Localized Error Detection Experiments Using Features Derived from ASR Hypotheses

Eli Pincus, Svetlana Stoyanchev, Julia Hirschberg
elipincus@gmail.com, sstoyanchev@cs.columbia.edu, julia@cs.columbia.edu

Columbia University
Spoken Language Processing Lab

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.

Generic Clarification Question uses threshold value of ASR confidence:

User: *I want to find a restaurant for XXX.*

Threshold confidence value calculated: #<ThresholdValue

System: *I did not understand that, please repeat.*

Targeted Clarification Question requires accurate localized error detection:

Mis-recong. word

User: *I want to find a restaurant for XXX.*

Correctly recognized context.

System: *You want to find a restaurant for what?*

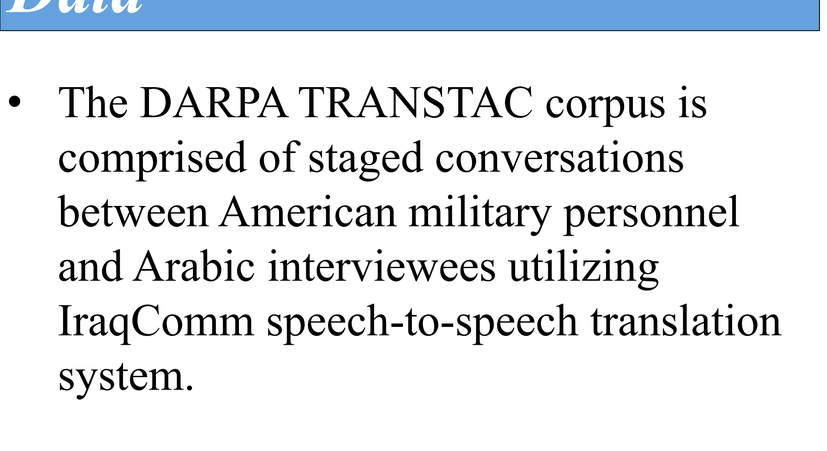
Data

- The DARPA TRANSTAC corpus is comprised of staged conversations between American military personnel and Arabic interviewees utilizing IraqComm speech-to-speech translation system.
- There are 3,952 total utterances and 25,333 total words in the corpus.

Example of misrecognized utterance in corpus:
Reference: do you know BAD people here

ASR Hypothesis: ***d***o you know THAT people here

Possible Reprise Clarification Question: do I know which people here?



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.
- For utterance experiments, all utterances from the corpus are used. For word experiments, only words from incorrect utterances are used.

Baseline Utterance Feature Set
- Average ASR confidence score for all words in utterance

Baseline Word Feature Set
- ASR confidence score for current word

Optimal Feature Set for Utterance Misrecognition Prediction
- Avg ASR conf score for all words in utt
- Average word-length in utterance
- Utterance length in words
- Utterance location within corpus
- POS unigram & bigram count
- Ratio of function words to total words in utterance

Optimal Feature Set for Word Misrecognition Prediction
- ASR conf score for current word
- Avg ASR conf score for current, previous, and next word if present
- Avg ASR conf score for all words in utt
- Word length in letters
- Frequency of longest word in utterance
- Utterance length in words
- Utterance location within corpus
- Word distance from sentence start
- POS tag (curr, prev, next)
- Func/Content tag (curr, prev, next)
- Ratio of func words to total words in utterance

Features Experimented with but not Present in Optimal Sets
- Information associated with minimum-length word In utterance
- Fraction of words in utt with greater length than avg-length word in utt
- Syntactic features such as dependency tag of current word
- Prosodic features such as jitter, shimmer, pitch, and phrase information
- Semantic information obtained from a semantic role labeling of data

Utterance Feature Experiment Results
(Precision, Recall, F-Measure for Correct & Incorrectly Recognized Utts)

Experiment	correct P-R-F	incorrect P-R-F	% F-Measure Incorrect Imp over ASR Only	Accur.
Baseline utt feature set	.893-.930-.911	.678-.571-.620	-	85.5%
Utt optimal feature set	.897-.941-.918	.719-.584-.644	3.9%	86.7%

Word Feature Experiment Results
(Precision, Recall, F-Measure for Correct & Incorrectly Recognized Words)

Experiment	correct P-R-F	incorrect P-R-F	% F-Measure Incorrect Imp over ASR Only	Accur.
Baseline word feature set	.845-.912-.877	.682-.531-.597	-	81.2%
Word optimal feature set	.851-.906-.878	.678-.555-.610	2.2%	83.3%

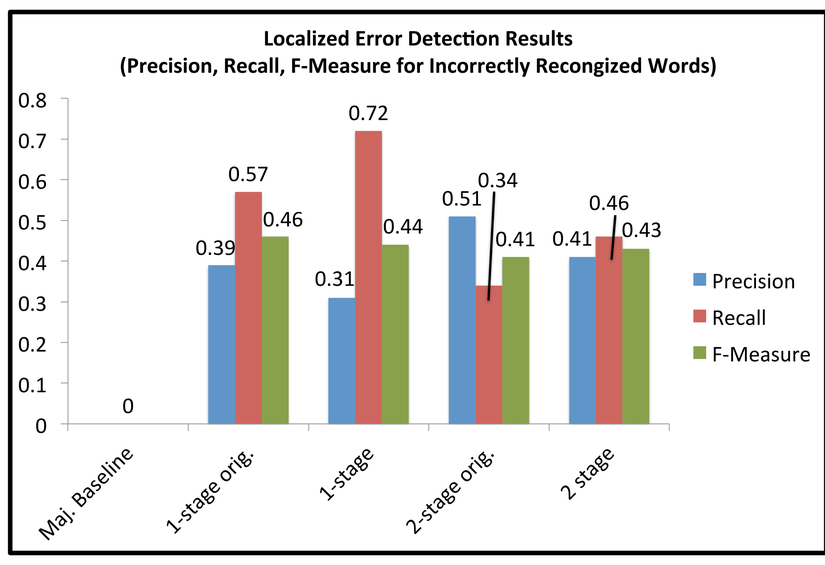
- To simulate actual performance we conduct 1-stage and 2-stage experiments by splitting up the data; 80% training, 20% test.
- For 1-stage Experiments, we classify each word in the corpus.
- For 2-stage experiments we first classify all utterances as correct or incorrect, and then only classify the words in the utterances classified as incorrect.

Experiment Results

- The 2-stage (no up-sampling) approach yields the highest precision for detection of word mis-recognition at 51%.

Localized Error Detection Results
(Precision, Recall, F-Meas. for Correct & Incorr. Recognized Words)

	Correct P-R-F	Incorrect P-R-F	Accuracy
Majority baseline	.94-1.00-.97	_ - 0 - _	94%
1-stage orig.	.97-.94-.96	.39-.57-.46	92%
1-stage (35% Up Sampling)	.98-.90-.94	.31-.72-.44	89%
2-stage orig.	.96-.98-.97	.51-.34-.41	94%
2-stage (35% Up Sampling)	.96-.96-.96	.41-.46-.43	93%



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