ISSN 1000-0054 清华大学学报（自然科学版）2011 年 第 51 卷 第 9 期 6/33
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2011, Vol. 51, No. 9 1187-1190,1195

面向语音识别错误恢复的澄清式人机对话系统

于 东[1]，贾 磊[2]，徐 波[2]

（1. 北京语言大学 汉语国际教育技术研发中心，北京 100083；2. 中国科学院 自动化研究所，模式识别国家重点实验室，北京 100190）

摘 要：在人机交互系统中，自动语音识别（ASR）错误将导致交互障碍，通过发起澄清式人机对话可以实现 ASR 错误恢复。该文提出澄清式人机对话系统结构，用于实现语音识别错误恢复，实现了系统的 4 个组成部分：ASR 错误检测、基于统计机器翻译（SMT）方法的澄清式疑问句生成模型、说话人响应分析、基于有限状态机（FSM）的对话管理模型。各模块均采用与特定任务无关的方法建立。实验结果表明：澄清式人机对话系统可以有效模拟口语中的澄清现象，在不同的错误环境中能够较好的实现 ASR 错误恢复任务。

关键词：澄清式人机对话系统；语音识别错误恢复；澄清式疑问句；对话管理

中图分类号：TP 391 **文献标志码**：A

文章编号：1000-0054(2011)09-1187-04

Human-machine dialogue clarification system for speech recognition error recovery

YU Dong[1], JIA Lei[2], XU Bo[2]

(1. International R&D Center for Chinese Education, Beijing Language and Culture University, Beijing 100083, China;
2. Digital Content Technology Research Centre, Institute of Automation, Chinese Academy of Science, Beijing 100190, China)

Abstract: Incorrect automatic speech recognition (ASR) result hinder interactions in human-machine systems. This problem can be solved by clarifying the dialogue for ASR error recovery. This paper presents a human-machine dialogue clarification system that includes ASR error detection, clarification question generation based on statistical machine translation (SMT), user response analysis, and dialogue clarification management based on a finite state machine (FSM). All metrics are not task specific. Tests show that the system can effectively clarify misunderstandings by handling mis-recognized utterances to achieve high accuracy error recovery.

Key words: human-machine clarification dialogue system; speech recognition error recovery; clarification question; dialogue management

人机交互系统通常利用自动语音识别（ASR）技术将用户的话语转换为文本串进行处理，ASR 错误将直接影响到整个系统的性能。在人际对话中，当说话者遇到理解或听音障碍时，会发起疑问请求对方重述或解释，这类疑问称为澄清式疑问，由其引发的对话称为澄清式对话，示例如下：

甲：包里面有酒和香烟吗？
乙：酒和什么？（未听清香烟二字）
甲：酒，和香烟。（对澄清疑问的解释）
乙：哦，没有没有。（已经明白甲的话语）

在人机交互中模拟澄清式对话，可以实现 ASR 错误恢复。目前，关于澄清式对话的研究均针对特定任务环境，系统根据话语语义分析结果发起澄清式对话。Schlangen[1] 计算语义置信度，对低置信度话语发起澄清式对话；Purver[2] 将话语转换为语义表达式，转换失败则发起澄清；Bohus[3] 根据语义分析结果判断错误类型发起澄清式对话。另外，任务知识的特定数据结构也可以用来指导澄清式对话。Lewis[4] 采用树状数据结构来表达查询任务；Misu[5] 计算对话的信息增益来选择对话动作，其中包括了澄清式对话。上述方法重点在于处理与特定任务相关的用户话语的歧义和模糊，不适合应用于非特定任务的 ASR 错误恢复。此外，上述方法均依赖人工设计对话策略，使系统缺乏灵活性，而且无法涵盖各种 ASR 错误模式。

本文提出一种澄清式人机对话系统，对语音识别结果进行错误分析，针对错误结果发起澄清式人机对话以实现 ASR 错误恢复。该方法可以作为一种 ASR 错误恢复的通用方法，具有灵活和自然的表达方式。实验结果表明：系统可以针对 ASR 错误

收稿日期：2011-07-15

基金项目：国家自然科学基金资助项目（90820303，60973062/F020605）

作者简介：于东（1982—），男（汉），山东，讲师。

通信作者：徐波，研究员，E-mail：xubo@hitic.ia.ac.cn

部分生成合理的澄清式疑问句,并通过发起澄清式对话有效实现 ASR 错误恢复。

1 澄清式疑问数据的获取和标注

本文利用 CASIA 语音识别系统[6]对 2 600 句中文口语语音进行识别,识别结果中有 265 句出现词错误。统计错误数据发现,74.31%的错误句子中仅有 1 词错误,出现 3 词以上错误的仅占 2.77%。

这些错误数据量无法满足机器学习的要求,需要通过在文本语料库中模拟 ASR 错误得到足够数据。错误模拟过程假设 1 个句子中至多存在 1 处错误,该错误包含若干连续错误词,可以出现在句子任意位置。按照真实错误数据中错误词分布模拟错误位置和错误词数,被选中的词将被屏蔽并代替为错误标志"Err",可以得到合理的模拟错误句子。标注人员根据模拟错误句子标注合适的澄清式疑问句,得到"错误语句—澄清疑问"句对。标注过程有如下约束:1) 不考虑句子语法错误和歧义;2) 疑问句必须包含至少 1 个正确词;3) 疑问句必须包含疑问语气词。考虑到模拟错误的随机性,允许标注人员拒绝标注难以澄清的样本。

本文共对 20 000 句中文模拟错误句子进行标注,共有 15 772 句被标注人员接受并标注对应的澄清式疑问句。标注结果的错误句子中,77.86%含 1 个错误词,15.52%含 2 个错误词,5.62%含 3 个及以上错误词,错误词数分布与真实情况下的分布接近。

2 系统框架和各部分建模

2.1 系统框架

澄清式人机对话系统需要分析 ASR 结果,判断其中的错误,针对错误发起疑问,并及时响应用户的回答,控制对话进程,最终实现错误恢复。系统框架见图 1,由如下 4 部分组成:

1) ASR 错误检测:检测 ASR 结果中是否出现错误并将错误标出。

2) 澄清式疑问句生成:模拟人的口语表达形式,根据句子中标识的错误部分生成适当的疑问句。

3) 澄清式人机对话管理:控制对话进程,更新用户话语和澄清状态,当错误被澄清后关闭对话进程。

4) 用户响应分析:分析用户回答信息,寻找有助于澄清错误的信息,更新用户话语。

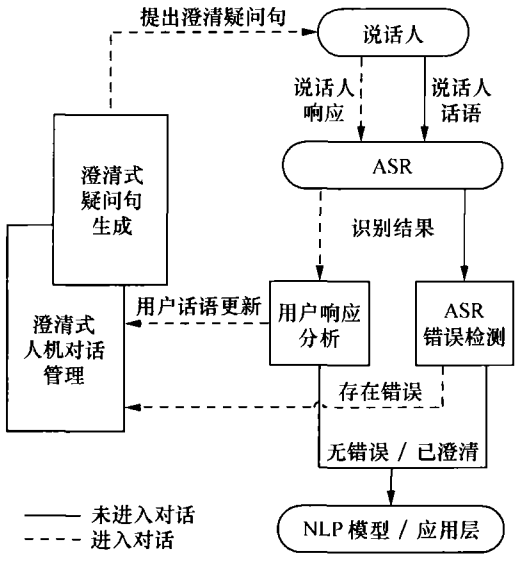

图 1 澄清式人机对话系统框架

澄清式人机对话系统可看作模拟的"听音人",在对 ASR 结果具体应用前进行错误判断,如果无错误则直接使用;否则发起澄清式人机对话,对错误进行恢复。

2.2 ASR 错误检测

本文计算 ASR 结果中每个词的词图后验概率作为置信度度量,设定置信度阈值作为判错依据。考虑到单阈值判错产生的误报和漏报,使用双阈值进行判错。图 2 中,ASR 置信度区间被划分为错误、不可信、可信 3 个子区间。阈值 T_1 用于指示错误词;阈值 T_2 用于找出错误词周围所有可能存在的错误,避免遗漏。

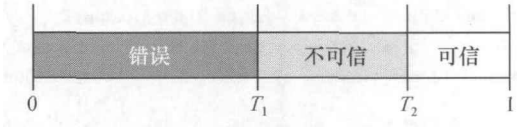

图 2 置信度区间的双阈值示意图

对于识别结果串 $W_1^l=(w_1,w_2,\cdots,w_l)$,其中每个词的置信度记为 $C(w_i)$。2 个置信度阈值分别为 T_1、T_2 且 $T_1<T_2$。T_1 过高会导致错误误判,而过低则会导致漏判。合适的 T_1 可以通过实验折中得到,T_2 在此基础上根据经验设定。记错误词集为 E,不可信词集为 U,可信词集为 B。错误判断过程可以分为 2 步:

1) 对 $\forall w_i$,如果有 $C(w_i)\leqslant T_1$,则 $i\in E$;

2) 对 $\forall w_j$,$j\notin E\vee U$,如果有 $j+1\in E\vee U$ 或 $j-1\in E\vee U$,且 $C(w_j)\leqslant T_2$,则 $j\in U$。

所有的错误词和不可信词都将被标记为"Err",

连续的错误词将被合并，原句 W_1^l 被改写为 $E_1^m|_{\mathrm{Err}(s,t)}$，其中 $\mathrm{Err}(s,t)$ 表示从 s 词到 t 词的部分被标注为错误。

2.3 基于 PBSMT 的澄清式疑问句生成

本文中，ASR 错误句和澄清式疑问句可以看作具有对应关系的 2 种“语言”，借鉴基于短语的统计机器翻译(phrase-based statistical machine translation，PBSMT)[7]方法可以实现澄清式疑问句的自动生成。在标注数据中，澄清式疑问句的平均句长为 4.35 词，而错误句子的平均长度为 7.13 词，这说明澄清式疑问句仅针对错误提出，句中正确部分会被忽略。这种信息省略具有倾向性：对于同类型错误，澄清疑问句的形式也往往相同。

因此，本文对 PBSMT 模型进行调整：首先允许源语言短语向邻接的对空词扩展，表达信息省略；其次将源语言短语中的对空词泛化为占位符，实现短语的非精确匹配。设短语对 $(T_i', Q_n^m) \in B(T,Q,A)$，$w_{\mathrm{Err}} \in T_i'$，$j-i \geqslant 2$，如果存在词串 γ 使 $T_i' = T'\gamma T''$，其中 $T' \vee T'' \notin \varnothing$，$\gamma \notin A$，则 γ 可以被重写为占位符，并有 $(T'X_1T'', Q_n^m) \in F(T,Q,A)$。其中 X_1 为占位符，只能由非泛化短语替换，$F(T,Q,A)$ 为泛化短语集。对齐短语的泛化过程有如下的约束：泛化短语中的占位符不超过 2 个；占位符不能相邻。

通过对占位符的替换，对齐泛化短语可以有效提高澄清式疑问表达能力。利用对齐泛化短语建立的澄清式疑问句生成模型结构如图 3 所示。该模型能够利用有限的澄清疑问数据来训练翻译模型，对不同领域有较好的自适应能力。同时，该方法生成的澄清式疑问句形式灵活，与用户的话语紧密相关，容易被用户理解和接受。

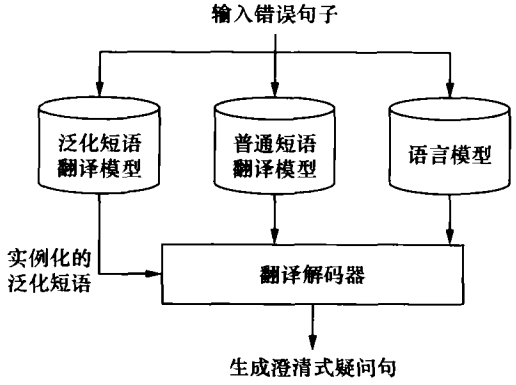

图 3 澄清式疑问句生成过程

2.4 固定澄清疑问策略

为构成完整的人机对话，除了澄清式疑问句外，系统还需要使用确认请求和重述请求这 2 种固定的对话策略，作为对澄清式疑问句的补充。当系统认为错误已被澄清时会发起确认请求，询问说话人是否接受该结果，如接受则结束对话，反之引发新一轮对话。当句子存在多处错误或所有词都被标记错误，系统放弃澄清，发出重述请求。

2.5 说话人响应分析

在澄清式人机对话中，系统不需“理解”用户回答话语的语义，而只需分析该话语是否有助于澄清 ASR 错误。因此，可以假设选取答句中某一片段替换原句的错误部分即可实现错误澄清。记用户话语为 W_1^l，错误标注结果为 $E_1^m|_{\mathrm{Err}(s,t)}$，说话人答句为 A_1^n，根据以上假设，有如下替换规则：

$$E_1^m|_{A_i'} = \mathrm{Err}(s,t) \rightarrow A_i', \quad \forall\, 1 \leqslant i \leqslant j \leqslant n. \tag{1}$$

最优替换 $\hat{A}_i'$ 可由动态规划算法计算得到：

$$\hat{A}_i' = \max_{1 \leqslant i \leqslant j \leqslant n} P_{LM}(E_1^m|_{A_i'}). \tag{2}$$

如果 $\hat{A}_i'$ 使错误句子语言模型得分提高，则认为答句有助于澄清错误，该替换有效；否则认为答句无效，系统将 E_1^m 扩展为 $E_1'^t|_{\mathrm{Err}(s-1,t+1)}$。模型如图 4 所示。

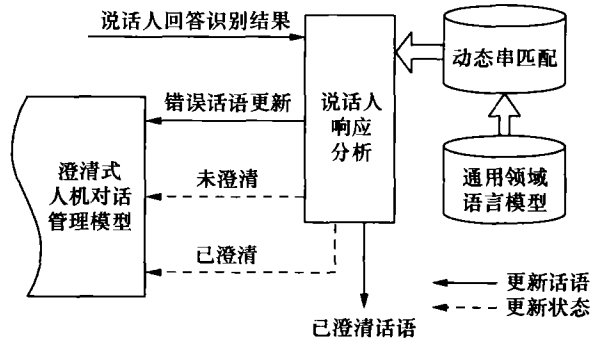

图 4 说话人响应分析模型

模型将话语分析转化为基于通用领域语言模型的字符串最优匹配问题。对于无效话语，模型通过扩展错误的方式发起新一轮澄清式对话，直至错误被澄清或者句中全部词都被标记为错误。

2.6 基于有限状态机的人机对话管理模型

本文定义二维状态特征即错误状态和对话状态，来描述澄清式人机对话系统状态。错误状态的初值由错误检测结果确定，之后根据说话人响应分析结果更新系统状态。错误状态和对话状态分别包括若干种状态值，如表 1 所示。系统状态可由二维向量表示，如(1,1)表示用户话语中存在部分词错误并且错误未被澄清。系统一共存在 7 种合理的状态。系统有 4 种对话动作：疑问式澄清、请求确认、

请求重述、结束对话。对话管理模型根据系统状态向量选择不同的动作来推进对话过程,如图5所示。

表1 澄清式人机对话系统状态定义

错误状态	状态值	对话状态	状态值
检测错误	0	已澄清	0
部分词错误	1	未澄清	1
全部词错误	2	已确认	2
无错误	3	未确认	3
		回合数超限	4

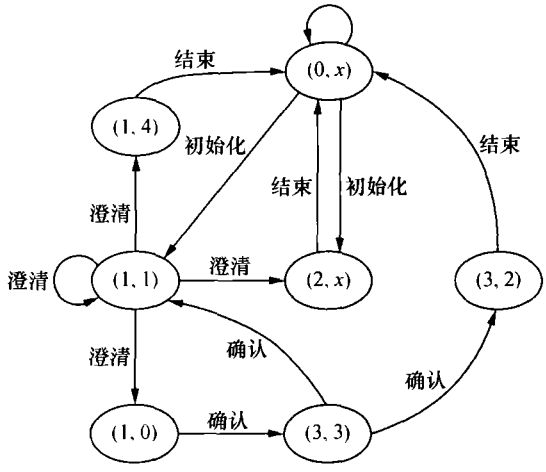

图5 对话管理模型状态转移网络

通过分析用户话语是否有助于澄清错误,从而展开对话使对话结构明确。澄清式疑问有灵活的表达方式。对话由系统主导,可以降低对话复杂度。

3 实验结果和分析

3.1 澄清式疑问句生成实验

实验中,澄清式疑问数据集被划分为训练集(13 772句对)和测试集T1(2 000句),另由真实ASR错误数据构成测试集T2(265句对)。翻译模型TM1中未使用泛化短语,翻译模型TM2中使用泛化短语。生成结果由人工评价其合理性,划分为3类:合理、可接受、不可接受。对测试集T1采用3元BLEU(bilingual evaluation understudy)打分评价方法。实验结果见表2。

表2 澄清式疑问句生成实验结果

测试集	翻译模型	3类人工评价占比/%			BLEU打分
		合理	可接受	不可接受	
T1	TM1	45.4	20.2	34.4	28.7
	TM2	68.8	16.4	14.8	41.3
T2	TM1	57.6	20.3	22.1	—
	TM2	70.7	16.3	13.0	—

实验结果表明:模型TM2的性能明显优于模型TM1,证明了泛化短语的有效性。测试集T1的3元BLEU打分也有同样的结果。测试集T2上的实验结果优于T1,是因为语音识别系统所产生的错误形式相对集中,降低了澄清难度。

3.2 澄清式人机对话实验

人机对话实验使用2 600句ASR结果句子进行,共458句被系统判定存在错误,其中233句确有错误。实验以澄清对话持续回合数作为衡量对话质量的定量指标,同时还记录了标注员对对话质量作出的优、中、差主观评级。当标注员强制终止对话时,认为对话质量为差。表3给出实验中澄清ASR错误持续的对话回合数分布情况。

表3 澄清错误所需对话回合数分布

错误类型	澄清错误所需回合数占比/%				
	2回合	3回合	4回合	>4回合	强制终止
系统判错	46.28	28.82	13.76	5.03	6.11
真实错误	44.21	27.47	14.59	5.15	8.58

实验结果中,约1/2的错误句子2回合可以澄清,约3/4的错误句子在3回合内被澄清,约10%的错误句子的澄清对话超过4回合或者被终止。这说明人对澄清式人机对话的忍耐程度大约为4个回合。

表4给出了澄清式对话的人工评价结果。80%以上的对话被评为优或中,表明对话过程比较接近人类的口语表达,系统能够较好地完成错误恢复任务。

表4 对话质量评级分布

错误类型	不同对话质量评价占比/%		
	优	中	差
系统判错	62.45	22.93	12.42
真实错误	55.36	28.33	16.31

4 结 论

本文建立了一种面向语音识别错误恢复的澄清式人机对话系统,能够根据ASR结果中错误部分以人机对话的方式实现错误恢复。系统由ASR错误检测、澄清式疑问句生成、说话人响应分析、对话管理4部分组成。实验表明:系统能够有效判断ASR错误并生成合理的澄清式疑问句发起对话,而且能够在各种错误情况下较好地澄清错误。该方法与领域无关,有较好的应用价值。

(下转第1195页)

女性和男性发音人起点的上线都为100%，为全局语调的上限。男性和女性起点的音高中线相差较小，女性为83.1%，男性为91.8%。男性和女性起点的音高下线相差较大，女性为66.2%，男性为83.6%。Q^1表示句中调群的起伏度，无论是女性还是男性，Q^1的最大值是句中调群的音高下线，最小值是音高上线。Q^2表示句末调群的起伏度，无论是女性还是男性，Q^2的最大值是句末调群的音高下线，最小值是音高上线。

从Q^1和Q^2的值的正负来看，Q^1和Q^2的值均为正值，说明维吾尔语标准话陈述句语调在时间轴上从前到后呈现出由高到低的状况，表现出陈述句语句音高总体下倾的共性规律。

U值表示全局的起伏度，无论是女性还是男性，均为正值，说明陈述句从整体上是音高下倾的。另外，从Q值大小的分布来看，从音高上线到中线到下线，Q值呈现出从小到大分布。另外，陈述句音高总体下倾也应该和发音生理有关，从句首到句末发音的气流逐渐减弱，从而导致各种语言的陈述句音高基本上都是普遍下倾的。

4 结 论

本文运用相对化的方法，采用起伏度的计算公式，对维吾尔语标准话陈述句音高进行了考察，发现维吾尔语标准话陈述句音高总体是下倾的，这种下倾是从句首调群到句末调群。这是维吾尔语标准语陈述句音高的主要特征。

陈述句音高的变化具有一定的发音生理基础。说话时气流由强到弱的变化应该是导致陈述句音高总体下倾的主要原因。

参考文献 (References)

[1] 曹剑芬. 现代语音研究与探索 [M]. 北京：商务印书馆出版，2007.
CAO jianfen. Modern Linguistic Research and Discovering [M]. Beijing: The Commercial Press, 2007. (in Chinese)

[2] 石峰. 中国语言学的新拓展 [M]. 香港：香港城市大学出版社，1999.
SHI feng. New Development of Chinese Linguistics [M], Hongkong: Hongkong City University Press, 1999. (in Chinese)

[3] 石峰，王萍，梁磊. 汉语普通话陈述句语调的起伏度 [J]. 南开语言学刊，2009，**2**：4－17.
SHI feng. The undulating scale of the intonation of declarative sentences in Standard Chinese [J]. *Nankai Linguistics*, 2009, **2**: 4－17. (in Chinese)

[4] 王萍，石峰. 汉语北京话疑问句语调的起伏度 [J]. 南开语言学刊，2010，**2**：14－22.
WANG ping, SHI feng. The undulating scale of the intonation of interrogative sentence in Beijing Chinese [J]. *Nankai Linguistics*, 2010, **2**: 14－22. (in Chinese)

[5] 根本晃，石峰. 日语声调核在陈述句语调中的表现 [J]. 南开语言学刊，2010，**1**：45－52.
GEN benhuang, SHI feng. The effect of accent nucleaus to the prosodic representation in Japanese [J]. *Nankai Linguistics*, 2010, **1**: 15－52. (in Chinese)

[6] 李爱军. 语调研究中心理和声学等价单位 [J]. 声学技术，2005，**24**(3)：13－17.
LI aijun. The psycho-acoustic units for intonation study [J]. *Technical Acoustics*, 2005, **24**(3): 13－17. (in Chinese)

（上接第1190页）

参考文献 (References)

[1] Schlangen D. Causes and strategies for requesting clarification in dialogue [C]// Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue. Boston: ACL, 2004: 136－143.

[2] Purver M. CLARIE: handling clarification requests in a dialogue system [J]. *Research on Language & Computation*, 2006, **4**(2): 259－288.

[3] Bohus D, Rudnicky A I. Error handling in the RavenClaw dialog management framework [C]// Proceedings of HLT/EMNLP. Vancouver: ACL, 2005: 225－232.

[4] Lewis C, Fabbrizio G D. A clarification algorithm for spoken dialogue system [C]// Proceedings of ICASSP2005. Philadelphia: IEEE Press, 2005: 37－40.

[5] Misu T, Kawahara T. Dialogue strategy to clarify user's queries for document retrieval system with speech interface [J]. *Speech Communication*, 2006, **48**(9): 1137－1150.

[6] GAO Sheng, XU Bo, HUANG Taiyi. A new framework for Mandarin LVCSR base on one-pass decoder [C]// Proceedings of ISCSLP. Beijing: IEEE Press, 2000: 49－52.

[7] Koehn P, Och F J, Marcu D. Statistical phrase-based translation [C]// Proceedings of NAACL/HLT. Edmonton: ACL, 2003: 48－54.