Incorporating ASR information in Spoken Dialog System confidence score

José Lopes [1,2], Maxine Eskenazi [3], Isabel Trancoso [1,2]

[1] INESC-ID Lisboa, Portugal
[2] Instituto Superior Técnico, Lisboa, Portugal
[3] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
jose.david.lopes@l2f.inesc-id.pt

Abstract. The reliability of the confidence score is very important in Spoken Dialog System performance. This paper describes a set of experiments with previously collected off-line data, regarding the set of features that should be used in the computation of the confidence score. Three different regression methods to weight the features were used and the results show that the incorporation of the confidence score given by the speech recognizer improves the confidence measure.

Keywords: Spoken Dialog Systems, Confidence measures

1 Introduction

Spoken Dialog Systems (SDS) have to deal with many sources of uncertainty before performing an action based on the user's input. First, there are errors introduced by the speech recognizer. Parsing may not be sufficient to resolve ambiguity, meaning that the same recognized word could represent different actions for the SDS. To reduce the effect of uncertainty it is very important to select a set features could help improve the accuracy of the confidence scores and consequently improve the quality of the interaction.

Since speech recognition is the major source of uncertainty, previous studies tried to add semantic and context information that could reduce the unpredictability of the speech recognition output.

The different approaches to the problem are very dependent on the quality of the specific modules used in the system. The study presented used a dialog system developed with the Olympus architecture [1], however to have speech recognition in European Portuguese, the in-house Automatic Speech Recognition (ASR) module [2] was plugged in. This introduced a new challenge to compute confidence scores, since in the previous architecture with PocketSphinx [3] the acoustic confidence score was not taken into account in the computation of the overall confidence score. The confidence annotator module already computes a set of features to help the Dialog Manager make a better decision, however only the ratio of uncovered words at each stage of the dialog has been used.

The introduction of confidence scores coming from this ASR module and other features was studied in order to observe their impact on the confidence

measure computed by the system. Three different methods for computing regression were compared, simultaneously with various rejection thresholds.

In addition, since the previous studies with the integration of the new ASR module target the improvement of ASR performance by choosing the lexical primes that the system has proposed [4], it is very important to have reliable information from the ASR module in order to decide whether or not retain a prime.

2 Related Work

Many studies have been carried out that approach to this problem from the point of view of error recover. The intuitive idea here is that for SDS, a misunderstanding costs more than a rejection. Bohus et al [5] used two features to build two different regression models, the number of correctly and incorrectly transfered concepts for each dialog state, since they believed that rejection thresholds may vary along dialog states. Then, a logistic regression optimization was trained for task success and a Poisson optimization model was created for dialog duration.

Sarikaya et al in [6] state that for domain constraint systems, semantic information can be helpful. Two different features were used in the studies, the first relying only on the parsing result is based on the intuition that a grammatically correct sentence is easier to parse. The second uses the language model posterior to incorporate information from the recognition process into the confidence measure calculation.

Litman and Pan [7] designed strategies for adaptive behavior in an SDS. The system should change its behavior according to the confidence measure computed by the system. A corpus was labeled with semantic accuracy, that is the percentage of concepts affected by speech recognition errors in each user turn. A threshold was set to classify a dialog as good or bad according to the ASR performance. Together with this, a set of 23 features divided into five different categories was used: acoustic confidence, dialog efficiency and quality, experimental parameters and lexical. Since the used thresholds often requires tuning and the system may recover from an ASR error by adding context information when binding the ASR output to the concepts. In order to deal with this, a new feature was computed that tries to predict the percentage of misrecognitions. This feature uses the log-likelihood score from the ASR to predict that a turn is going to be misrecognized. If the log-likehood falls below a threshold, the turn is predicted to be a misrecognition.

Raymond et al. [8] tried to improve belief confirmation using confidence measures. They combine linguistic information, using a ratio between the intersection of the observed trigrams in the training corpus and all the trigrams occuring in a determined utterance, with the acoustic information given by the log-likelihood score given by the speech recognizer.

3 Data

The data set used for this study was collected from a set of tests done with an SDS that gives bus schedule information which is described in more detail in

[4]. The study was carried out for two weeks where the users were supposed to complete three different usage scenarios in each week. The system received 256 calls during that period of time. This corresponded to 1592 user turns. These turns were labeled as correct if the system concepts were correctly acquired from the user turn, or incorrect if they were not. The data set was further divided into a training and a test set. 1144 turns were used for training and the remaining 448 were used for testing.

4 Experimental Procedure

This study aims at selecting a set of features that could help to improve the performance of the Helios confidence annotator, giving more accurate confidence scores to help in the dialog manager decision process. First, the data was analyzed selecting numeric features that could be used to compute the confidence score. The set of features selected included acoustic features that come from the ASR (average word confidence and average word confidence greater than 50% averaged within the turn), dialog performance feature (last turn was marked non-understood) and parsing features (the number of words in the turn and the number of words not covered by the parser).

Since this is a binary problem, logistic regression seemed to be adequate. There are several algorithms to run a logistic regression for feature selection. For the maximum entropy (MaxEnt) method we have used MegaM[9], for prior-weighted Logistic Regression we have used FoCal [10] with 0.5 prior and Stepwise regression functions available in MATLAB's statistical toolbox to compute stepwise regression. Then the results are computed for several rejections thresholds.

The performance of these new weighted feature selections was compared with the baseline system which confidence score was given by:

$$logit(confidence) = 1.6886 - 5.5482 \cdot Ratio\,of\,Uncovered\,Words \quad (1)$$

5 Results

On Table 1 we see the average word confidence is indeed the most weighted feature in every method. Non-understanding in the last turn, the number of uncovered words and the ratio of uncovered words are good predictors for misunderstandood turns. Some of the values are quite unexpected, such as the fact that the average word confidences greater that 0.5 give a negative contribution. In fact, the threshold of 0.5 could mean that it is not adjusted to the new ASR module. The fragment ratio gives an interesting clue: the more words, the less likely the concept is to be correctly bound.

The graphics in this section show Accuracy, Precision and Recall simultaneously, computed with the different algorithms mentioned before, Helios (the baseline), MegaM, FoCal and Stepwise regression computed with MATLAB.

The best performance in terms of accuracy is achieved with the weights computed with MegaM setting the rejection threshold at 0.6. All the methods, except Helios, have their best performance at 0.6 threshold.

	MegaM	FoCal	Stepwise
Intercept	-0.61	-0.94	0.30
Word Confidence	2.95	4.09	0.71
Word Confidence > 0.5	-0.3	-0.97	-0.17
Fragment Ratio	-0.11	-0.36	0.00
Fragment Ratio > mean	0.38	0.57	0.00
Non-understanding in last turn	-0.57	-0.60	-0.10
Ratio of the number of parses	-0.62	-0.98	0.00
Number of uncovered words	-0.88	-1.80	-0.12
Number of uncovered words > 0	-0.38	-0.38	-0.31
Number of uncovered words > 1	-0.71	-0.03	0.00
Normalized number of uncovered words	0.97	2.13	0.00
Ratio of uncovered words	-0.62	-1.34	-0.20
Ratio of uncovered words > mean	-0.38	-0.39	0.00

Table 1. Weights for the features used in the confidence annotator training procedure.

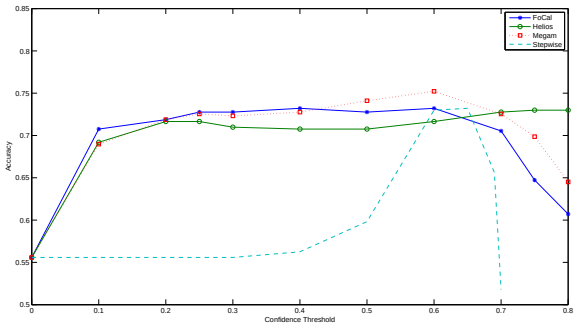

Fig. 1. Accuracy results for the compared methods.

In a dialog system it is very important to minimize the number of misrecognitions. Precision gives an indication of the vulnerability of the system to misrecognitions. Figure 2 shows that FoCal shows clearly the best performance for all the thresholds. Both MegaM and FoCal outperform the Helios baseline.

Although rejections are not as problematic as misunderstandings in SDS, they can reveal if the system is minimizing them. In Figure 3 the Stepwise method has the best performance between 0.4 and 0.6 confidence thresholds. After that, Helios baseline has the best performance.

Among the methods presented the MegaM has a more balanced performance and it has best accuracy. We see that the rejection threshold of 0.6 seems to be the ideal point. The performance of all the methods that include the ASR confidence measure outperform the Helios baseline at this point.

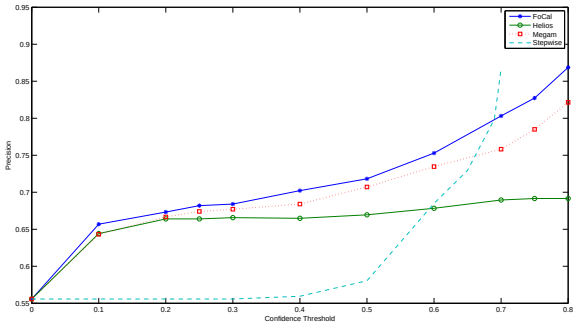

Fig. 2. Precision results for the compared methods.

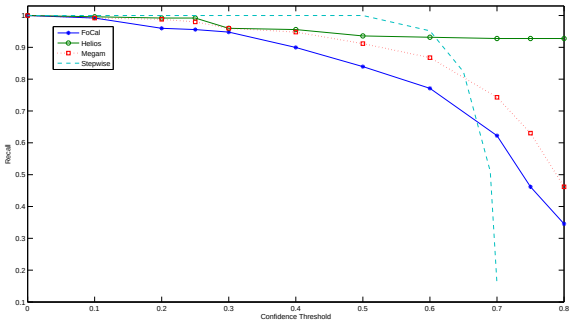

Fig. 3. Recall results for the compared methods.

6 Conclusions and Future Work

This paper describes a study of the improvement of the confidence score in an SDS, namely by including ASR confidence scores. A set of features was selected from the features computed by Helios. Three methods have been compared to baseline Helios confidence annotator: MaxEnt, prior-weighted Logistic Regression and stepwise regression. All the new approaches have outperformed the baseline if 0.6 confidence threshold is considered. The best performance was obtained using MaxEnt to compute the feature weights.

In the future, this study could be extended to other feature selection approaches. Semantic accuracy used in [7] and Word Error Rate could also be used to label the data set, instead of the binary classification that was used. It would also be interesting to see the impact of the results of this study with off-line data in an on-line system.

Acknowledgments This work was partly supported by FCT project CMU-PT/005/2007.

References

1. Bohus,D., Raux, A., Harris, T. K., Eskenazi, M. and Rudnicky, A. I., "Olympus: an open-source framework for conversational spoken language interface research", Bridging the Gap: Academic and Industrial Research in Dialog Technology workshop at HLT/NAACL 2007.
2. J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins and D. Caseiro, "Broadcast News Subtitling System in Portuguese", Proc. ICASSP 2008, Las Vegas.
3. Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., Rudnicky, A.I., "PocketShpinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices", Proc. ICASSP 2006, Toulouse, France.
4. Lopes, J., Eskenazi, M., Trancoso, I., "Towards Choosing Better Primes for Spoken Dialog Systems", Proc. ASRU 2011, Hawaii, USA.
5. Bohus, D., Rudnicky, A. I., "A principled approach for rejection threshold optimization in spoken dialog systems", Proc. Interspeech 2005, Lisbon, Portugal.
6. Sarikaya, R., Gao, Y., Picheny, M. and Erdogan, H.,"Semantic Confidence Measurement for Spoken Dialog Systems", IEEE Trans. on Speech and Audio Processing, Volume 13, July 2005.
7. Litman, D. and Pan, S, "Designing and evaluating an adaptive spoken dialogue system.", User Modeling and User-Adapted Interaction, Volume 12, pages 111-137, 2002.
8. Raymond, C., Estève, Y., Béchet, F., De Mori, R. and Damnati, G., "Belief confirmation in Spoken Dialogue Systems using confidence measures", Proc. ASRU 2003, St. Thomas, US Virgin Islands.
9. Daumé III, Hal., "Notes on CG and LM-BFGS Optimization of Logistic Regression", Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`, August, 2004.
10. Brummer, N., "Focal: Tools for Fusion and Calibration of automatic speaker detection systems", URL:`http://www.dsp.sun.ac.za/nbrummer/focal/`