经过语音识别(ASR)转化成文字，文字经过理解算法的处理转化成机器可处理的语义表达.

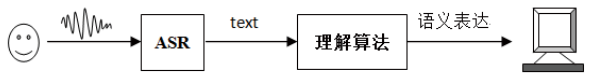

图 3　口语理解过程

本节余下部分首先将介绍本文对任务语义表达的定义，接下来具体描述本文提出的理解算法，最后介绍本文涉及的支持向量机模型和条件随机场模型特征的定义.

2.1 任务语义表达

人机交互中，机器要完成人类指定的任务，不仅要准确理解任务的类别，还要知道任务的一些例如时间、地点等的信息. 如“打电话”任务，机器需要知道被呼叫人的姓名才能完成任务. 不同的任务需要不同的语义信息，我们采用手工定义的方式对八个任务定义了语义信息列表，具体见表 1. 表 1 中必须要素定义为机器完成此任务必须的要素，但是一些任务中，某些要素设置用户当前所处状态为默认值，在用户输入中没有提供的情况下取默认值. 可选要素是使任务更加精确的描述，可以使机器更加精确的完成任务，但不是必须的. 为了更好的表达任务，本文将任务表达为“<任务类别，语义信息，…>”语义向量形式.

表 1　任务语义信息的定义

任务类别	语义信息	
	必须要素	可选要素
打电话	人名	
天气查询	时间，地点	
飞机票查询	时间，出发地，目的地	
火车票查询	时间，出发地，目的地	
打开手机应用	手机应用名称	
查找附近饮食	地点、店类别	店名、菜名
音乐播放	歌手、歌名	流派
导航	目的地	

2.2 理解算法

本文采用支持向量机和条件随机场串行结合的方法，把口语任务理解分解为任务发现和语义信息抽取两个过程.

2.2.1 任务发现模块

在面向任务的中文口语理解中,用户的自然口语输入分为任务(*T*)和噪声(*N*). 在领域内任务类别必定是一个有限的集合，则用户的口语输入必属于集合 $\{T_1,T_2,...,T_m,N\}$，其中 m 为任务类别总数. 本文采用支持向量机分类的方法将各个任务从噪声中区分开，实现任务的发现. 从标注好类别的语料训练 m+1 个类的 SVM 分类器，对用户的输入进行识别.

2.2.2 语义信息抽取模块

完成一个任务，在知道任务类别的情况下，还需要一些语义信息，任务的语义信息定义见表 1. 由标注好语义信息序列的任务语料，采用条件随机场训练每种任务的语义信息抽取模型. 对发现的任务调用对应任务的信息抽取模型，将经过分词的任务从关键词序列转化为语义序列，抽取对应的语义信息，并最终将任务表达为任务名和语义信息的向量形式.

系统结构如图 4 所示，输入为用户自然情况下口语表达经过语音识别转化成的文本，经过分词得到问句关键词向量表达. 任务发现模块发现任务 *I*，根据任务类别 *I*，调用不同任务类别的信息抽取模型，抽取该任务的语义信息，得到任务的语义表达. 如用户输入“国购广场附近的农家乐饭店”，经过分词得到问句向量 *X*=(国购广场，附近，的，农家乐，饭店). 经过任务

发现模块发现任务为“查找附近饮食”. 调用“查找附近饮食”的信息抽取模型, 得到标注结果 y={地点, 其他, 其他, 店名, 店种类}, 抽取得到任务的语义表达向量: (任务=“查找附近饮食”、地点=“国购广场”、店名=“农家乐”、菜名=“”、种类=“饭店”).

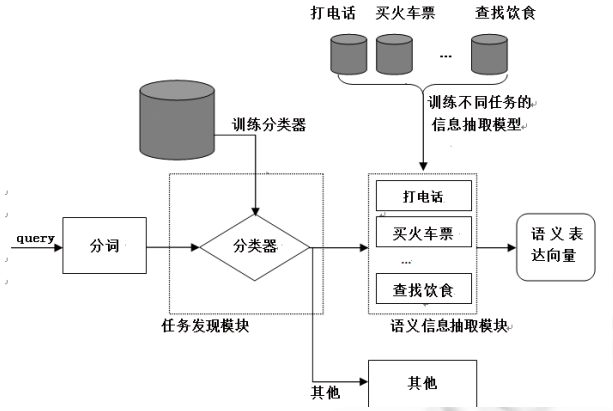

图 4 系统结构

2.3 特征的定义

2.3.1 支持向量机模型特征

关键词特征, 定义为用户输入中所含的词语. 除了关键词特征, 本文还使用了用户输入的语义特征, 定义为用户的输入中是否含有某种语义, 例如时间、地点等.

2.3.2 条件随机场特征模板

为了反映语言的内在规律, 充分考虑上下文的信息, 本文设定的滑动窗口大小为 2, 既只考虑当前词的前后两个词作为特征. 本文考虑了一元特征、二元特征和转移特征, 定义了相应的特征模板来筛选这些特征. 特征模板定义如表 2 所示, 筛选出来的特征都将定义成二值函数的形式.

表 2 条件随机场特征模板的定义

特征模板	模板意义
$x_{i-2}=w$	当前词的前第二个词
$x_{i-1}=w$	当前词的前一个词
$x_i=w$	当前词
$x_{i+1}=w$	当前词的下一个词
$x_{i+2}=w$	当前词的下下一个词
$pos_{i-2}=t$	当前词的前第二个词的词性
$pos_{i-1}=t$	当前词的前一个词的词性
$pos_i=t$	当前词的词性
$pos_{i+1}=t$	当前词的下一个词的词性
$pos_{i+2}=t$	当前词的下下一个词的词性
$x_{i-1}=w_1 \& x_i=w_2$	前一个词与当前词的组合
$x_i=w_1 \& x_{i+1}=w_2$	当前词与下一个词的组合
$pos_{i-1}=t_1 \& pos_i=t_2$	前一个词的词性与当前词的词性的组合
$pos_i=t_1 \& pos_{i+1}=t_2$	当前词的词性与下一个词的词性的组合
$y_{i-1}=s_1 \& y_i=s_2$	前一状态到当前状态的转换

3 实验

3.1 实验描述

3.1.1 实验环境描述

实验是基于 windows 环境, 编译环境为 VS2008.

3.1.2 实验语料描述

为了能有效反应本文方法的实用性, 本文所采用的实验数据全部来自“讯飞语点”语音助手用户使用语音数据经过语音识别转换成的文本. 其中本文选择的最常用的 8 个任务为经过人工标注的公司内部数据集, 语料总数为 22982 个. 为了训练和测试系统任务发现模块, 本实验加入了与任务语料等比例的噪声数据, 总数为 23000 个. 在实验中, 将整个语料分为训练集和测试集, 训练集和测试集的比列为 2:1, 具体实验语料见表 3.

表 3 实验语料描述

任务类别	问句总数	训练集	测试集
打电话	4001	2640	1361
天气查询	4231	2792	1439
飞机票查询	646	426	220
火车票查询	925	610	315
打开手机应用	4620	3049	1571
查找附近饮食	921	607	314
播放音乐	5346	3658	1688
导航	2292	1512	780
总数	22982	15294	7688
噪声	23000	15332	7666

3.1.3 实验评价标准

为了评价实验结果, 本文采用准确率(*P*)、召回率(*R*)和综合指标 *F* 值(*F*)对实验结果进行评价. 任务发现中准确率定义为(系统正确发现任务数/系统发现任务总数)*100; 召回率定义为(系统正确发现任务数/语料中任务总数)*100; F 值定义为(2*准确率*召回率)/(准确率+召回率). 语义信息抽取和系统评价标准和此类似.

3.2 实验及分析

3.2.1 任务发现模块

训练任务和噪声共九个类的支持向量机分类器，训练集大小为 30626，测试集大小为 15354. 为了得到分类器的最佳参数(惩罚因子 C，核参数 g)，使分类器能够精确地预测未知数据，本文采用 5 重交叉验证的方式，即将训练集分成五份，轮流将其中四份作为训练集，一份作为测试集，五次测试结果的平均值作为性能指标. 实验针对两个特征集合，交叉验证结果见图 5，任务发现实验结果见表 4.

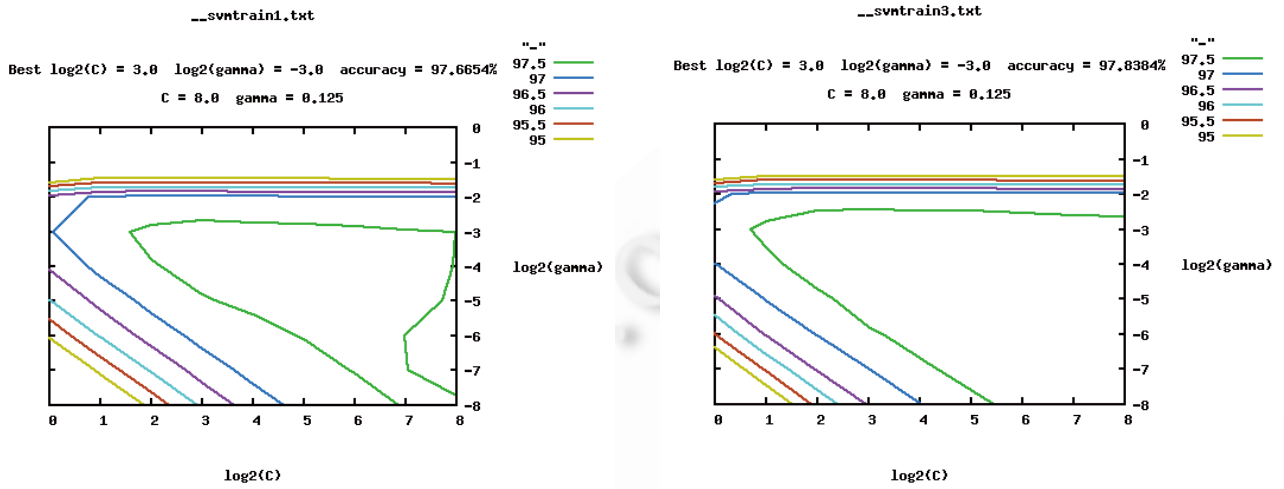

图 5　左右图分别为特征集合 1、2 交叉验证结果

表 4　任务发现实验结果

编号	特征选择	测试集	任务总数	发现任务总数	正确个数	准确率	召回率	F 值
1	关键词特征	15354	7688	7576	7421	97.95%	96.53%	97.23%
2	关键词特征+语义特征	15354	7688	7567	7421	98.07%	96.53%	97.29%

从表 4 可以看出，当只选择关键词作为特征时，实验已经达到非常好的效果，准确率为 97.9%，召回率达到 96.5%. 特征集合为关键词特征和语义特征时，准确率有了进一步的提高，但不明显.

3.2.2 语义信息提取模块

训练每种任务的条件随机场的信息抽取模型. 八种任务 7688 个输入中共有语义信息 8357 个. 实验结果见表 5，测试得到 8258 个语义信息，正确 7747，平均正确率为 93.81%，平均召回率为 92.70%. 从语义信息总数可以看出并不是每一个任务输入都有完整的所需语义信息. 从表 5 可以看出，对于像打开应用和导航任务，所需语义信息复杂，长度较长，我们的训练集过小数据过于稀疏，导致识别准确率和召回率较低.

3.2.3 系统实验结果

系统实验结果如表 6 所示，系统共识别任务总数 7567，其中正确表达任务向量的数量为 6832. 由于基于规则中文口语理解的文法规则的标注需要大量的人力和物力，在实验室的条件下很难实现，我们用“讯飞语点”对测试集进行了测试. 从实验结果得出，不需要大量人力物力去标注文法规则，只需要学习的方式系统的准确率和召回率分别达到了 90.28%和 88.86%，达到了和企业上线系统相似的系统表现 1.

表 5　语义信息抽取实验结果

任务类别	总数	结果个数	正确个数	正确率	召回率
打电话	1342	1347	1292	95.92%	96.27%
天气查询	1858	1852	1830	98.81%	98.49%
飞机票查询	460	461	447	96.96%	97.17%
火车票查询	599	593	593	100.00%	99.00%
打开手机应用	1571	1528	1311	85.80%	83.45%
查找附近饮食	341	337	306	90.80%	89.74%
播放音乐	1626	1563	1464	93.67%	90.04%
导航	560	577	504	87.35%	90.00%
	8357	**8258**	**7747**	**93.81%**	**92.70%**

表 6　系统测试结果

	任务总数	识别任务数	正确表达任务向量数	准确率(%)	召回率(%)	F1 值(%)
结果	7688	7567	6832	90.29%	88.87%	89.57%

4 结论与工作展望

中文口语的理解是一个非常困难有挑战的问题. 对于面向相对开放领域的中文口语理解问题, 本文提出了采用支持向量机和条件随机场串行结合的方式, 支持向量机分类来实现用户任务的发现, 条件随机场序列标注的方式来实现对用户口语输入任务附加语义信息的提取. 最后实验完全是基于用户真实使用数据可以充分的反映该方法的可行性. 相比于基于规则的方法, 本方法在不需要大量人力和物力去标注规则, 解决冲突的情况下, 系统的准确率达到了 90.3%, 召回率达到 88.8%, 达到了和企业系统的相似的系统表现. 实验结果表明该方法有效可行, 能够很好的适应中文的口语表达方式的复杂多变, 句式不工整等问题.

从实验中, 我们发现两个问题, 一是对于像导航和打开手机应用任务, 语义信息复杂冗长, 语义信息抽取的准确率不高, 下一步我们将尝试使用半马尔科夫条件随机场[16](Semi-Markov Conditional Random Fields), 在实现对观察序列分段的同时标注各个段落的语义, 提高信息抽取的准确率. 二是对于用户口语输入中没有必须语义信息的问题, 如用户输入“打电话”, 没有人名或者号码, 无法完成任务, 下一步我们会采用对话的形式获取完成任务的必要信息.

参考文献

1 Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proc. of ICML, 2001: 282–289.

2 Ward W. Recent improvements in the CMU spoken language understanding system. Proc. Human Language Technology Workshop, Plainsboro, NJ, 1994: 213–216.

3 Seneff S. TINA: A natural language system for spoken language applications. Comput. Linguistics,1992,18(1):61–86.

4 Dowding J, Gawron JM, Appelt D, Bear J, Cherny L, Moore R, Moran D. Gemini: A natural language system for spoken-language understanding. Proc. 31st Ann. Meeting Association for Computational Linguistics, Columbus, Ohio, 1993:54–61.

5 Price P. Evaluation of spoken language system: The ATIS domain. Proc. DARPA Speech and Natural Language Workshop, Hidden Valley, PA, 1990: 91–95.

6 Miller S, Bobrow R, Ingria R, Schwartz R. Hidden understanding models of natural language. Proc.31st Ann. Meeting Association for Computational Linguistics, New Mexico State University, 1994: 25–32.

7 Pietra SD, Epstein M, Roukos S, Ward T. Fertility models for statistical natural language understanding. Proc. 35th Ann. Meeting Association for Computational Linguistics, Madrid, Spain, 1997: 168–173.

8 He Y, Young S. Semantic processing using the hidden vector state model. Journal of Computer Speech & Language. Elsevier, 2005,19(1):85–106.

9 Wang YY, Acero A. Combination of CFG and N-gram modeling in semantic grammar learning. Proc. Eurospeech 2003, Geneva, Switzerland, 2003: 2809–2812.

10 Wang Y, Acero A, Mahajan M, Lee J. Combining statistical and knowledge-based spoken language understanding in conditional models. Proc. of COLING/ACL, July 2006.

11 Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers.In Haussler D. ed, 5th Annual ACM Workshop on COLT, Pittsburgh, PA, 1992. ACM Press: 144–152.

12 Cortes C, Vapnik V. Support-Vector Networks. Machine Learning,20,1995.

13 Leek TR. Information extraction using hidden Markov models [Master's thesis]. U.C.San Diego, 1997.

14 McCallum A, Freitag D, Pereira F. Maximum entropy Markov models for information extraction and segmentation. Proc. ICML 2000. Stanford,California. 591−598.

15 McCallum A, Li W. Early results for Named Entity Recognition with Conditional Random Fields, Feature Induction and WebEnhanced Lexicons. Proc. of CoNLL-2003.

16 Sarawagi S, Cohen WW. Semi-Markov conditional random fields for information extraction. Saul LK, Weiss Y, Bottou L, eds. Advances in Neural Information Processing Systems 17, pages1185−1192. MIT Press, Cambridge, MA, 2005.

17 Wang YY, Deng L, Acero A. Spoken language understanding. IEEE Signal Processing Mag.,Sept.2005,22(5):16–31.

18 Chang CC, Lin CJ. LIBSVM:a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

19 http://crfpp.sourceforge.net/.