基于条件随机场的自然口语语义理解方法

李成华[1],张世娟[1,*],刘 磊[2],江小平[1]
(1 中南民族大学 电子信息工程学院,武汉 430074;2 武汉民大信息科技有限公司 研发部,武汉 430074)

摘 要 采用条件随机场技术将面向智能手机用户的自然口语语义理解分为操作任务分类和语义组块提取两个主要步骤,收集并分析了口语语料库的特征,根据归纳出的任务种类和语义组块特征规律设计了任务分类标记集和语义组块标记集;通过基于规则的组块分析得到了中间语义表示格式,从而实现了对用户口语语义理解的目的.实验结果表明:任务分类准确率及语义组块提取平均正确率分别达到 98.85% 和 94.53%,系统综合性能测试的准确率达到 91.86%.

关键词 人工智能;自然语言处理;口语理解;条件随机场;中间语义表示格式(IF)
中图分类号 TP39 **文献标识码** A **文章编号** 1672-4321(2017)02-0060-06

Approach to Understand Chinese Oral Task for Mobile Terminals Based on Conditional Random Fields

Li Chenghua[1], *Zhang Shijuan*[1,*], *Liu Lei*[2], *Jiang Xiaoping*[1]
(1 College of Electronics and Information Engineering, South-Central University for Nationalities, Wuhan 430074, China;
2 Information Technology Limited Company of South-Central University for Nationalities, Wuhan 430074, China)

Abstract Conditional random fields technology is applied to this paper and semantic understanding of the natural spoken language of smart phone users is divided into two processes: classification of operating instructions and extraction of semantic chunking information. Characteristics of the spoken language corpus are collected and analyzed. Tag sets of task classification and semantic chunking are designed according to the inductive types of tasks and the rule of the semantic chunking characteristics. The middle semantic representation format are obtained through the chunking analysis based on rules, so as to realize the goal of the semantic understanding spoken language of the users. In the experiment, the accuracy rate of tasks classification reaches 98.85%, and the average accuracy rate of semantic chunking extract reaches 94.53%. In the end, the accuracy rate of system's comprehensive performance test reaches 91.86%.

Keywords artificial intelligence; natural language processing; spoken language understanding; conditional random fields; middle semantic representation format(Interchange Format)

相比键盘、按键和触摸屏等交互方式,自然语言对话是最自然高效的人机交互方式.近年来苹果公司在 Iphone 和 Ipad 上推出的 Siri 智能语音助手,以其强大的自然语言理解能力给用户带来了全新的体验感.它集成了语音识别、语义理解和语音合成等技术,成功地将人工智能技术带进普通大众的生活,同时,也再度引起学术界和工业界对自然语言处理技术研究的极大关注.

口语理解的任务是理解用户的意图并抽取用户输入语句所包含的关键信息[1].面向智能手机终端对用户口语任务语义理解,最终目的是将用户的语音输入转化成智能手机可处理的表达形式.口语语

收稿日期 2016-11-20 * 通讯作者 张世娟,研究方向:语言信息处理,E-mail:1303326202@qq.com
作者简介 李成华(1972-),男,副教授,博士,研究方向:云计算、信息安全,E-mail:103776901@qq.com
基金项目 中央高校基本科研业务费专项资金资助项目(CZW15043,CZQ14001);文物保护装备产业化及应用示范项目(2015-427)

句往往不符合语法规范,存在句子成分重复、省略和颠倒等现象[2]. 口语理解与传统自然语言理解技术密切相关,又有其自身的特点和难点[3]. 陈俊燕等[4]采用概念级和句子级两级理解的方式,在第一级按照概念规则进行概念捆绑,得到句中所有可能的概念候选,生成概念图. 第二级根据句子级规则集使用可跳过非关键成份的基于树扩展的稳健句法分析算法搜索最优的整句理解结果. 吴尉林等[1]提出了基于两阶段分类的口语理解方法: 第 1 阶段为主题分类,用来识别用户输入语句的主题; 第 2 阶段为主题相关的语义槽分类,根据识别的主题抽取相应的语义槽/值对. 本文将口语理解分为操作任务分类和语义组块提取两大步,操作任务分类的目的是获得用户口语的任务类型,根据识别的任务类型进行语义组块提取以获得执行该任务的相应的的参数信息.

郭群等[5]采用支持向量机(SVM)和条件随机场(CRF)串行结合的方法实现了口语理解的任务发现和信息抽取,并将任务最终表达成语义向量的形式. 但由于 SVM 本质上属于两类分类算法,同时还存在核函数及其核参数的选择以及松弛系数和惩罚因子等参数确定上的困难[6],导致文献[5]在任务发现模块中构造了 9 个基于 SVM 的分类器,采用 5 重交叉验证的方法以获得惩罚因子 C 和核参数 g 的最佳取值. 此外,在文献[5]中,仅考虑了用户输入的关键词特征和语义特征,而没有考虑到上下文特征. 朱敏[7]就音乐类命名实体进行识别分类时对 CRF 和 SVM 进行了对比实验,结果表明采用 CRF 技术识别的准确率要高出将近 12%. 相比文献[5],本文在任务分类和语义组块提取两个步骤中均采用了 CRF 技术,并最终得到的口语语义理解结果采用中间语义表示格式(IF)[2]进行表示.

1　采用的技术

1.1　条件随机场

条件随机场模型是在 2001 年被提出的一种典型的判别式模型,它在观测序列的基础上对目标序列进行建模,重点解决序列化标注的问题[8]. 条件随机场模型既具有判别式模型的优点,又具有产生式模型考虑到上下文标记间的转移概率、以序列化形式进行全局参数优化和解码的特点,解决了其他判别式模型(如最大熵模型,隐马尔科夫模型)难以避免的标记偏置问题[9]. 它常被用于完成句法分析[10]、分词标注[11]、命名实体识别[12]等自然语言处理(NLP)任务. 刘泽文等[11]使用 CRF 进行中文短文本分词实验的 F 值达到 95% 以上. 张朝胜等[12]在英文产品命名实体识别实验中采用 CRF 技术 F 值达到 93.0%. CRF 技术详细资料可以参考文献[13]的描述.

1.2　中间语义表示格式 IF

用户口语解析的结果是 IF 表示式,它是一种基于中间语义关系的人造体系,由国际语音翻译联盟(C-STAR)制定[2]. IF 是一种中间格式,可以供不同的应用目标系统使用(如语言翻译),智能手机程序开发员利用该格式能快速生成目标程序代码. 其理论基础是对话行为理论,基本观点是: 语言不仅仅用来陈述事实,还带有说话者的意图[14]. IF 具体代表用户的一句话,通常由四部分组成: 说话者(Speaker)、语句意图(Speach Act)、概念(Concept)、具体参数(Arguments). 一个 IF 表达式由一个说话者标志和至少一个的语句意图以及数目可选的概念和参数组成,语句意图,概念和参数之间可以如下进行组合 Speaker: Speech Act + Concept* (Argument *),其中* 表示它所限定的左边成分可重复出现多次.

"说话者"用于说话人的身份标注,c 代表顾客(client),a 代表代理(agent). 本文只用到 c 标志.

"语句意图"表示用户的语句意图或语言行为. 本文将其划分为三种"动作请求、询问请求、未知请求",其中"动作请求"是用户让机器执行相关的应用程序,"询问请求"是用户需要机器返回有关信息的查询结果,"未知请求"表示机器不能理解的请求.

"概念"是用户请求的类别,也是语句的主题概念. 它在本文表示口语任务的语义类别,如设定闹钟、发短信等任务.

"具体参数"即语句的具体内容描述,如人名、时间、地点等组块内容,它在本文表示完成某条任务所需的具体参数信息,表达形式由 Argument 和对应的 Value 构成,比如: ContactName = 张三.

1.3　语义组块分析

目前,浅层语义分析一般采用角色标注的方法进行处理,由于语料库的规模限制,如果沿用常用的语义角色提取方法,将会显现出数据稀疏问题,进而影响分析效果. 本文将采用语义组块分析方法. 对语义组块的概念,学者们从不同角度给出了不同的定义. 本文采用文献[14]中的解释: 语义组块是指口

语句子中不依赖于其他词汇而能表示某种特定语义的最小部分. 界定语义组块遵从两个基本原则: 1)能够承担某种确定语义的普通短语; 2)按照口语表达习惯,以比较固定的形式出现、具有明确含义的组合语.

CRF 技术在语义组块分析中也有成功应用的报道. 孙广路等[15]的基于条件随机域模型的中文组块分析的算法性能优于基于最大熵马尔科夫模型的组块分析方法,将语义类特征加入分析模型后取得了 F 值为 92.77% 的实验结果. 李新德等[16]通过构造层叠条件随机场(CCRF),先后提取名词短语组块及语义组块,最后组成完整路径信息,其平均准确率和召回率分别达到 81.98% 和 82.32%.

2 采用方法描述

本文两次采用 CRF 技术,第一次用于识别用户口语所要表达的操作任务类型,第二次用于提取语义组块内容,最终输出口语语义的 IF 表示式. 图 1 是本文实现方法框图,总体分为训练阶段和解析阶段,训练阶段包括: 预处理模块和训练模型模块,解析阶段包括: 任务分类模块、语义组块提取模块和语义生成模块.

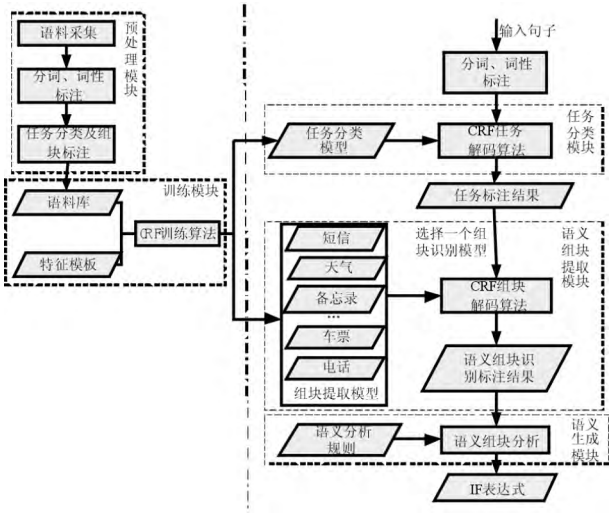

图 1 方法框图

Fig. 1 Block diagram

预处理模块: 采集用户口语语音语料并转化为文本; 对文本进行分词、命名实体识别、词性标注; 根据归纳的任务种类和语义组块特征规律设计任务分类标记集和语义组块标记集,对语料手工标注任务类型和语义组块. 标注语料是建立统计模型的第一步. 预处理后,语句由词序列变成语义类序列,得到语料库.

训练模块: 采用自定义特征模板,采用 CRF 技术训练语料库得到任务分类模型; 将语料按任务类型进行分类,再次用 CRF 技术分别训练每个任务的语料,得到各任务对应的语义组块提取模型.

任务分类模块: 从操作任务分类的角度,将用户的自然口语输入分为任务 T 和噪声 U,用户的口语输入的任务类型必对应有限集合{$T_1, T_2, \cdots, T_m, U$}的一种,其中 m 为任务类别总数. 对进行测试的语句使用训练得到的任务分类模型对其执行 CRF 解码算法,得到任务分类的标注结果.

语义组块提取模块: 根据任务分类标注结果选择该任务对应的语义组块识别模型,再次对词性标注过的用户输入语句执行 CRF 解码算法,得到语义组块识别标注结果.

语义生成模块: 运用语义组块分析规则分析上一步的语义组块识别标注结果,得到该输入语句的中间语义表示格式 IF 表达式.

3 关键过程描述

3.1 语料采集、分析与标注

语料采集: 文献[5]采用的科大讯飞公司的口语语料并没有公开. 本文自行收集口语语料,方法是通过自建网站在线问卷调查的形式,调查对象是武汉两所高校的在校大学生. 为获得被调查人员真实的用语习惯,未设置任何限制条件. 在剔除错字、重复以及其他错误后,共获得了 4109 个有效句子.

分析: 经过对收集语料的整理和分析,归纳出了 10 个任务框架概念,见表 1 第 2 列,如果经过任务分类不能把它归于这 10 个任务的其中一个的话,那么就把它归到“噪声”类别中. 每种任务框架都包含必需要素和可选要素. 其中,必需要素是智能手机完成该任务操作所必须具备的参数,在某些任务框架中必需要素如果没有在口语中显示表达出来,就设置相应的默认值,如在“天气”任务框架中,地点信息默认为当地,时间默认为今天; “导航”任务框架中,起点位置默认为当前位置. 可选要素是对具体指令任务的更加精确的描述要素,比如“定一个早上 6 点的闹钟,震动模式”,其中“震动”是可选要素.

标注符号设计: 根据上面的分析,并结合中文语义组块分析标注方法以及本文所采用 CRF + + 工具标注格式要求,设计了如表 1 所示的标注符号.

表 1　标注符号

Tab. 1　Tagging symbols

语句意图	任务框架		语义组块必需要素		语义组块可选要素		其他成分
	类型	标注符号	内容	对应的标注符号	内容	标注符号	
动作请求	打电话	P	人名或号码	PN			N
	发短信	S	人名,内容	SP,SC			
	设置管理	ST	动作(关/开),名称	STM,STN	状态	SMS	
	备忘录	M	时间,内容	MT,MC			
	闹钟	AC	动作,时间	ACM,ACT	模式	AMM	
	应用管理	AP	动作,应用名称	APC,APN			
	音乐播放	MS	歌手,歌名	MSS,MSN	流派	MSG	
询问请求	天气	W	时间,地点	WT,WL			
	导航	NA	出行方式,目的地,出发地	NAY,NAD,NAO			
	车票	T	乘车类型,时间,出发地,目的地	TY,TT,TD,TO			
未知请求	其他	U	搜索引擎检索内容	UC			

表 2 给出了两个标注样例,两个语句之间用空行隔开. 其中,第 1、2 列是 BosonNLP 分词标注工具对口语文本进行标注的结果,第 3 列是任务框架标注,第 4 列是语义组块标注. 在任务分类模型训练中使用了前 3 列特征,在语义组块提取模型训练中使用了第 1、2 和第 4 列特征.

3.2　特征模板设计

条件随机域模型的性能在很大程度上依赖于特征模板的选取. 在 CRF 模型中,上下文是以当前单词为中心的一个“观察窗口”,窗口大小会在很大程度上影响最终的识别效果. 结合本文语料库特征,并经过反复试验对比,确定了上下文单词特征和词性特征的窗口长度. 本文考虑了一元、二元特征和转移特征,所设计的特征模板及意义描述见表 3 描述. 其中,在任务分类训练时采用表 3 中除 U10 和 U14 之外的特征,训练得到任务分类模型; 在语义组块提取训练时采用表 3 中除 U00 和 U04 之外的特征,对每种任务类型的语料分别进行训练,得到对应的 10 个语义组块识别模型.

表 2　标注样例

Tab. 2　Sample of tagging

词语	词性标注	任务框架标注	组块标注
早上	t	AC	ACT
6 点半	nz	AC	ACT
叫	vi	AC	N
我	rr	AC	N
起床	vi	AC	N
震动	vn	AC	AMM
今天	t	W	WT
白天	t	W	WT
三亚	ns	W	WL
的	udel	W	N
气温	n	W	N
多	a	W	N
高	a	W	N

表 3　特征模板意义

Tab. 3　The meaning of feature template

特征模板	意义	举例说明
U00:%x [-2,0]	前第二个词形	今天
U01:%x [-1,0]	前第一个词形	白天
U02:%x [0,0]	当前词形	三亚
U03:%x [1,0]	后第一个词形	的
U04:%x [2,0]	后第二个词形	气温
U05:%x [-1,0] /%x [0,0]	前第一和当前词形	白天/三亚
U06:%x [0,0] /%x [1,0]	当前和后第一词形	三亚/的
U10:%x [-2,1]	前第二个词性	t
U11:%x [-1,1]	前第一个词性	t
U12:%x [0,1]	当前词性	ns
U13:%x [1,1]	后第一个词性	udel
U14:%x [2,1]	后第二个词性	n
U15:%x [-1,1] /%x [0,1]	前第一和当前词性	t/ns
U16:%x [0,1] /%x [1,1]	当前和后第一词性	ns/udel
U17:%x [-1,1] /%x [0,1] /%x [1,1]	前第一、当前和后第一词性	t/ns/udel
U20:%x [0,0] /%x [0,1]	当前词形和和当前词性	三亚/ns
U21:%x [0,0] /%x [1,1] /%x [2,1]	当前词形、后第一和后第二词性	三亚/udel/n
U22:%x [-1,0] /%x [0,0] /%x [1,0] /%x [0,1]	前第一、当前、后第一词形和当前词性	白天/三亚/的/ns

3.3 IF的生成

语义组块提取后得到标注符号序列,我们采用基于规则的方法进行标注序列到IF表达式的映射转换.针对表1中的11种任务框架,我们对应定义了11个语义规则树,从标注符号到IF表达式的映射过程就是语义规则树的赋值过程.

以闹钟任务为例,定义了如图2所示的语义规则树结构(虚线以上部分).表4以标注样例1给出的句子为例,任务分类模块得到结果是"闹钟"任务,在进行语义组块标注后,对图2语义规则树进行赋值(虚线以下部分),得到供APP应用程序解析的IF表达式.

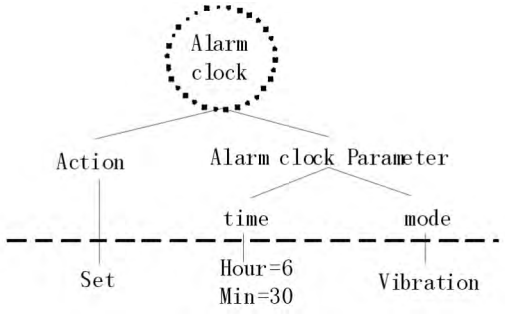

图2 闹钟设定任务的树形表示

Fig. 2 The tree of Alarm setting tasks

表4 IF生成过程示例

Tab. 4 Example of the IF generation process

步骤	输入	例子
1	口语句子	早上6点半叫我起床震动
2	任务分类标记结果	早上 AC 6 点半 AC 叫 AC 我 AC 起床 AC 震动 AC
3	语义组块标记结果	早上 ACT 6 点半 ACT 叫 N 我 N 起床 N 震动 AMM
4	IF结果表示式	IF: c: Request - Action (AlarmClock (Action = "Set", time = (hours = "6", min = "30"), Mode = "Vibration"))

4 实验结果及分析

1) 收集用户口语文本语料库,包括10任务类型,总样本数为4109条,利用BosonNLP工具对该语料库进行分词及词性标注.然后,对该语料进行手工任务分类标注及语义组块标注,完成满足CRF标注格式要求的包含全部样本的语料库.

2) 将该语料库按比例7:3分为训练语料和测试语料,且样本数分别为2876条和1233条.根据前文提到任务分类用的特征模板,通过CRF++工具对训练语料进行训练,得到操作任务识别模型,并完成对测试语料的操作任务分类测试,测试结果见表5.

表5 操作任务分类结果表

Tab. 5 Classification results of operating instructions

实验	任务总数	识别任务数	正确表达任务向量数	准确率/%	召回率/%	F值/%
实验结果	1233	1218	1204	98.85	97.65	98.25
对比实验[5]	7688	7567	7421	98.07	96.53	97.29

3) 按任务类型将语料库分成10个子语料库,每个库的语料按比例7:3分为训练语料和测试语料.根据前文提到语义组块提取用的特征模板,通过CRF++工具分别对各任务训练语料进行训练,得到各个任务对应的语义组块提取模型,并完成对各自对应的测试语料的语义组块提取测试,实验统计结果见表6.

表6 任务语义组块提取结果表

Tab. 6 Information extraction results of semantic chunk

任务类别	任务总数	发现任务数	正确数	准确率/%	召回率/%	F值/%
打电话	139	138	136	98.55	97.84	98.19
发短信	93	92	86	93.48	92.47	92.97
手机设置	152	147	141	95.92	92.76	94.31
备忘录	136	132	122	92.42	89.71	91.04
闹钟	111	110	109	99.09	98.20	98.64
手机应用	162	160	143	89.38	88.27	88.82
音乐	138	134	125	93.28	90.58	91.91
天气	137	136	136	100.00	99.27	99.63
导航	88	86	84	97.67	95.45	96.55
车票	77	74	63	85.14	81.82	83.44
平均值	1233	1209	1145	94.71	92.86	93.78

4) 对任务分类和语义组块进行综合性能测试,实验结果如表7所示,共输入1233条口语任务,系统能够识别出来的任务数有1216条,其中获得正确的IF表达式有1117条,准确率,召回率以及F值分

别达到 91.86%，90.59%，91.22%.

表 7　系统测试结果表

Tab. 7　Results of testing system

实验	任务总数	识别任务数	正确表达任务向量数	准确率/%	召回率/%	F 值/%
实验结果	1233	1216	1117	91.86	90.59	91.22
对比实验[5]	7688	7567	6832	90.29	88.87	89.57

任务发现过程中，文献[5]采用 SVM 技术建立模型所用样本数为 15294 条，而本文采用 CRF 技术建立模型所用样本数为 2876 条，仅为文献[5]样本数的 18.8%，但与文献[5]的实验结果数据（表 7 中的对比实验）相比较，F 值提高了 1.65%.

实验过程中，还发现对语料进行分词及词性标注所采用工具软件的性能对本文方法测试结果影响大. 通过提高分词及词性标注工具的分词准确率，特别是提高口语中机构名、地名、人名等命名实体的识别率能提高本文方法的性能.

5　结语

本文将口语理解分为操作任务分类和语义组块提取，两步均采用了 CRF 技术，并将最终得到的口语语义理解结果采用中间语义表示格式表示. 先进行任务分类能缩小语义分类的搜索空间，从而提高语义分析的性能. 操作任务分类中，相比文献[5]的 SVM 技术，本文采用 CRF 技术在用更少的语料训练分类模型后，反而取得了比 SVM 技术更好的实验效果. 以后可按以下思路去开展进一步研究：一是要采集多类人群的口语习惯，建立更加完善的语料库；二是考虑如何提升分词及词性标注、命名实体识别的准确率；三是建立合适的外部扩展词库，提升系统在文本检错、分词、实体识别等方面的性能.

参　考　文　献

[1] 吴尉林，陆汝占，段建勇，等. 基于两阶段分类的口语理解方法[J]. 计算机研究与发展，2008，45(5)：861-868.

[2] 解国栋，宗成庆，徐 波. 面向中间语义表示格式的汉语口语解析方法[J]. 中文信息学报，2003，17(1)：1-6.

[3] 王　彬. 汉语人机对话系统中口语处理的研究[D]. 北京：清华大学，2004：1-6.

[4] 陈俊燕，吴　及，王　侠，等. 口语对话系统中的一种稳健语言理解算法[J]. 清华大学学报(自然科学版)，2005，(01)：21-24.

[5] 郭　群，李剑锋，陈小平，等. 一种面向移动终端的自然口语任务理解方法[J]. 计算机应用系统，2013，22(8)：124-129.

[6] 罗　瑜. 支持向量机在机器学习中的应用研究[D]. 成都：西南交通大学，2007：11-14.

[7] 朱　敏. 面向多领域大规模知识库的自然语言自动问答研究[D]. 成都：. 西南交通大学，2015：25-28.

[8] Wallach H M. Conditional random fields：an introduction[J]. Technical Reports，2004，53(2)：267-272.

[9] Sutton C，Mccallum A. An introduction to conditional random fields[J]. Foundations & Trends(r) in Machine Learning，2010，4(4)：267-373.

[10] 李新德，张秀龙. 一种面向室内智能机器人导航的路径自然语言处理方法[J]. 自动化学报，2014，40(2)：289-305.

[11] 刘泽文，丁　冬，李春文. 基于条件随机场的中文短文本分词方法[J]. 清华大学学报(自然科学版)，2015，55(08)：906-910.

[12] 张朝胜，郭剑毅，线岩团，等. 基于条件随机场的英文产品命名实体识别[J]. 计算机工程与科学，2010，32(6)：115-117.

[13] 熊　英. 中文自然语言理解中基于条件随机场理论的词法分析研究[D]. 上海：上海交通大学，2009：11-36.

[14] 宗成庆. 统计自然语言处理[M]. 2 版. 北京：清华大学出版社，2013：480-485.

[15] 孙广路，郎　非，薛一波. 基于条件随机域和语义类的中文组块分析方法[J]. 哈尔滨工业大学学报，2011，43(7)：135-139.