**Introduction**

- We investigate replacing generic clarification questions in automatic spoken dialog systems with targeted clarification questions.
- We conduct machine learning experiments to determine an optimal feature set for performing localized error detection.
- We experiment with lexical, positional, prosodic, semantic, and syntactic features.

**Current State of Dialog Systems:** Ask generic clarification questions. Use recognizer’s confidence for whole utterance.

**Goal of Localized Error Detection:** Tokenize ASR hypothesis into correctly recognized segment(s) and incorrectly recognized segment(s) based on features derived from the hypotheses.

- Use correctly recognized segments to generate a targeted clarification question.

**Method & Feature Selection**

- For all experiments we use a J48 decision tree classifier boosted with MultiBoostAB method.
- In order to derive optimal feature sets for incorrect utterance and incorrect word detection we perform 10-fold cross validation classification experiments.
- We compare classification results from experiments using a baseline feature set to results from experiments using an expanded feature set.

**Data**

- The DARPA TRANSCAT corpus is comprised of staged conversations between American military personnel and Iraqi interviewees utilizing IraqComm speech-to-speech translation system.

**Conclusion & Future Work**

- We have conducted feature selection experiments to find optimal feature sets to train classifiers for utterance and word mis-recognition prediction.
- We find that certain lexical, positional, and syntactic features improve classification results over a baseline feature set containing only ASR posterior score features.

- In future work we will experiment with additional corpora as well as further investigate the construction of reprise clarification questions by conducting much turk experiments.

- We will also experiment with new features derived from the word lattice result of the ASR.

**Table: Feature Selection Results**

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Correct P@F</th>
<th>Incorrect P@F</th>
<th>% Measure over ASR Only</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline utterance feature set</td>
<td>0.84-0.87</td>
<td>0.57-0.60</td>
<td>0.89</td>
<td>94%</td>
</tr>
<tr>
<td>1-stage orig.</td>
<td>0.95-0.96</td>
<td>0.78-0.80</td>
<td>0.92</td>
<td>92%</td>
</tr>
<tr>
<td>2-stage (50% Up Sampling)</td>
<td>0.95-0.96</td>
<td>0.93-0.96</td>
<td>0.95</td>
<td>95%</td>
</tr>
</tbody>
</table>

**Figures:**

- Localization Error Detection Results (Precision, Recall, F-Measure for Correct & Incorrect Recognized Words)
- Utterance Feature Experiment Results (Precision, Recall, F-Measure for Correct & Incorrectly Recognized Utterances)
- Word Feature Experiment Results (Precision, Recall, F-Measure for Correct & Incorrectly Recognized Words)