

When Calls Go Wrong: How to Detect Problematic Calls Based on Log-Files and Emotions?

Ota Herm¹, Alexander Schmitt², Jackson Liscombe³

¹Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

²Institute of Information Technology, University of Ulm, Germany

³SpeechCycle Inc., New York City, USA

hermol@fel.cvut.cz, alexander.schmitt@uni-ulm.de, jackson@speechcycle.com

Abstract

Traditionally, the prediction of problematic calls in Interactive Voice Response systems in call centers has been based either on dialog state transitions and recognition log data, or on caller emotion. We present a combined model incorporating both types of feature sets that achieved 79.22% classification accuracy of problematic and non-problematic calls after only the first four turns in a human-computer dialogue. We found that using acoustic features to indicate caller emotion did not yield any significant increase of accuracy.

Index Terms: online call quality monitoring, emotion recognition, problematic dialogue prediction

1. Introduction

Interactive Voice Response (IVR) systems, dialog systems in which human callers converse with automated agents, are being deployed with ever-increasing frequency and are becoming more and more complex. For these reasons, it is imperative to develop new means to detect and handle problematic dialogues between callers and IVRs. By “problematic” we mean calls that run the risk of ending with an unsatisfactory result. With respect to customer satisfaction, the employment of IVR systems in call centers is a double-edged sword. On the one hand, they benefit callers by allowing them to avoid being put on hold while waiting for a human agent. On the other hand, most callers prefer talking to a human operator instead of talking to a machine. It is not uncommon for the latter type of person to do whatever it takes to reach a human operator, even if it means waiting for some time for a human agent. There are other reasons to direct callers to a human agent instead of keeping them in an IVR. Some callers have needs that are not currently addressed by the IVR, others are not used to handling this emerging technology due to low media competence, and still others engage in willfully uncooperative or misleading ways. Of course, problematic dialogs can also ensue from persistent poor speech recognition performance. All of these situations can result in potentially long calls that do not end up satisfying the caller’s needs. Table 1 shows an example of a problematic conversation between a caller and an automated agent that is able to solve internet related problems serving for an ISP as a customer support system.

Most current IVRs are rather inflexible and constrained. The vision and the standard for the next generation of IVRs should be one that is more dynamic by detecting problematic calls and, ultimately, solving such problems by changing the dialog strategy. One way to do this may be to monitor the perceived emotional state of the caller and, though the ultimate goal

may be to positively alter a negative emotional state, a successful intermediate goal would be to escalate upset callers to a human agent at the earliest possible detection point. All of these actions would improve the caller experience of IVRs.

2. Related Work

Some of the first models to predict problematic dialogues in IVR systems were proposed by Walker et al. ([1], [2], [3]). Walker et al. employed RIPPER, a rule-learning algorithm, to implement a Problematic Dialogue Predictor forecasting the outcome of calls in the HMIHY (How May I Help You) call routing system from AT&T. Their classifier was able to determine whether a call was “problematic” or “not problematic” and this information was used to adapt the dialogue strategy of the dialogue manager and to fix the problem. The accuracy of the classifier after two turns between the user and the system was reported to be 79.9% using automatic features; an improvement of 24.4% over the majority class baseline of 64.0%.

Levin and Pieraccini[4] combined a classifier with various business models to arrive at a decision to escalate a caller depending upon expected cost savings in so doing. The target application was that of a technical support automated agent. Horvitz and Paek[5] consider the influence of an agent queue model on the call outcome and include the availability of human operators in their decision process. Similarly, Kim[6] presents a problematic/non-problematic classifier that is trained with 5-grams of utterances from callers that reaches an accuracy of 83% after 5 turns. Escalation is performed when the quality falls below a certain threshold.

Recognition of caller emotional state for evaluating human-computer dialogue has been proposed in [7]. In this study the corpus of DARPA Communicator consisting of simulated travel planning dialogues is used. The assumption of grouping negative emotions together and analyzing them against remaining emotions is studied. Yacoub et al.[8] presents a study on emotion recognition targeting Interactive Voice Response systems with special attention paid to the short nature of IVR utterances. While the classification performance for neutral and strong anger is over 90%, the employed data set is not suited for online classification in IVR systems. First, instead of real dialogue examples, the data set is an artificial corpus recorded with actors playing 15 different emotions. Second, IVR utterances show typically only neutral or negative emotions - emotions such as disgust, panic, and anxiety are very unlikely to appear. Third, there is no decision point employed where a specific action is performed, such as escalation to a human operator. Similarly, Petrushin[9] focusses on call center applications

Table 1: Example dialogue of a problematic call from an automated agent. Caller’s utterances are transcribed using the predicted (ASR) transcriptions.

Agent:	(chime) This is our automated Internet troubleshooter. Please hold on while I look up your account.
Agent:	Okay. I found it.
Agent:	Are you calling because you’ve lost your Internet connection? Please say yes or no.
Caller:	um sure thanks yes
Agent:	All right. Lets get that fixed. I’m going to check your modem from my end; one moment please.
Caller:	all ack [nomatch]
Agent:	Oh! Sorry, I’ll just say this again. I just need to check your modem.
Caller:	all ack [nomatch]
Agent:	Thanks for waiting. I do see your modem on our network, but I need to make sure it can communicate correctly.
Caller:	have map [nomatch]
Agent:	Oh! Sorry, I’ll just say this again. I just wanted to say that I can see your modem but I understand you cannot connect.
Caller:	have map [nomatch]
Agent:	I’m going to reset it in a moment, but before I do I need to know... Are you calling using Acme Telephone Service, a Regional Phone Service, or an Internet-Based Telephone Service? In other words, is your phone plugged into a piece of computer equipment?
Caller:	yes mam yes
Agent:	Okay, Acme’s name for this service is Acme Telephone. Do you use Acme’s telephone service?
Caller:	no no
Agent:	And who is your Internet Based Telephone Service provider?
Caller:	oh that is

Caller hangs up

with his research on emotion recognition but also uses an artificially created corpus based on 58 utterances.

Our approach accounts for both points of view considered in the aforementioned studies. First, we introduce a classification approach that can either serve as indicator for the dialogue manager to change the dialogue strategy or to escalate to a human agent. Second, we discuss the outcome of integrating acoustic features according to the emotional states to the group for classification.

3. Corpus

The following experiments are based on data collected from an automated troubleshooting agent that is currently deployed for a large High Speed Internet provider. The automated agent troubleshoots problems related to billing, provisioning, outages, integration and equipment. Callers are encouraged to describe their problems in natural spoken English. The application belongs to a new generation of dialog systems that goes beyond classical FAQ IVRs in that it takes similar steps as a human agent would to solve the issue and fix the problem. Figure 1 shows a dialog that we refer to as problematic.

The sample dialogue exemplifies a problematic call by showing an interaction in which the automated agent was unable to troubleshoot the caller’s Internet problems. As we can

see, the customer hung up after several turns because the conversation didn’t run smoothly. While the automated agent was attempting to ping the customer’s modem, the customer continued talking to the agent. Since the agent did not expect an answer from the customer at this point in time and the utterance did not match the activated standard grammars, the ASR module output nomatch-events. Evidently, the customer was not comfortable and/or experienced in dealing with IVRs. At least, the caller appeared to be uncooperative. Our aim is to detect such problematic dialogues and intervene before it is “too late.”

For our classification purposes, we employed 1,911 calls containing 22,774 caller utterances, i.e. an average about 12 utterances per call. Each caller utterance was hand labeled for strong negative emotional state (i.e. *angry*). The emotional state labeling process represents a big challenge in this field. Most emotions are expressed weakly and are difficult to account for given variations of human speech. The label was assigned only if there was a clear evidence of the speaker being angry.

In addition, each call was divided into three groups, each indicative of the success of the call. An unsuccessful call was one in which the caller problem was not solved and had to be *escalated* to a human operator. A successful call was one in which the IVR *deflected* the caller from a human agent by completely solving the caller’s problem or *partially deflected* it when some issues were solved by the IVR and the rest of the call was escalated to the human agent. Such labeling could also be referred to as *not automated*, *automated* and *partially automated*, respectively.

Such call classification is important for our domain because it is directly tied to the business model. Deflected (and, to a lesser extent, partially deflected) calls yield positive income. Escalated calls not only do not yield income, they actually cost money because the company is charged for the time spent on the call. Thus, it is highly preferable to escalate bad calls as soon as possible to minimize the loss.

4. SLIPPER

We employed SLIPPER, a fast and effective rule learner for call resolution classification. SLIPPER stands for **S**imple **L**earner with **I**terative **P**runing to **P**roduce **E**rroR **R**eduction [10] and is an instance of a supervised learning algorithm. During training it creates easily understandable and compact if-then-else-rules for classification.

Using SLIPPER provides several advantages: it is fast in classification and is thus ideally suited for integration in a real-time system such as a dialogue manager. Furthermore, its learned rules are easy to understand and facilitate comprehensible hypotheses that enable good classification accuracies. SLIPPER is based on AdaBoost, a special form of boosting algorithm that combines a number of weak classifiers to arrive at a precise predictor for classification.

For a binary classification task, SLIPPER creates a strong hypothesis $H(x)$ comprising combination of weak hypotheses $h_t(x)$ that are weighted (α) in an iterative process depending on their significance to the overall outcome. Given a set of features x the classifier outputs “-1” for class 1 and “+1” for class 2 (*sign*-function).

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right) \quad (1)$$

This final hypothesis $H(x)$ contains all weighted weak rules that make up the classification algorithm.

Table 2: Labeled corpus: non-automated calls are escalated (“E”), completely automated and partially automated are not escalated (“DE”)

	E	DE	E+DE
Number of calls (subset)	759	1,152	1,911
Average turns (subset)	17	38	30
Average length/min (subset)	1:32	5:59	4:12

In the call classification domain those weak rules for detecting a bad call can be, for example:

- if the average recognition accuracy from the ASR is below 60%
- if we observe the words “operator” or “agent” at least two times
- if the user did not respond more than three times to a question by from the IVR

SLIPPER determines such rules automatically. The model is trained with a set of features and an according to label.

5. Classification based on ASR, NLU and DM Features

Our aim is to prevent callers from hanging up prematurely without receiving help from the system. Thus we can distinguish between three call types: *completely automated*, i.e. all caller problems were solved by the system; *partially automated*, i.e. caller’s problems were solved by the system but others had to be handled by a human agent; or *not automated*, i.e. the system could not help, the caller hung up prematurely or asked for an agent. We labeled all automated and partially automated calls as “don’t escalate” (DE) and all not automated calls as “escalate” (E). Table 2 lists the call resolution distribution of the calls in our corpus.

For training the model, features from a subset of 1261 calls were used. The test set consists of 649 disjunct calls.

We extracted the following features from each dialogue turn:

ASR features Raw ASR transcription of the caller’s utterance (*utterance*). ASR confidence of returned utterance transcription, from 0–100 (*confidence*). Name of the grammar that returned the parse (*triggeredgrammarname*). Whether the caller communicated with speech or keypad (*inputmode*). Whether the speech recognizer returned a valid parse (‘Complete’) or not (‘No Input’ or ‘No Match’) (*recognition status*). Whether or not the caller began speaking before the prompt completed (*bargedin*).

NLU features The semantic parse of the caller utterance as returned by the activated grammar in the current dialogue module (*interpretation*). A number of systems tries to retrieve parsable caller input (*loopname*). The first time the system prompts the caller, *loopname* has the value ‘Initial.’ Subsequent re-prompts can either be ‘Timeout’ or ‘Retry.’

Dialogue Manager features The number of previous tries to elicit a valid response from the caller (*roleindex*). Whether the system requested substantive user’s input or confirmed previous substantive input (*rolename*). Type of dialogue activity (*activitytype*), e.g. ‘Question’ or ‘Announcement.’ Duration of the dialogue turn, in seconds (*duration*).

Separated models were created for each turn in the dialogues. The model according to turn 1 was trained with features from first turn only, the model according to turn 2 was trained with features from turns 1 and 2 and so on. The results are presented in Table 3 and Figure 1. After the fourth turn the classification accuracy is 79.22%. At this point it is the best time to decide, if the call shall be processed further or escalated to the human agent.

Figure 1: Accuracy of trained models on the testing set according to the baseline - random guess of majority class.

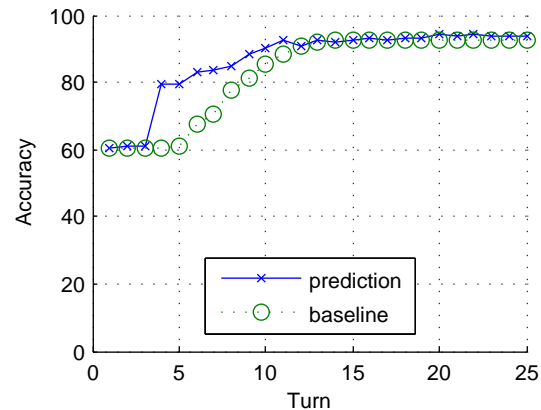


Table 3: Accuracy of each classifier at turn n .

Turns	Accuracy	Test Error Rate	Baseline
3	60.86%	39.14%	60.28%
4	79.22%	20.78%	60.66%
5	79.28%	20.72%	60.92%
6	82.96%	17.04%	67.37%
7	83.91%	16.09%	70.81%
8	84.97%	15.03%	77.52%
9	88.50%	11.50%	81.49%
10	90.40%	9.60%	85.20%
11	92.42%	7.58%	88.12%
12	90.70%	9.30%	90.61%
13	92.35%	7.65%	91.84%
14	91.69%	8.31%	92.31%
15	92.71%	7.29%	92.46%
16	92.99%	7.01%	92.64%
17	92.59%	7.41%	92.33%
18	93.24%	6.76%	92.64%

6. Emotion Recognition for IVRs

Many studies in the field of Emotion Recognition concern the emotional corpuses based on high quality studio recordings acted by trained professionals, as could be seen for instance in

Table 4: Dialogue result versus recognizable emotional state: evaluated on testing subset of 651 dialogues

State	Neutral	Negative
Escalated	84.4 %	14.0 %
Not escalated	86.1 %	13.7 %
Total	85.4 %	13.9 %

[11]. By this approach great accuracy over five basic emotions or their combinations could be reached. But this situation is far away of real applications, because people typically do not express their emotions explicitly enough.

For the task of predicting dialogue failures in IVR systems the situation is significantly different. One assumes that positive emotion plays a relatively unimportant role. However, it can also be postulated that problematic calls correlate with negative caller affect.

The contribution of the acoustic information for call result classification was tested by adding automatically extracted features to the feature vectors used by the SLIPPER. But this step added no benefit and the classification results based only on acoustic features stayed under the baseline.

7. Correlation of user emotional state and dialogue result

The distribution of negative emotion in escalated and deflected calls is illustrated by Table 4. Surprisingly, the amount of “negative” expressions in escalated calls is very similar to the amount in the deflected calls. At least in our corpus, the recognition of the caller emotional state does not bring much helpful information for predicting the dialogue result.

8. Conclusion

This study has explored a powerful classification model for predicting problematic dialogues in IVR systems. Successful classification is based on ASR features, NLU features and Dialogue Manager features. Using this method, we observed a classification accuracy of 79% after only four turn of dialogue. This is 31% better than baseline performance. Using this result for decision of further call processing or escalating to a human operator can bring significant cost reduction based on call-duration basis. Another important contribution is the potential for significantly lowering caller dissatisfaction, caused for example by non standard wishes of the customer, lack of user experience or cooperativeness, or ASR problems. Early detection of problematic calls, then, can bring not only cost reduction, but can also lead to a better acceptance and usability of IVR systems. This study is based on a corpus of real recordings made in a recent call center IVR system, deployed as a trouble shooting application for several U.S. High Speed Internet service providers. All recordings were taken over telephone lines. We were surprised to find that negative caller emotions were equally prevalent in successful calls as were in problematic calls. In our case, it meant that information logging semantic parses and dialog transitions was more effective than emotion monitoring in the automatic detection of problematic calls.

Future work with this corpus will include employment of other supervised learning methods such as neural networks and

support vector machines. Extension of the feature space on lexical and acoustic information will be accomplished. Distribution of the callers’ emotional state, especially the changes within the dialogue, will be further investigated, but it depends on more detailed labeling of the speech corpus, that is still going on.

9. References

- [1] M. Walker, J. Wright, and I. Langkilde, “Using natural language processing and discourse features to identify understanding errors in a spoken dialogue system,” in *Proc. 17th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 2000, pp. 1111–1118. [Online]. Available: citeseer.ist.psu.edu/walker00using.html
- [2] M. Walker, I. Langkilde, J. Wright, A. Gorin, and D. Litman, “Learning to predict problematic situations in a spoken dialogue system: Experiments with how may i help you,” 2000. [Online]. Available: citeseer.ist.psu.edu/541013.html
- [3] M. Walker, I. Langkilde-Geary, H. Wright, J. Wright, and A. Gorin, “Automatically training a problematic dialogue predictor for a spoken dialogue system,” 2002. [Online]. Available: citeseer.ist.psu.edu/article/walker02automatically.html
- [4] E. Levin and R. Pieraccini, “Value-based optimal decision for dialog systems,” in *Workshop on Spoken Language Technologies (SLT 06)*, December 2006.
- [5] E. Horvitz and T. Paek, “Complementary computing: policies for transferring callers from dialog systems to human receptionists,” *User Modeling and User-Adapted Interaction*, vol. 17, no. 1-2, pp. 159–182, 2007.
- [6] W. Kim, “Online call quality monitoring for automating agent-based call centers,” in *Proceedings Interspeech 2007 ICSLP*, Antwerp, Belgium, 2007.
- [7] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, “Prosody-based automatic detection of annoyance and frustration in human-computer dialog,” in *Proceedings Interspeech 2002 ICSLP*, Denver, Colorado, 2002.
- [8] S. Yacoub, S. Simske, X. Lin, and J. Burns, “Recognition of emotions in interactive voice response systems,” in *Proceedings of EUROSPEECH, Interspeech*, 2003, pp. 729–732.
- [9] V. Petrushin, “Emotion in speech: Recognition and application to call centers,” 1999. [Online]. Available: citeseer.ist.psu.edu/petrushin99emotion.html
- [10] W. W. Cohen and Y. Singer, “A simple, fast, and effective rule learner,” in *In Proceedings of the Sixteenth National Conference of Artificial Intelligence, 1999*, 1999.
- [11] “A database of german emotional speech,” in *Proceedings Interspeech 2005 ICSLP*, Lisboa, Portugal, 2005.