口语对话状态追踪的研究*

任 航 徐为群 颜永红

(中国科学院声学研究所语言声学与内容理解重点实验室 北京 100190)

摘要: 口语对话系统是最自然的人机交互界面之一。然而语音识别和口语理解模块带来的级联错误会对用户体验造成很大影响,在嘈杂的环境中更为严重。对话状态跟踪器可根据对话的上下文和可观测到的语音识别、理解结果对各个回合的对话状态做出估计。因此,提出一种由数据驱动基于鉴别式模型的对话状态追踪方法,能够处理更大规模的特征集,特征函数依赖于可观测的全部 N - best 结果。通过在真实语音数据集上进行评测,实验结果表明,该方法比单纯使用 1 - best 结果的基线系统具有更强的性能。

关键词: 语音对话系统,鉴别式模型,对话状态跟踪

Studies on Spoken Dialog State Tracking

REN Huang, XU Weiqun, YAN Yonghong

(The Key Laboratory of Speech Acoustics and Content Understanding Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China)

Abstract: Spoken dialog system is a natural and intuitive human - computer interface. But the errors resulted from automatic speech recognition and spoken lanuage understanding will harm user experience, which is more serious in noisy environment. Dialog state trackers make estimation of the current dialog states by observations in dialog history. In this paper we propose a data - driven dialog state tracking method by discriminative modeling. This method can handle large feature sets which can utilize full observed N - best results. The proposed method show better performance than baseline operating on 1 - best results.

Keywords: spoken dialog system, discriminative modeling, dialog state tracking

1 对话追踪简介

口语对话系统(Spoken Dialog Systems, SDS)在人们的日常生活中已经得到了非常广泛的使用。移动语音助手是 SDS 最具代表性的应用,最近几年来,许多商业公司都推出了这类应用,作为各自软件平台的一大卖点推向市场,最为著名的有苹果公司的 Siri,谷歌公司的 Google Now 以及微软公司的 Cortana。这类语音助手的共同特点是以语音交互对话的形式作为人机交互界面,完成过去需要靠键盘或触摸操作来完成的任务。语音对话系统的使用大大的降低了计算设备的门槛,使得人们能够以更为自然的方式来使用计算机。但目前的口语对话系统还存在一些问题阻碍其大规模的普及和应用,其中由语音识别系统(Automatic

本文于 2016 - 05 - 11 收到。

* 国家自然科学基金:(编号:11461141004,61271426,11504406,11590770,11590771,11590772,11590773,11590774),中国科学院战略性先导科技专项(面向感知中国的新一代信息技术研究,编号:XDA06030100,XDA06030500,XDA06040603),国家 863 计划(编号:2015AA016306),国家 973 计划(编号:2013CB329302)和新疆维吾尔自治区科技重大专项(编号:201230118 - 3)。

Speech Recognition，ASR）和口语理解系统（Spoken Language Understanding，SLU）引入的级联错误会大大降低系统的用户体验。对话状态追踪的目的就在于从包含错误的识别和理解结果中对用户的实际意图进行估计。尽管近年来深度学习等新的机器学习方法大大降低了识别和理解的错误率，但良好的对话状态追踪方法可有效地克服识别和理解错误，从而进一步改善用户体验，所以有必要对其进行深入研究。目前常用的商用系统仅仅使用识别和理解的 N－best 输出中的最优结果，考虑完整的 N－best 结果可有效提高状态追踪的准确度。

1.1 基于统计学习原理的对话状态追踪方法

基于统计学习原理的对话状态追踪方法由数据驱动，在大规模对话语料库上进行模型训练，调整参数，性能上超越了基于规则的方法[1]。对话状态追踪可视为一个分类问题，而常见的用于统计学习模型，可分为对联合概率进行建模的生成式模型和对后验条件概率直接进行建模的鉴别式模型。在早期的对话状态追踪的研究中常使用基于动态贝叶斯网络（Dynamic Bayesian Networks，DBN）的生成式模型[2]。而近几年的研究表明，使用基于最大熵模型[3]，深层神经网络模型[4]，条件随机场模型[5]等鉴别式模型的状态追踪器可获得比生成式模型更好的性能。本文提出的对话状态追踪方法基于鉴别式模型。

2 鉴别式对话状态追踪

在任务驱动的对话中，对话状态可视为对各个预定义的语义槽的赋值。此时，对话状态追踪可视为一个分类问题，即从候选的语义值中挑选出唯一正确的结果。此时如果使用可能出现的所有语义值作为分类域会导致严重的数据稀疏，因为语义值的集合常常很大，而且有些分类结果在训练集中不会出现。不仅如此，较大的分类域会提高计算复杂度，降低了系统的使用性。于是我们对分类域进行限制，使其仅仅包含那些在 N－best 结果中出现的语义值，记为 y_t，即在回合 t 的分类域。鉴别式模型对后验条件概率 $P(S_t \mid O_1^t)$ 进行建模，其中 $S_t \in y_t$，O_1^t 为从首个回合到第 t 回合的所有观测。常用的分类器包含最大熵、条件随机场、深层神经网络等模型。这类模型的优点是可以加入任何相互依赖的特征，而不必对特征之间的概率关系建模，可以达到优于生成式模型的分类效果。

2.1 结构化鉴别式模型

上述的鉴别式模型是基于每个对话回合的，而没有利用回合间的对话状态关系，而实际上前一回合的对话状态估计和本回合状态具有较强的统计相关性，故有必要利用这部分信息，这里我们假设相邻的对话状态具有马尔科夫性。此时，条件概率 $P(S_t \mid O_1^t)$ 可分解为式（1）：

$$P(S_t \mid O_1^t) = \sum_{S_{t-1} \in y_{t-1}} P(S_t \mid O_1^t, S_{t-1})\, P(S_{t-1} \mid O_1^{t-1}) \tag{1}$$

与静态的鉴别式模型相对比，结构化模型对条件概率 $P(S_t \mid O_1^t, S_{t-1})$ 进行建模，故特征函数可利用前一回合的对话状态，我们将依赖于前一对话状态的特征称为转移性特征，反之称为普通特征。

2.2 多层神经网络分类器

我们使用多层神经网络作为分类器，神经网络中包含的丰富的非线性变换可使得模型对输入特征间的复杂的相互作用具有更强的建模能力[6]。由于预测域为动态生成，个数不能事先确定，我们对模型的计算方法进行一定的修改。在预测过程中将分类域中的每一类对应的特征向量输入网络中进行前馈计算，得到未归一化的分数，最后将各个类别对应的分数共同输入到 Softmax 函数中归一化为合法的概率值，大小在 0－1 之间，得到各个回合中对话状态的边缘概率分布，如式（2）所示。

$$\begin{aligned} y_i &= W_{l-1} \cdot g_{l-1}(\cdots g_1(W_1 \cdot X_i)\cdots) \\ P_Y &= \mathrm{Softmax}(y_1, \cdots, y_{|y_t|}) \end{aligned} \tag{2}$$

其中 g_1 到 g_{l-1} 为进行非线性变换的 Sigmoid 函数，W_i 为第 i 层线性变换的权矩阵，X_i 为第 i 类对应的输入特征 $X_i = f(O_1^t, y_i)$，y_i 为经过 Somftmax 层归一化前的输入值。与常见的深度神经网络训练方式类似，在模型的训练中使用后向传播方法计算梯度值，利用随机梯度下降进行参数优化。

3　实验验证

在口语对话系统研究中常常需要构建包括语音识别、理解、对话管理等模块的完整对话系统,由此对所研究方法进行实验比较。近几年来随着开放对话数据集的兴起,对话相关的研究可以直接在标注数据集上展开。本文实验部分使用带有对话状态标注的开放数据进行模型训练和参数调优,在测试集上进行性能测试。

3.1　数据集

本文实验部分使用 DSTC2 开放数据集(Dialog State Tracking Challenge 2) [7] ,该语料库使用真实的人机对话记录,对话领域是餐馆信息查询。对话领域信息包括了餐馆名、地区、饮食风味、价位等多个语义槽,用户可使用不同组合方式对餐馆进行查询。关于 DSTC2 的详细信息以及评价指标参考 DSTC2 手册。

3.2　实验设定

考虑到相邻对话回合所对应的对话状态具有统计相关性,我们使用鉴别式建模方式,并对模型结构进行马尔科夫假设,即将模型在前一回合输出的对话状态,与当前回合提取的特征值共同作为状态跟踪模型的输入。为了验证不同模型结构对结果的影响,共训练 4 个不同形式的模型。每种模型设定如表 1 所示,其中 MEMM 为无隐层的神经网络,实际上退化为最大熵马尔科夫模型; SNN1 和 SNN2 分别为单隐层和双隐层的神经网络; NN 网络结构与 SNN2 相同,但去除了所有转移性特征,退化为普通的神经网模型。这里选取作为对比的基线系统只使用每个回合中 N - best 结果中的最优结果。

3.3　实验结果

在 DSTC2 测试集上的结果如表 2 所示。再测试指标中 ACC 为追踪准确率,L2 度量估计的概率分数与真实概率值的欧几里得距离,CA05 度量 ROC 性能,表示当错误接受率为 5% 时的正确接受率。除了 L2 外,其他参数越大越好。ACC 与 L2 同常被认为是最重要的指标,他们直接评价了跟踪模型的准确度和输出概率分数的好坏。比较模型 MEMM、SNN1 和 SNN2,3 个模型结构复杂程度递增,由表 2 中结果可见在最重要的准确率指标上 SNN2 取得了最佳结果,综合考虑模型在结构上的差别,这体现了多层神经网络在建模中的优势。而对比 SNN2 和 NN,两者具有相同的网络结构。相比 SNN2,NN 缺少了转移性特征而使得马尔科夫结构失效并退化为静态的神经网络模型。而结果显示 NN 在准确率上比 SNN2 有所降低,证明了对话回合间的马尔科夫假设对下一回合的预测起到了一定的作用。

表 1　模型设定

模型	模型类别	隐层单元个数
MEMM	最大熵马尔科夫模型	-
SNN1	结构化神经网(单隐层	[50]
SNN2	结构化神经网(双隐层	[50, 30]
NN	普通神经网(双隐层	[50, 30]

表 2　DSTC2 测试集性能

模型	ACC	L2	CA05
基线	0.619	0.738	0.000
MEMM	0.707	0.447	0.223
SNN1	0.713	0.437	0.207
SNN2	0.718	0.461	0.100
NN	0.713	0.448	0.128

4　结束语

本文基于相邻回合对话状态的马尔科夫性假设,提出了一种针对对话状态跟踪任务的结构化鉴别性模型。分析了鉴别性模型相比产生式模型在状态跟踪中的优势,描述了结构化模型对普通的鉴别性模型的改进。通过在真实的对话语料库 DSTC2 数据集上进行实验验证,初步证明了结构化鉴别性模型的优势。

参　考　文　献

[1] Williams J, Raux A, Ramachandran D, et al. The Dialog State Tracking Challenge [C] //Proceedings of the SIGDIAL 2013 Conference. Metz, France, 2013.

(下转第 38 页)

试例的设计、验证过程的自动化执行、验证结果分析等方面进行了研究。目前，HINOC2.0 SOC 商用样片已经研发成功，经测试，包括 HINOC PHY 在内的各项功能性能均达到了设计要求，说明了本文所采用的基于 UVM 进行 HINOC PHY 验证方法是有效的且大大提高了验证效率，基于 UVM 的验证方法对大规模数字逻辑电路的测试和验证工作具有积极作用。

Scope	TOTAL	Statement	Branch	FEC Expression	FEC Condition	FSM State	FSM Trans
TOTAL	91.27%	99.25%	97.64%	95.53%	75.72%	100.00%	76.47%
U_HIPHY_TOP	100.00%	100.00%	100.00%	100.00%	100.00%	--	--
u_interface	83.49%	93.78%	90.67%	77.20%	72.31%	--	--
u_interrupt_hiphy_cpu	82.98%	94.50%	92.51%	73.33%	71.56%	--	--
U_HIPHY_CONTROL	88.30%	96.45%	94.81%	87.96%	73.99%	--	--
U_TOP_SEND	92.34%	98.36%	94.59%	97.93%	78.47%	--	--
U_TOP_RECV	91.64%	99.57%	98.36%	95.51%	76.54%	100.00%	76.47%

图 6 HINOC PHY 模块覆盖率统计

参 考 文 献

[1] 国家新闻出版广电总局广播科学研究院，北京大学，西安电子科技大学等.（GY/T 265—2012）NGB 宽带接入系统 - HINOC 传输和媒质接入控制技术规范[S]. 国家新闻出版广电总局行业标准. 2012.8

[2] 国家新闻出版广电总局广播科学研究院，北京大学，西安电子科技大学等.（GY/T 297—2016）NGB 宽带接入系统 - HINOC2.0 物理层和媒体接入控制层技术规范[S]. 国家新闻出版广电总局行业标准，2016.3

[3] 张强. UVM 实战[M]. 北京：机械工业出版社，2014.07

[4] 任宇 等. VLSI 设计中一种新型的功能验证方法[J]. 微计算机信息. 2006.

作者简介

欧阳峰，（1979 年 7 月 -），男，高级工程师，研究方向：通信工程，网络信息安全技术。

郭乐，（1987 年 11 月 -），男，工学硕士，研究方向：芯片验证技术，通信协议设计，FPGA 设计。

金淼，（1989 年 7 月 -），男，工学硕士，研究方向：宽带接入技术，芯片验证技术，通信协议设计。

（上接第 21 页）

[2] Young S，Gašić M，Keizer S，et al. The Hidden Information State model：A practical framework for POMDP - based spoken dialogue management[J]. Computer Speech & Language，2010，24(2)：150 - 174

[3] Metallinou A，Bohus D，Williams J. Discriminative state tracking for spoken dialog systems[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia，Bulgaria，2013.

[4] Henderson M，Thomson B，Young S. Deep neural network approach for the dialog state tracking challenge[C]//Proceedings of the SIGDIAL 2013 Conference. Metz，France，2013.

[5] Lee S. Structured Discriminative Model For Dialog State Tracking[C]//Proceedings of the SIGDIAL 2013 Conference. Metz，France，2013.

[6] Bengio Y. Learning deep architectures for AI[J]. Foundations and trends® in Machine Learning，2009，2(1)：1 - 127

[7] Henderson M，Thomson B，Williams J. The Second Dialog State Tracking Challenge[C]//Proceedings of the SIGDIAL 2014 Conference. Philadelphia，Pennsylvania，2014.

作者简介

任航，男，1989 年生，博士研究生，研究方向：口语对话系统。