面向语音识别错误恢复的澄清式疑问句生成*

于 东[1]，贾 磊[1]，徐 波[1,2]

[1]中国科学院 自动化研究所 数字内容技术研究中心，北京 100190

[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

E-mail: {dyu, jialei, xubo}@hitic.ia.ac.cn

摘 要：人机对话系统中的语音识别错误将导致人机交互障碍。通过发起澄清式疑问是实现语音识别错误恢复的新思路。本文研究了澄清式疑问句生成问题，建立了人工标注的澄清疑问数据库，提出基于 SVM 分类器的截取模型和对齐泛化短语模型两种方法为澄清疑问模式建模，建立了基于统计机器翻译方法的澄清式疑问句生成模型。针对不同生成模型和测试集的实验证明，该模型可以有效模拟口语澄清现象，并可以有效根据错误的语音识别结果生成生成合理的澄清式疑问句。

关键词：语音识别错误恢复；澄清疑问模式；澄清式疑问句；SVM 分类器；基于短语的统计机器翻译

An Approach to Clarification Question Generation for ASR Error Recovery

Yu Dong [1], Jia Lei [1], Xu Bo[1,2]

[1] Digital Content Technology Research Centre, Institute of Automation, Chinese Academy of Sciences, Beijing 100190

[2] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190

E-mail: {dyu, jialei, xubo}@hitic.ia.ac.cn

Abstract: In human-computer dialogue system, incorrect speech recognition result may hinder human-computer interaction. A new idea of solving the problem is to prompt Clarification Question (CQ) for speech recognition error recovery. This paper presents a new approach to CQ generation. A clarification database is collected and annotated manually. Two methods of modeling Clarification Pattern (CP) are proposed, which are SVM classifier based interception model and generalized parallel phrase model. Phrase-based SMT (PB-SMT) framework is introduced to generate CQs. Experimental results show that our methods can effectively simulate clarification phenomenon in spoken language, and can generate reasonable clarification question from incorrect speech recognition results.

Keywords: speech recognition error recovery; clarification pattern; clarification question; SVM classifier; phrase-based SMT

1 引言

人机对话系统通常利用自动语音识别(ASR)技术将用户的话语转换成文字串进行处理，语音识别错误将直接影响到整个系统的性能。在人与人的口语对话中，当说话者遇到理解或听音障碍时，会用疑问的形式要求对方对话语进行重述或者解释，这种疑问句称为澄清式疑问句(Clarification Question)。作为对该语言现象的模拟，在人机对话中根据语音识别错误生成合适的澄清式疑问句，可以引导纠错对话实现有效的错误恢复。在这方面，相关研究通常预先设计一组固定的澄清策略，研究重点是如何选择最优策略。策略选择依据包括语义分析结果和任务结构知识等。Purver[1]、Bohus[2] 、Krum[3]、Sangkeun[4]分别根据语义分析结果选取合适的策略发起疑问。Lewis[5]将用户的意图映射到任务树，并根据该信息选择澄清策略。Misu[6]根据用户话语对数据库的信息增益来选择澄清策略。以上方法依赖于语义分析结果，并且需要人工编写澄清策略，因此不适合作为一种通用方法来处理语音识别系统产生的错误。

针对语音识别错误的特点，本文利用 SVM 分类器和对齐泛化短语两种方式表示澄清疑问模

* 本文承国家自然科学基金资助项目（项目号：90820303）的资助。

式，并利用统计机器翻译方法动态生成澄清式疑问句。该方法不依赖于语义分析模型，因而与特定任务无关；与固定的澄清策略相比，具有更加灵活和自然的表达方式。我们采用机器评估和人工评估来验证方法的有效性，在对口语领域的实验结果表明，我们的方法可以针对语音识别中错误部分生成合理的澄清式疑问句。

2 澄清式疑问数据的获取和标注

2.1 模拟语音识别错误数据的获取

本文基于CASIA语音识别系统[7]对中文口语语音进行识别，语音库包含2600个句子，其中有265句的One-Best识别结果出现词错误。表1中给出错误识别结果中错误词数量的概率分布和错误位置的概率分布。

表1 语音识别错误分析数据统计

错误位置	句开头	句中	句尾	错误词数	1词	2词	3词	3词以上
分布	43.75%	31.25%	25.00%	分布	74.31%	16.67%	6.25%	2.77%

由于真实语音识别过程得到的错误数据量无法满足机器学习的要求，本文由文本语料库模拟语音识别错误来得到足够的数据样本。我们假设一个句子中至多存在一处识别错误，该错误包含若干个连续的错误词，可以出现在句子的任意位置。根据表1的统计数据，首先对语料库中的每个句子按照上述两个特征分布随机选择错误位置和连续错误词数，然后被选中的词将被屏蔽并代替为错误标志“Err”，由此可以得到具有真实语音识别错误特征的模拟错误。

2.2 澄清式疑问数据的标注

本文选用IWSLT2005提供的中文口语语料作为基础语料库生成模拟语音识别错误数据。标注人员根据每一条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”句子对。标注过程有如下约束：(1)不考虑句子语法错误和歧义；(2)生成疑问句必须包含至少一个识别正确的词；(3)生成疑问句必须包含带有明显疑问语气的词。同时，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。表2给出了标注数据库的统计信息。约3/4的模拟错误数据被标注人员接受并标注澄清疑问句。标注数据中，错误词数的分布情况与真实语音识别错误中的分布接近。

表2 澄清式疑问句标注语料库数据统计

属性	数据	属性	数据
语料库句子总数	20000	1个错误词样本比例	77.86%
标注样本总数	15772	2个错误词样本比例	15.52%
模拟错误句子平均长度	7.13	3个及以上错误词样本比例	5.62%
澄清式疑问句平均长度	4.35	词典规模	6213

3 澄清疑问模式的建模和识别

表2中澄清式疑问句的平均长度仅为错误句子的一半左右。这说明澄清式疑问仅针对错误话语内容发起疑问，而忽略正确部分和具有完整意义的部分。而且这种现象有明显的倾向性：对于同一种错误情况，标注员会采用相同或类似的方式发起澄清，可以定义为澄清疑问模式：对带有语音识别错误的句子 $T_1^l=(t_1,t_2,\cdots,t_{err},\cdots,t_l)$，其对应的澄清式疑问句为 Q_1^k，其中 t_{err} 为错误位置。如果存在 T_1^l 的一个子串 T_i^j 满足 $0\leqslant i\leqslant t_{err}\leqslant j\leqslant l, j-i\geqslant 2$，且 Q_1^k 仅依赖于 T_i^j，则称 T_i^j 为错误句子 T_1^l 的一个澄清疑问模式。

澄清疑问模式可以有效描述口语中澄清疑问现象的逻辑思维过程，如表 3 中例子所示。在例(a)中，按照习惯的口语表达方式，提问者以“你想什么？”发起澄清，即确立了一个澄清疑问模式：“我 想 Err”，该模式可以完全确定澄清式疑问句“你想什么？”而且该模式在其他的错误环境中，依然能有效进行澄清。同样，在例(b)中提问者选择使用“Err 的 目光”发起疑问，而忽略了错误点前面的信息。在例(c)中，提问者同时利用了错误点两边的信息“在 Err 前面”发起疑问，其余的部分被忽略。

表 3　澄清式疑问类型举例

	例(a)	例(b)	例(c)
甲:	我想要些清洁用的绵纸。	没有女孩能抵抗你的目光。	它就在那栋楼前面。
乙听到:	我想 Error 清洁用的绵纸。	没有女孩 Error 的目光。	它就在 Error 前面。
乙提问:	你想什么?	怎样的目光?	在什么前面?

澄清疑问模式表达模型可以从两个角度分别实现。第一种为截取模型，根据错误情况和错误出现的位置进行判断，截取与错误相关的信息，删除无关的信息，从而得到澄清模式。第二种为泛化模型，将句子中无关的信息泛化为非终结符，在生成疑问句时，非终结符中的内容将被忽略。

3.1　基于 SVM 分类器的截取模型

澄清式疑问句通常只与错误句子中某一片段相关，截取句子不同的片段可以得到不同的澄清疑问模式。本文根据错误出现的位置定义三种截取方式：(1)截取前端模式：句子开头到错误点作为澄清疑问模式；(2)截取后端模式：错误点到句子结尾作为澄清疑问模式；(3)截取前后词模式：错误点及前后相邻词构成的片段作为澄清疑问模式。三种方式在标注数据库中的分布分别为 41.37%，15.54%，34.68%，占全部数据的 91.59%，即绝大多数数据都可以通过该三种方式进行截取。标注人员更倾向于使用错误前面的话语进行提问，错误点后面的话语作为辅助信息可以使提问更具体，以错误开头的提问方式较少被使用。

本文利用 SVM 分类器建立以上三种截取方式的分类器模型，分类结果将用于截取错误句子的相应部分作为澄清疑问模式。按照错误点前后词是否在疑问句中出现，可以获得截取类型标签。由于澄清式疑问句与错误点紧密相关，且距离错误点越远的词与错误的相关性越小，因此我们选择错误点位置前后 3 词范围抽取特征向量。句子长度和错误点位置也影响到澄清类型，也作为特征使用。最终共抽取 35 维特征对澄清类型进行判断。特征列表见表 4。

表 4　SVM 分类器使用特征列表

类型	特征描述	类型	特征描述
一元词特征	w_f3, w_f2, w_f1, w_b1, w_b2, w_b3*	二元词特征	w_f3f2, w_f2f1, w_b1b2, w_b2b3
一元词性特征	p_f3,p_f2,p_f1,p_b1,p_b2,p_b3	二元词性特征	p_f3f2, p_f2f1, p_b1b2, p_b2b3
前后词联合出现特征	w_f1b1, w_f1b1b2, w_f2b1b2, w_f2f1b1, w_f2f1b2, w_f2f1b1b2	前后词性联合出现特征	p_f1b1, p_f1b1b2, p_f2b1b2, p_f2f1b1, p_f2f1b2, p_f2f1b1b2
错误位置	Position	句子长度	length
位置/长度比	position/length		

*w_f3 和 p_f3 分别指示错误点前第三个词及其词性，其余可类推。

截取模型是一种获取澄清疑问模式的间接方法：首先通过有监督学习建立三种截取方式的分类模型，在应用中根据分类器结果对原句进行截取，从而得到合适的澄清疑问模式。该方法可以在小规模样本基础上建立模型，结构简单容易操作，能够迅速找到对应于错误的澄清疑问模式。

3.2 基于 SVM 分类器的截取模型

本文中，澄清式疑问标注数据可以看作是错误句子和澄清疑问句组成的双语并行语料。利用两者之间的对齐关系可以直接对澄清疑问模式建模。为此，我们借鉴 Och[8]提出的抽取对齐短语方法，根据双语句对的词对齐关系抽取对齐短语作为澄清疑问模式。在对齐短语边界的确定问题上存在两种准则："硬边界"准则禁止扩展包含对空的词(NULL)；"软边界"准则允许短语边界向邻接的一个或多个对空词扩展。澄清疑问数据中存在大量的信息省略，即存在大量的对空词汇。在抽取其中的对齐短语时，通过使用"软边界"对源语言短语进行扩展，可以使这些无关信息包含于短语中，从而通过短语的对齐关系表达澄清疑问模式。

另一方面，Och 的方法抽取的对齐短语为严格短语，只允许短语的精确匹配，而在实际情况中，一个澄清疑问模式可以应用于类似错误的澄清。为此，我们将 Och 方法抽取的初始对齐短语泛化。设短语对 $(T_i^j, Q_n^m) \in B(T,Q,A)$，$i \leqslant err \leqslant j, j-i \geqslant 2$，如果存在词串 γ，使 $T_i^j = T'\gamma T''$，其中 $T' \vee T'' \notin \varnothing, \gamma \notin A$，则 γ 可以被重写为非终结符，并有生成规则：$(T'X_1T'', Q_n^m) \in F(T,Q,A)$。其中 X_1 为非终结符，$F(T,Q,A)$ 为泛化短语集。穷举所有可能的泛化方式可以得到大量泛化短语，因此我们对泛化过程和生成过程进行如下的约束：(1)可以泛化的初始短语长度限制为 5~10；(2)一个泛化短语非终结符不超过 2 个；(3)非终结符不能相邻，连续的对空词泛化将被合并；(4)在生成过程中，非终结符只能被初始短语替换。对齐泛化短语可以有效提高对齐短语表达澄清疑问模式的能力。在遇到相似的错误环境时，通过对错误无关信息的泛化，仍可以有效忽略无关的信息，而仅针对错误信息进行澄清。

4 基于 SMT 方法的澄清式疑问句生成模型

在得到澄清疑问模式之后，需要进一步建立映射模型将其转换为澄清式疑问句，这个过程非常类似于两种语言之间的翻译转换过程，本文选择基于短语的统计机器翻译模型[9](Phrase-based SMT)来实现该生成模型。该方法能够充分利用标注的澄清疑问数据进行文本映射，可以根据不同的错误情况产生灵活的表达方式。翻译模型对不同领域有较好的自适应能力，因此模型适用于非特定任务下的错误澄清问题。在具体实现时，针对前文所述两种澄清疑问模式模型，我们分别建立对应的翻译模型。

4.1 基于 SVM 分类器的截取模型

对于截取模型，我们使用训练 SVM 分类器的数据作为翻译模型的训练语料。在模型训练阶段，我们将所有错误句子按照截取类型标签进行截取，将得到的澄清疑问模式作为源语言，参考文献[11]中的方法将对应的澄清疑问句作为目标语言训练翻译模型，并利用 SRILM 工具训练 3 元语言模型。在澄清式疑问句生成阶段，首先对 SVM 分类器得到该错误句子的截取类型，然后将截取部分输入翻译模型生成对应的澄清式疑问句。整个系统框架如图 1 所示。为了保证所生成的句子是疑问句，我们对翻译模型的解码进行了调整，包括：(1)标志错误的标记必须被翻译为疑问词，否则该翻译假设无效；(2)翻译过程面向同种语言，因此解码采用单调解码。

4.2 基于对齐泛化短语的澄清疑问句生成

对齐泛化短语能够直接表达澄清疑问模式，因此针对该模型建立的翻译模型可以直接由标注数据中的双语对齐语料训练得到。翻译模型包括两部分，分别是非泛化短语翻译模型和泛化短语翻译模型。两个模型均采用文献[11]所描述的双向短语概率特征和词汇化概率，共 4 个特征来指示短语的翻译质量。模型使用 3 元语言模型用来评价生成句子的质量。

在生成阶段，错误句子将直接作为模型输入。与一般翻译任务不同，我们需要在解码过程中

使用对齐的泛化短语，因此我们将解码过程分为两步。第一步首先搜寻能够匹配当前输入错误句子的所有泛化短语，并根据当前错误句子的具体环境替换泛化短语中的非终结符；第二步搜寻所有非泛化短语，与之前得到的实例化的泛化短语一同解码。该过程如图 2 所示。

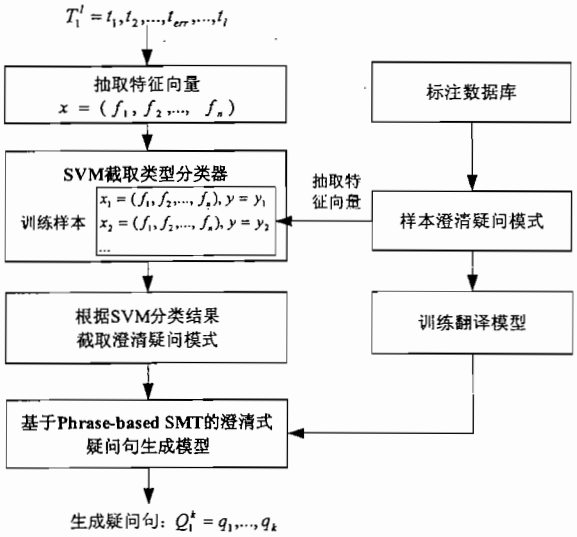

图 1　基于截取模型的澄清疑问句生成过程

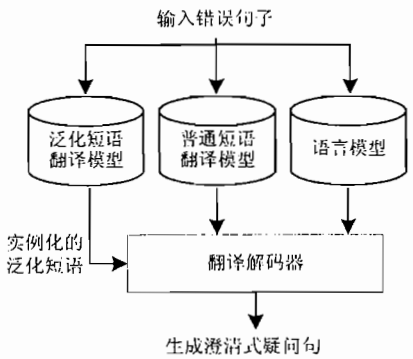

图 2　基于泛化短语模型的澄清疑问句生成过程

5　实验结果和分析

5.1　实验设定

表 5 中给出了实验数据集统计信息。标注数据被划分为训练集和测试集 1，其中翻译模型训练数据与 SVM 分类器训练和测试所用数据可以复用。实际的语音识别错误数据作为测试集 2。

表 5　实验数据集样本统计

数据集		样本数	数据集	样本数
机器翻译模型和语言模型训练集	SVM 分类器训练集	9000	测试集 1：模拟语音识别错误集	2000
	SVM 分类器测试集	4772	测试集 2：实际语音识别错误集	208

实验首先测试 SVM 分类器性能，然后测试澄清疑问模式表达模型的生成能力。在评价准则方面，对于测试集 1 我们采用 3 元 BLEU 打分作为自动评价方法。对所有三组实验结果我们采用人工评价方法评价其合理性，生成的澄清式疑问句将被划分为 3 类：(1)合理：生成的澄清式疑问句合理，可以理解且符合口语习惯，能够表达疑问；(2)可接受：生成的澄清式疑问句可以理解，能够表达疑问，但不太符合口语习惯；(3)不可接受：生成的澄清式疑问句无法理解，不能表达疑问。

5.2　SVM 分类器性能实验

我们分别测试了三个类别的正确率，错误率和 F 值测度，并计算了总的正确率。同时我们分别统计了各个类别中错误的分布。实验结果见表 6。SVM 分类器对类型 1 样本的分类正确率较高，但召回率并不高，在对类型 3 的分类上则相反。这说明分类器对类型 1 和 3 具有一定的混淆性，两者交叉错误较多。这是因为类型 3 样本可以看作是对类型 1 样本的信息补充，但这种混淆度对最终生成句子并无太大影响。分类器对于类型 2 的分类性能较为稳定，这表明采用第 2 类方式进行

澄清的现象比较固定。在所有错分样本中，分类器倾向于选择第一种类型做为分类结果，这也与汉语语言习惯接近。

表 6　SVM 分类器性能实验结果

类型	正确率	召回率	F 值	错分样本分布	归入截取前端类型	归入截取后端类型	归入截取前后词类型
截取前端	91.28%	85.00%	0.88	截取前端错误	-	18.3%	81.7%
截取后端	87.85%	84.86%	0.86	截取后端错误	66.2%	-	33.8%
截取前后词	78.80%	88.21%	0.83	截取前后词错误	80.4%	19.6%	-
总测试集	86.06%						

5.3　澄清式疑问句生成实验

为了测试澄清式疑问句生成模型的性能，我们一共训练三个翻译模型。首先直接利用训练样本中的错误句和澄清疑问句对训练 Baseline 翻译模型(TM1)。TM1 直接以错误语音识别结果为输入来生成澄清式疑问句。我们将将所有的训练集样本按照对应的澄清疑问模式标签进行截取，组成澄清疑问模式与澄清式疑问句对，训练翻译模型 2(TM2)。TM2 的输入数据为截取模型获得的澄清疑问模式。最后我们对 Baseline 系统抽取的短语表进行泛化，得到泛化翻译模型(TM3)。TM3 也以错误句子作为输入。

表 7 给出了以上三种翻译模型在模拟错误数据集，测试集 1 上的性能。在人工评测性能中，Baseline 系统得到的澄清式疑问句有约 2/3 的结果被人工评定为合理或者可接受，这证实了 PB-SMT 方法对于生成澄清式疑问句的可行性。在 TM2 的实验中，当我们用截取模型获取澄清疑问模式之后再进行疑问句生成，得到的生成结果性能有明显提高。在 TM3 的实验中，我们用泛化短语方法表达澄清疑问模式，得到了最好的结果。

表 7　模拟数据集上的生成实验结果

翻译模型	人工评测			BLEU-3 打分
	合理	可接受	不可接受	
TM1	45.4%	20.2%	34.4%	28.7
TM2	56.2%	23.6%	20.2%	36.6
TM3	68.8%	16.4%	14.8%	41.3

TM3 模型的测试性能要优于 TM2，一方面因为 SVM 分类器本身存在固有的分类错误，会导致错误传递和叠加，从而降低了系统性能；另一方面，泛化短语与翻译模型的兼容性好，在解码时可以充分考虑不同澄清疑问模式的优劣，因此使生成疑问句质量得到提高。最后，BLEU-3 得分给出的自动评价结果也得到与人工评测类似的结果，TM3 得到的澄清式疑问句与人工标注数据最为接近。这也证明了本文提出方法的有效性。

表 8 给出了三个翻译模型在真实数据集，测试集 2 上的性能。对比三个模型的性能，我们能得到与测试集 1 上类似的结果。这说明本文的方法对于真实数语音识别错误结果的有效性。另外，

表 8　真实数据集上的生成实验结果

翻译模型	人工评测		
	合理	可接受	不可接受
TM1	57.6%	20.3%	22.1%
TM2	65.3%	17.9%	16.8%
TM3	70.7%	16.3%	13.0%

在实际环境中，系统的性能要高于在模拟识别错误测试集上的结果。这是因为添加模拟错误会导致错误形式分散，任何形式的错误都会出现，从而增大了澄清难度。而一种语音识别系统所产生的错误种类是有倾向性的，因此其错误形式相对集中，降低了澄清难度。

6 结论

本文研究了面向语音识别错误恢复的澄清式疑问句生成方法。该方法能够根据语音识别结果中错误的部分提出符合口语表达习惯的澄清式疑问句，由此可以引发澄清式对话，完成错误恢复。为了克服数据量的不足，我们在模拟错误数据基础上进行人工标注得到澄清疑问数据库。为了模拟人类口语中的澄清现象，我们建立两种澄清疑问模式的识别模型，并在此基础上我们采用统计机器翻译的方法建立了澄清式疑问句生成模型，该模型根据分类器产生的澄清类型生成澄清式疑问句。实验证明，我们的方法在模拟数据和真实数据上都能够有效表达澄清疑问模式，人工评测结果显示，基于统计机器翻译的澄清疑问句生成方法可以有效生成合理的结果。该系统建立在通用口语语言领域中，可以作为一种手段对语音识别结果展开澄清式交互，在基于语音界面的人机交互系统中具有很好的应用价值。未来我们将完善澄清式疑问现象和澄清式交互的研究。

参 考 文 献

[1] Purver M., CLARIE: Handling Clari fication Requests in a Dialogue System [J]. Research on Language & Computa-tion, 2006, 4(2): 259-288.

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

[3] Krum U., Holzapfel H., Waibel A., Clarification questions to improve dialogue flow and speech recognition in spoken dialogue systems [C]//Proceeding of INTERSPEECH2005, Lisbon, 2005: 3417-3420.

[4] Sangkeun J., Cheongjae L., Gary G. L., Three Phase Ve-rification for Spoken Dialog Clarification [C]// Proceedings of IUI2006, Sydney, 2006: 55-61.

[5] Lewis C., Fabbrizio G.D., A clarification algorithm for spoken dialogue system [C]//Proceedings of ICASSP2005, Philadelphia, 2005.

[6] 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.

[7] 高升. 语境相关的声学模型和搜索策略的研究[D]. 北京:中科院自动化所博士学位论文, 2001.

[8] Och F. J. and Ney H., The alignment template approach to machine translation [J]. Computational Linguistics, 2004, 30(4): 417-449.

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