**Using Context to Improve Emotion Detection in Spoken Dialog Systems**

阅读笔记

**Abstract**

大多数探索对话系统用户情绪状态的研究并没有充分利用对话结构所提供的上下文性质。本文报道了利用“我能如何帮助你?”语音对话系统收集的5690个对话语料库，设计的自动分类用户回合情绪状态的机器学习实验结果。结果表明，利用语音对话结构和跟踪用户状态的上下文特征对标准词汇和韵律特征进行增强，可以提高2.6%的分类准确率。

**Introduction**

在客服应答系统处于自动（部分/全部）托管时，顾客可能1.一开始就是来带着申诉的情绪;2.在人机对话中因为机器无法理解人的意思而逐渐表达出消极的情绪。而现有的系统无法检测这个问题点，所以无法修改对话策略去修复这个问题或者将呼叫转移给人工操作员。问题的指标即用户的负面情绪程度。作者认为，跟踪呼叫者的情绪状态将提高客户满意度，增加成功的互动次数。

介绍了一个商用客服电话中心用户情绪状态的分类研究。

Section2 介绍了领域其他工作

Section3 介绍了使用的语料库等

Section 4、5 详细介绍实验设计、特征描述和结果

Section6 总结、展望

**Related Work**

Research has tended to focus on either on lexical or prosodic/acoustic cues such as intonation, speaking rate, and loudness

已有情感分类研究往往集中在词汇或韵律/声学/音准，说话速度，音量。

但是训练数据都是从actors中人为提炼出来的，而且分类范围比较广泛。而在语音对话系统中，情绪检测领域的研究试图对实际系统用户更微妙的自然发生的情绪进行分类。因此，这一领域的许多研究人员使用词汇和韵律特征作为情感分类的基础，但随后使用额外的特征来增加他们的特征集，这些特征集通常旨在利用数据的会话性质。

同时基于自然语料库的研究和使用actor-elicited speech 引出的研究之间的另一个区别是前者通常只尝试识别一个人的语言的情感价值，或积极/消极的影响。而后者分类的类别更细，如：愤怒、恐惧、悲伤等等。

引用10使用来自商业部署的呼叫中心的数据，自动将呼叫者转向分为负面或非负面情绪。研究使用了一套广泛的词汇和韵律/声学特征，包括一个新颖的词汇特征，以情感显著性词汇评分的形式，表达了一个特定的单词和情感之间的相互信息。此外，研究人员还引入了对话行为的用户转向作为一个话语级的功能。对话行为可以是以下几种:拒绝，重复，重述，要求重新开始，以上都不是。从他们的研究结果可以计算出，增加的话语信息比单独使用词汇和韵律特征增加了大约3%的相对改善。

**4. Automatic Emotion Classification**

我们的实验将机器学习程序BOOSTEXTER应用于自动分类每个用户回合所传达的情绪。BOOSTEXTER是一种增强算法，它通过结合若干次迭代的弱学习者决策[15]的结果形成分类假设。对于这里报告的所有实验，我们进行了2000次这样的迭代。BOOSTEXTER允许输入特性同时接受连续值和离散值。

语料库分为训练集和测试集。训练集包含15013个用户回合(语料库的75%)，测试集由剩余的5000个回合组成。利用时间信息对语料库进行分割;用户提交训练集的时间早于测试集。此外，训练集和测试集之间没有对话框。这样划分语料库是为了模拟实际的系统开发，首先从现场采集训练数据，然后利用这些数据构建系统，最后得到性能e

为了应用BOOSTEXTER，用户在语料库中的提交被编码为一组类和一组用于类预测器的输入特性。这些课程是根据第三节中描述的七种情绪来选择的。然而，由于情绪标签的分布不均匀(73.1%为阳性/中性)，我们采用了二元分类方案:将阳性/中性重新标记为非阴性，将第3节剩余的情绪全部归为阴性。

每个用户回合都有一组80个特征，这些特征要么是自动派生的，要么是手工注释的。如以下章节所述，这些特征被分为以下四个连贯的特征集:词汇特征  
(LEX)、韵律特征(PROS)、对话行为(DA)和上下文特征(CONTEXT)

4.1. Lexical Features

LEX特征集只包含一个特征:每个用户话语的手工抄写。BOOSTEXTER的配置是这样的，即每个用户抄写的所有unigram、bigram和trigram都以“单词包”的方式考虑。除了词汇之外，抄写还包含非言语的人类声音，如笑声和叹息声。

在训练语料库中，我们注意到在用户转录中发现的某些词与情绪状态相关。虽然这些相关性很小(最高小于0.2)，但它们非常显著(p0.001)。这似乎表明，人们说的话在他们的情绪状态中起了一定作用，尽管它们可能不是唯一的指标。一些比较有趣的相关性与消极的用户状态词提到账单的一些特征(美元美分,调用)和那些表明,调用者希望被转移到人类操作员(人、人类说话,说话,机器)。此外，数据显示，充满停顿，如oh和非言语的人类噪音，如叹息，也与消极的用户状态相关。

4.2. Prosodic Features

PROS特性集包括17个特性，用于捕获用户的声学、韵律和音质信息。这个特性集中的特性的动机是试图捕获用户说话的方式，作为他们情绪状态的指示。

使用语音分析和合成[16]程序Praat，在整个用户周期内自动提取了以下10个特征:中位数和标准差，以近似响度信息;总基频(最小、最大、中、标准偏差、平均绝对偏差)近似的基音轮廓;浊帧数占总浊帧数的比例帧，以近似的说话速度。

其余7个特征是半自动提取的。通过使用AT&T WATSON的一个特殊应用程序，即一个实时语音识别器[17]，手机和静音通过用户转弯的手动抄录来识别。这些特性包括:f0在最后一个元音后的斜率，用来模拟转折-最后一个音的音高轮廓;平均f0和能量除以最长的归一化元音，近似音高重音信息;每秒音节数，平均元音长度，内部沉默百分比，近似说话速度和犹豫;以及最长归一化元音的局部抖动，作为音质的一个参数。每个元音的归一化长度取决于训练语料库中发现的历时语境和同位语境。

提取技术产生的原始特征值通常对于泛化学习算法的应用来说过于具体。因此，我们考虑了几种不同的归一化技术。最优的解决方案是通过说话人使之正常化。然而，由于平均对话框只包含3.5个用户回合，这就造成了数据稀疏性问题。但是，我们认为整个语料库的正常化将是太广泛的。因此，我们选择了一个中间地带，根据性别来规范。归一化特征值以离均值标准差(z分数)为单位表示。标准化所需的信息(均值和标准差)仅在训练语料库中计算。

4.3. Dialog Act Features

DA特征集包括一个表示当前用户回合的对话行为的特征。对话行为可以看作是话语在对话上下文中所起的作用，它可以代表人机交互的当前状态。有不同的方法来标记对话行为，从一般到特殊。在本研究中，我们使用了HMIHY 0300语料库的预注释呼叫类型。这些是一些特定的、依赖于域的对话框行为标记。每个用户回合都标有一个或多个call-type，共65个。语料库中最常见的调用类型有:Yes，当调用者确认系统发起的问题时;客户代表，当来电者要求与客户代表通话时;和帐户余额，当来电者要求听到有关他们的帐户余额的信息。

4.4. Contextual Features

介绍了上下文特性集作为一种模型现象水平超出当前的用户。用户是嵌入在一个更大的结构一个对话框,因此似乎自然使用过去的用户活动的证据来帮助通知当前用户的情感分类。因为对话框是相对较短的在我们的语料库,我们决定使用上下文信息,扩展到前面的两个用户。这个特性集包含61特性描述设计跟踪特性的其他特性集比较之前。