MODELING THE COST OF MISUNDERSTANDING ERRORS IN THE CMU COMMUNICATOR DIALOG SYSTEM

Dan Bohus *Alexander I. Rudnicky*

School of Computer Science,
Carnegie Mellon University,
Pittsburgh, PA, USA

ABSTRACT

We describe a data-driven approach that allows us to quantify the costs of various types of errors made by the utterance-level confidence annotator in the Carnegie Mellon Communicator system. Knowing these costs we can determine the optimal tradeoff point between these errors, and tune the confidence annotator accordingly. We describe several models, based on concept transmission efficiency. The models fit our data quite well and the relative costs of errors are in accordance with our intuition. We also find, surprisingly, that for a mixed-initiative system such as the CMU Communicator, false positive and false negative errors trade-off equally over a wide operating range.

1. INTRODUCTION

Misunderstanding of user input, often precipitated by recognition errors, can be a major source of user frustration in dialog systems. More concretely, when a system misunderstands the user, and then acts erroneously on the misunderstanding, the user is forced to take corrective action, either by explicitly invoking a repair sub-dialog (for example, through speaking a correction keyword) or by restating the input. It follows that if the system is capable of monitoring its performance and identifying situations in which the likelihood of misunderstanding is high, then it can choose the most efficient response (i.e., one least damaging to the progress of the dialog). In general, the ability to monitor one's own performance and to "know that you don't know" can contribute to a more fluent and intelligent dialog, at the very least one that minimizes the introduction of incorrect information.

Effective performance monitoring requires a solution to three problems: 1) misunderstandings need to be detected, 2) a cost needs to be assigned to action alternatives in a given context, and 3) an appropriate recovery strategy needs to be selected. The present paper focuses on the second of these issues.

The work described in this paper makes use of an utterance level confidence annotator which we have previously described in [1]. The annotator employs features from different sources in the dialog system (decoder, parser, and dialog manager) to classify an utterance as understood or misunderstood, and achieves a 53% relative reduction in error rate from the baseline system concept error.

We have since integrated this classifier into the CMU Communicator system [2] and have continued to improve its accuracy. The greatest gains were obtained by a cleaner re-annotation of the training corpus, and by differentiating the binary *expected_slot* feature into a three-level feature: expected, accepted and unaccepted. We further investigated classification approaches, concentrating on the AdaBoost classifier, and on a logistic regression model. Although the two classifiers produce similar error rates (around 14%) in a 10-fold cross-validation on a new dataset, the logistic regression model performs better on a soft-metric: the average log-likelihood of the test data is -0.52, while for AdaBoost is -0.88. These experiments confirm that a density estimator model (e.g. logistic regression) is more appropriate if one intends to use the confidence rating as a probability (rather than make hard decisions) in the dialog management process.

2. MODELING THE COST OF MISUNDERSTANDINGS

An issue not addressed in our work so far is that of modeling the costs of the various types of errors made by the confidence annotator. When training a classifier, we typically minimize the total error rate, i.e. the sum of the false-positives (false acceptances) and false-negatives (false rejections). The classification is therefore optimized under the implicit assumption that the costs for these two types of errors are the same.

However, intuition tells us that this assumption is probably violated in most spoken dialog systems: a false-positive error should generally cost more than a false-negative, as the system will accept and possibly use incorrect information. This will require correction and will thereby slow down the progress of the dialog. Although the cost clearly depends on details of the system and on the dialog strategy chosen, we generally believe that accepting incorrect information is likely to lead to greater costs than simply rejecting a correct user input.

We describe a data-driven investigation of the costs of these various types of errors. We propose the following approach for computing the costs: first, identify a suitable performance metric (e.g. efficiency, completion, user satisfaction), which we intend to optimize. Next, create a statistical regression model[1], relating this performance metric to the counts of the various types of errors that occur in a dialog. Obtaining a good fit will give us a robust quantitative assessment of the (negative) contribution of these different types of errors to our metric. Finally, use the costs determined in the regression to optimally tune the classifier, so that the chosen performance metric is maximized.

There are several advantages in keeping the model for the cost of errors decoupled from the confidence annotator: first, it allows us to obtain a quantitative assessment of the costs. Moreover, it allows us to target global performance metrics, and thus capture the effects of the confidence errors across an entire session rather

[1] using whole dialogs as datapoints

than within any single utterance. We currently assume that error cost is constant throughout the dialog. Given a sufficiently large corpus this assumption could be relaxed.

Smith and Hipp [3] propose the use of dialog work analysis to determine the optimal tradeoff point between these types of errors. Compared to their approach, ours is entirely data-driven. Regression models have been used previously to evaluate dialog performance in the PARADISE framework [4, 5]. Our work is different in that it is targeted at assessing the cost of several precise types of errors with the final goal of optimizing the performance of the confidence annotator for a particular spoken dialogue system. To our knowledge, this is the first empirical investigation of the costs of misunderstanding errors in spoken dialog systems.

3. EXPERIMENTS AND RESULTS

In this section we detail the experiments performed and the results obtained. After a brief description of the corpus used, we illustrate the incremental development of three successively more detailed models that use dialog efficiency as the targeted response variable. Next, we briefly describe two additional models which target completion and user satisfaction as performance metrics. Finally, we show how to use the obtained cost model to determine the optimal tradeoff point between various types of errors committed by the confidence annotator.

3.1. Dataset

A total of 134 dialogs (2561 utterances) were used, collected mostly using 4 different scenarios. The scenarios varied across dimensions such as number of legs, and hotel and car requirements.

User satisfaction scores (on a scale from 1 to 5) were obtained for 35 of these dialogs. A human annotator manually labeled the dialogs for task completion. Each utterance was also manually labeled at the concept level, and whole-utterance labels were automatically generated. The annotation scheme provided for 4 labels, applied to each concept identified in the user input. Concepts corresponded to slots in the semantic grammar used in the Communicator system, which in turn was based on an ontology of the travel domain. The labels used were: OK, RBAD (recognition-based error), PBAD (parse-based error) and OOD (out-of-domain utterance). The generated aggregate utterance labels[2] were compared with the logged decisions of the confidence annotator running in the system, and the counts for each type of error were computed.

Dataset statistics	Total	Mean per dialog	Std.Dev. per dialog
# of dialogs	134	-	-
# of utterances	2561	19.11	9.34
# CTC	1983	14.80	7.64
# ITC	166	1.24	1.61
# REC	2373	17.71	9.25
CTC / Turn	-	0.77	0.23
CTC-ITC / Turn	-	0.71	0.28

Table 1. Dataset statistics

[2]For the present investigation, mixtures of OK and BAD labels were considered BAD at the utterance level

3.2. Optimizing dialogue efficiency

The primary objective metric used was the efficiency of the dialog, as measured by the rate at which the system obtained accurate information from the user. This is a reasonable choice, as the timely completion of the Communicator task requires the system to correctly acquire flight constraints and to efficiently navigate possible solutions.

3.2.1. Model 1: CTC = FP+FN+TN

The response variable for this model is the number of correctly transferred concepts (**CTC**) per turn. For example, in the utterance below, there is only one correctly transferred concept: [Depart_Loc]. Although the label for [I_want] is OK, we count only those concepts that the system uses. Note that if the confidence annotator's decision had been to reject this utterance, then CTC would be 0, as no transfer of information from the user to the system would have occurred.

User says:	I want to fly from Pittsburgh to Boston
Sys. recognizes:	I want to fly from Pittsburgh to Austin
Concepts:	[I_want/OK] [Depart_Loc/OK] [Arrive_Loc/BAD]
Decision:	Accept

The predictor variables are the proportion of false positives (**FP**), false negatives (**FN**), and true negatives (**TN**) in the session. Although the true-negatives are not errors per se, their inclusion provides a better fit for the model.

We constructed a linear regression model and the results are illustrated on the first line in Table 2. The R^2 value of 0.81 indicates a good fit. The robustness of the model was verified using a 10-fold cross-validation experiment. The means of the R^2 for the 10 runs on the training and testing set are also shown in Table 2.

3.2.2. Model 2: CTC-ITC=(REC+)FP+FN+TN

We can refine the first model by also minimizing the number of incorrect concepts transmitted (e.g. in the utterance above, there is one incorrectly transferred concept: [Arrive_Loc]). To do this, we extend the response variable to take into account the number of incorrectly transferred concepts (**ITC**) per turn . Using **CTC-ITC** for the response variable improves the fit.

Furthermore, we can add another predictor variable: the number of relevantly expressed concepts (**REC**) per turn, regardless of whether the system perceives them correctly or not (e.g. in the utterance above, there are 2 relevantly expressed concepts: [Arrive_Loc] and [Depart_Loc]). This variable contributes to the model by capturing the user's verbosity (a user who expresses more relevant concepts in an utterance is likely to have a higher CTC).

This model provides a better fit: R^2 = 0.89 (Table 2, third line). An inspection of the coefficients computed in the regression shows that the costs for the false-positives and for the false-negatives were very similar (-1.46 and -1.44 respectively). An analysis of this somewhat counterintuitive result suggested an additional refinement to the model.

3.2.3. Model 3: CTC-ITC=REC+FPNC+FPC+FN+TN

An important observation is that there are two conceptually different types of false-positive errors in the Communicator system.

Model	R^2 **on entire dataset**	**Mean** R^2 **on training set**	**Mean** R^2 **on testing set**
CTC = FP+FN+TN	0.8160	0.8169	0.7336
CTC-ITC = FP+FN+TN	0.8650	0.8657	0.7866
CTC-ITC = REC+FP+FN+TN	0.8910	0.8912	0.8325
CTC-ITC = REC+FPNC+FPC+FN+TN	0.9436	0.9439	0.9014

Table 2. Models for cost of confidence errors. See text for meaning of symbols.

If the utterance contains relevant concepts, and the confidence annotator commits a false-positive, the system will accept and use invalid information (e.g. using Austin as the arrival city in the example above). We call this type of error a *false-positive with concepts* (**FPC**). If there are no relevant concepts in the utterance, then the system will inform the user that it misunderstood, acting exactly the same as on a true-negative. We call this last error a *false-positive with no concepts* (**FPNC**).

The impact of these two types of false-positives on the dialog is clearly different. Therefore, in the third model we replaced the **FP** predictor variable with **FPC** and **FPNC**. This model provides an even better fit ($R^2 = 0.94$). The resulting coefficients and their 95% confidence intervals are listed in Table 3.

	Coef.	**Confidence interval**
Constant term	0.4188	0.3075 – 0.5302
REC	0.6254	0.5269 – 0.7239
FPNC	-0.4820	-0.6934 – -0.2707
FPC	-2.1222	-2.2894 – -1.9550
FN	-1.3302	-1.5429 – -1.1175
TN	-0.5588	-0.7025 – -0.4151

Table 3. Regression coefficients

The relative costs confirm the intuition: false-positives with concepts are most expensive, while false-positives with no concepts cost about the same as true negatives.

3.3. Other models

While a model of net concept transmission is of immediate interest in determining how to effectively use a confidence annotator, we can also consider other response variables that appear to be correlated with "good" dialogs, such as task completion. Since completion is defined as a binary variable, we can use logistic regression rather than linear regression for the model. The model did not provide a very good fit, which is not very surprising given that factors other than utterance rejection will likely also affect task completion.

We also constructed a linear regression model with user satisfaction as the response variable. Following the PARADISE framework [4, 5], we used completion and accuracy as the predictor variables. As we are interested in the individual contributions of the various types of errors that the confidence annotator commits, we decomposed accuracy into the **FP**, **FN** and **TN** factors.

Unfortunately, we were able to obtain user satisfaction scores for only 35 dialogs. The fit for the model constructed using these datapoints – $R^2 = 0.61$ – is comparable with results reported in the literature [4]. Since user satisfaction is probably the ultimate performance metric for a dialog system, we intend to collect additional data, with the goal of understanding whether the factors of interest in this paper have significant impact on user satisfaction.

3.4. Tuning the confidence annotator

Now we illustrate how to optimally tune the confidence annotation classifier with regard to the costs determined by the previous models.

In order to make a hard decision, most classifiers compare the output of the classification process with a threshold. By changing the threshold, we can bias the classifier towards more false-positive or more false-negative errors. Figure 1 illustrates the error rates for the different types of errors (**FPNC**, **FPC**, **FN**, **TN**) that the logistic regression confidence annotator makes, as a function of the classification threshold.

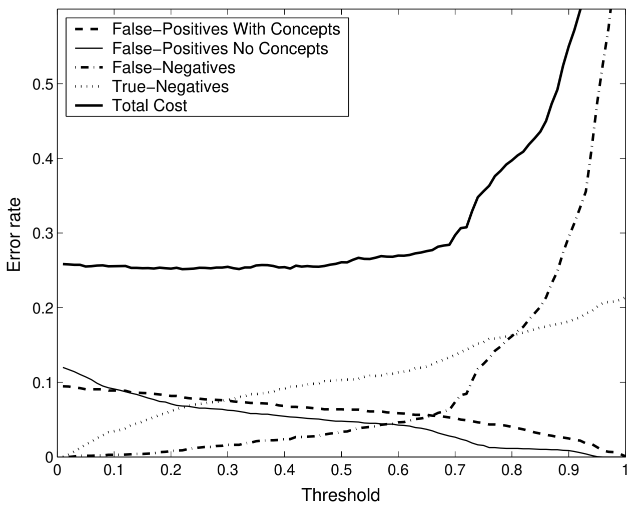

Fig. 1. Errors tradeoff and Total Cost as a function of classification threshold

To determine the optimal tradeoff between false positives and false negatives, we identify the threshold value that maximizes the regression expression, and thus implicitly the response variable – dialog efficiency. Since the **REC** factor (user's verbosity) is independent of the chosen threshold, and since the constant factor does not influence the location of the maximum, we only need to minimize the following cost:

$$TotalCost = 0.48FPNC + 2.12FPC + 1.33FN + 0.56TN$$

We plotted this function (Figure 1), but no minimum could be clearly identified. This is a surprising, somewhat counterintuitive, and very interesting result. The fact that the cost function is almost constant across a wide range [0-0.5] of the threshold values indicates that, to a large extent, the efficiency of the dialog stays about the same (at least in terms of the metric we have chosen to investigate), regardless of the ratio of false-positives and false-negatives that the system makes. Even when the threshold is set to zero, which is equivalent to completely eliminating the utterance level confidence annotator, the degradation in efficiency measured as **CTC-ITC** would be insignificant. A very similar result was obtained for the AdaBoost-based confidence annotator.

4. FURTHER ANALYSIS

In trying to better understand this unexpected result, we performed several additional experiments and checks.

First, we questioned the appropriateness of **CTC-ITC** as a response variable. An analysis of the distribution of this variable showed a rather large variance across dialogs (see Table 1). Furthermore, the mean values for the completed and the uncompleted dialogs were 0.82 and 0.57 respectively. A t-test showed that these means are statistically different with a very high level of confidence ($p = 7.23 \cdot 10^{-9}$). These results, together with the robust fit suggest that indeed, **CTC-ITC** is an appropriate response variable.

The next issue we addressed was the coverage for the model in terms of predictor variable values. Since the training data for the cost model was collected from the system running the confidence annotator with a threshold of 0.5 (which implies on average a certain proportion between FP- and FN-errors), it could be argued that the data does not allow us to construct a model which extrapolates correctly to other ratios between FP- and FN-errors (e. g., at extreme threshold values). However, an analysis of the distribution of the number of these errors in the dialogs showed that this was not the case.

We also evaluated the impact of the baseline error rate. A plot of the cost function determined based only on the dialogs with a low error rate, indicated that in this setting, the optimal threshold for the classifier is at zero (equivalent with eliminating the classifier). This observation, corroborated our previous results, and seems to indicate that for spoken dialog systems in which the user can easily override incorrectly captured information, the confidence annotator does not improve efficiency if the baseline error rate is low.

We can perhaps understand this result in the following way: in a mixed-initiative system, the user is able to correct errors by simple re-statement of the input which if now correctly understood can overwrite the incorrect previous entry. Thus the effective cost of a false-positive is essentially equal to that of a false-negative (for which restatement is naturally indicated). At low error-rates the likelihood of repeated misrecognition is low enough so that simple repetition will be able to move the dialog forward; the same strategy works of course for the false negative condition.

5. CONCLUSION

It is generally believed that tracking confidence of understanding and having dialog strategies take confidence into account leads to better dialogs. In this paper we present a data-driven approach to quantitatively assess the costs of the various types of errors committed by an utterance-level confidence annotator. We found that models based on net concept transfer efficiency fit our data quite well and that the relative cost of false positive and false negative confidence decisions are in accordance with our intuitions (i.e., false positives being on the whole more costly that false negatives).

For the classifier used in our work, however, we found that across a wide range of the receiver operating characteristic curve, the total cost stays the same. Moreover, the result indicates that, even without an utterance level confidence annotator, the efficiency of the dialog (as measured by the net correctly transmitted concepts per turn) would be the same. In a sense, this result is specific to the classifier we have developed and to the repair strategies supported by the Communicator system.

Given the counterintuitive nature of this result, we are conducting further experiments (for example, running the system with a very low confidence annotation threshold) to empirically check and further explore the predictions made by the cost model.

6. ACKNOWLEDGEMENTS

We would like to thank Roni Rosenfeld, Tania Liebowitz, Kayur Patel and Amit Mirpuri for their help throughout various stages in this work. This research was sponsored in part by the Space and Naval Warfare Systems Center, San Diego, under Grant No. N66001-99-1-8905. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

7. REFERENCES

[1] Carpenter P., Jin C., Wilson D., Zhang R., Bohus D., Rudnicky A., “Is This Conversation On Track?”, Proceedings of Eurospeech, 2001.

[2] Rudnicky, A., Thayer, E., Constantinides, P., Tchou, C., Shern, R., Lenzo, K., Xu W., Oh, A. “Creating natural dialogs in the Carnegie Mellon Communicator system”, Proceedings of Eurospeech, 1999, 4, 1531-1534.

[3] Smith R.W., and Hipp D.R. “Spoken Natural Language Dialog Systems”, Oxford University Press, 1994

[4] Walker M., Kamm C., Litman D. “Towards Developing General Models of Usability with PARADISE”, Natural Language Engineering, 1998

[5] Walker M.A., Litman D.J., Kamm C.A., Abella A. “PARADISE: A Framework for Evaluating Spoken Dialogue Agents”, Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, 1997.