Corrections and Repairs

Predicting Corrections in Spoken Dialogue Systems*

Motivation

- Linear relationship between WER and User Satisfaction (Sanders et al, 2002)
- Difficulty of making corrections has more effect on system assessment than actual error rate (Levow)
- Immediate detection followed by some strategy for repair

Summary of Litman et al.

- Understand how user initiate corrections, and level of success
- Hyperarticulation is useful in the automatic detection of corrections
- Corrections are misrecognized more frequently than non-corrections
- System can make better use of existing ASR

SNL Robot Flight Attendants*

SNL Robot Flight Attendants*

ROBOT: "Would you like me to place one blanket in your hand or in your leg?"

PASSENGER: In my hands please.

ROBOT: I'm sorry. I'm having trouble hearing you. Can you speak clearly and loudly into my face?

PASSENGER: In my *hands!*

ROBOT: I'm sorry. One more time.

PASSENGER: "HANDS, WOMAN, HANDS!"

Challenges

- Corrections are misrecognized more frequently than non-corrections
  - Corrections are easier to detect, but more likely to be misunderstood by the system
- As corrections get more distant from the original error, the prosodic differences get more extreme
- Corrections are more likely to exhibit some form of hyperarticulation

Repair

- Run ASR that is tuned for hyperarticulation
- Create prompts that naturally lead to certain types of corrections that are more successful
- Change the system initiative, confirmation strategy
  - “Fall back” to explicit confirmation when producing an error is more likely
- Multiple stages of the SDS can provide useful information (ASR, NLU, DM)

Further Study

- Cost based approaches (Skantze)
  - Data-driven thresholds
  - Derive cost on *principle of least effort*; correlation to user satisfaction
- RavenClaw (Bohus, Rudnicky)
  - Architecture that implements machine learning process
  - System that can tune error handling to domain
- Error prediction (HMIHY)
  - Focus on “prediction”
  - First couple of exchanges could predict problematic dialogs
  - 196 features were used, ASR, NLU, Dialog Manager, Hand-Labelled, Whole-Dialog

Discussion

- Baseline evaluation
- Many types of miscommunication / correction. User may self-correct, NLU, DM.
- Users may modify behavior (exploit system features)