PREDICTING USER REACTIONS TO SYSTEM ERROR

Diane Litman[†], Julia Hirschberg[†] and Marc Swerts[*]

[†]AT&T Research Laboratories, Florham Park (USA)
[*]IPO, Eindhoven (Neth.) and CNTS, Antwerp (Belgium)

Overview

- Introduction
  - Error handling, prosody
  - Aware sites
- Actual research
  - Toot data
  - Descriptive results
  - Prediction results
- Discussion and perspectives

Errors in speech recognition

- Spoken dialogue systems (SDS) are prone to error, mainly because of errors in the automatic speech recognition (ASR) component of these systems
- It is impossible to fully prevent ASR errors, especially when the system has to operate in noisy conditions or when the domain of the system is very large
- We have investigated to what extent prosodic information can be used as a potential resource for error recovery.

Prosody and errors

- **Our previous study showed that prosody is helpful to identify**
  - **turns in which an error occurred (*misrecognitions*)**
  - **turns in which users try to correct an error (*corrections*)**
- **This paper focuses on (*'aware' sites*), i.e., turns where a user first becomes aware of a system error**

User:	**I would like to leave at 10 pm (*misrecognition*)**
System:	**Do you want to leave at 10 am?**
User:	**No (*aware site*)**
System:	**When do you want to leave?**
User:	**10 pm (*correction*)**

Relevance of aware detection

- **If aware sites can be detected with a reasonable accuracy, they are potentially useful for the dialogue manager in an SDS**

  as Backward-looking signals
  that make it clear to the system that something has gone wrong in the preceding context ('go-back' signals)

  as Forward-looking signals
  indicating upcoming corrections or more drastic changes in user behavior, such as complete restarts of the task.

Goals

- **Descriptive analysis:**
  Analyse aware sites to see if they exhibit prosodic features that distinguish them from other user turns
- **Prediction results:**
  Try to predict aware sites with these prosodic features, alone and in combination with other automatically generated features (ASR confidence score, recognized string,...)

TOOT IVR Dialogues

- Collected in 1998 to investigate differences in dialogue strategy wrt initiative (system, user, mixed) and type of confirmation provided (explicit. implicit, none)
- Domain: train information over the phone
- Subjects: 39 summer students
  (16 male - 23 female; 20 native - 19 non-native)
- Task: find train information for 4 scenarios over the phone via speech
- ASR: BLASR
  - 1152 context-dependent subword models trained on telephone speech
  - barge-in
  - different grammars at each dialogue state

Data

- **152 dialogues giving 2328 user turns:**
  - **recognizer was able to produce a recognized string for 2320 of these turns**
  - **1415 of 2328 (61%) had a word error rate of 0**
  - **1618 of 2328 (70%) were semantically correct**
- **Labels (consensus)**

Corrections: **turns which correct information but are not themselves aware sites (13%)**

Awares: **turns which are aware sites but not also corrections (14%)**

Corrawares: **turns which are both awares and corrections (16%)**

Norm: **remaining turns (57%)**

Features

Prosodic (raw and normalized): f0 max, f0 mean, rms maximum, rms mean, duration, preceding pause, tempo, % silence;

ASR: grammar, string, subcases of string (cancel, help,...), confidence score, rejbool;

System Experimental: initiative, confirmation strategy, adaptive, realstrat;

Dialogue context:

position of current turn in the dialogue

features for preceding turn (e.g., preff0max) and for preceding preceding turn (e.g., ppref0max);

the number/percentage of prior turns exhibiting a feature (e.g., priorynstrnum/priorynstrpct);

the mean of the feature's value over all prior turns (e.g., pmnf0max);

Descriptive analysis

- Procedure
  - For each speaker, calculate mean values for each prosodic feature for each of the four turn categories
  - Then, run a one-factor within subjects anova on these speaker means to see whether there is an overall effect of turn category
  - Perform post-hoc t-tests (Bonferonni method) to see which turn categories were significantly different
- Results
  - There is an significant overall effect of category on F0 max, RMS max, RMS mean, duration, tempo and %Silence
  - Post-hoc analyses show that there is at least one prosodic feature that distinguishes each pair of utterance categories

Predicting utterance categories

- Machine learning experiments to automatically induce rule sets for predicting the different utterance categories
- Ripper (Cohen '96) uses greedy search guided by a measure of information gain to produce ordered rule sets for input vectors of feature values and observations of different utterance categories
- We used 10-fold crossvalidation to give results in terms of accuracy (with estimation error), precision and recall
- Feature sets included all automatically available prosodic/acoustic features, experimental condition features, ASR features and features of the dialogue context

4-way classification performance

	Precision (%)	Recall (%)
norm	80.09	89.39
corr	72.86	61.66
aware	61.01	39.79
corraware	61.76	61.72
Accuracy: 74.23% ($\pm$ 0.96%); baseline: 57%		

A confusion martrix shows that the categories aware and corraware are often confused

3-way classification performance

- isaware = original corraware and aware collapsed
- the 3-way classification would be useful in identifying the existence of a prior error, and identifying those turns which simply represent corrections

	Precision (%)	Recall(%)
norm	84.49	87.48
corr	72.07	67.38
isaware	80.52	77.07
Accuracy: 81.14% ($\pm$ 0.83%); baseline: 57%		

2-way classification performance

- isaware = original corraware and aware collapsed
  !isaware = norm and corr collapsed
- the 2-way classification is useful in flagging a prior system problem, even though it fails to target the turn intended to correct the problem

	Precision (%)	Recall (%)
!isaware	91.7	91.6
isaware	80.7	81.1
Accuracy: 87.80% ($\pm$ 0.61%); baseline: 70%		

Rules for Predicting 2 Turn Categories: ISAWARE (T) versus the rest (F)

if (preconf $\leq$ -4.06) $\wedge$ (pretempo $\leq$ 2.65) $\wedge$ (ppau $\geq$ 0.25) **then** T
if (preconf $\leq$ -3.59) $\wedge$ (prerejbool=T) **then** T
if (preconf $\leq$ -2.85) $\wedge$ (predur $\geq$ 1.039) $\wedge$ (tempo2 $\geq$ 1.04) $\wedge$ (preppau $\geq$ 0.57) $\wedge$ (pretempo $\leq$ 2.18) **then** T
if (preconf $\leq$ -3.78) $\wedge$ (pmnsyls $\geq$ 4.04) **then** T
if (preconf $\leq$ -2.75) $\wedge$ (prestr contains the word 'help') **then** T
if (pregram=universal) $\wedge$ (pprewordsstr $\geq$ 2) **then** T
if (preconf $\leq$ -2.60) $\wedge$ (predur $\geq$ 1.04) $\wedge$ (%sil1 $\leq$ 1.06) $\wedge$ (prermsmean $\geq$ 370.65) **then** T
if (pretempo $\leq$ 0.13) **then** T
if (predur $\geq$ 1.27) $\wedge$ (pretempo $\leq$ 2.36) $\wedge$ (prermsmean $\geq$ 245.36) **then** T
if (pretempo $\leq$ 0.80) $\wedge$ (pmntempo $\leq$ 1.75) $\wedge$ (ppretempo2 $\leq$ 1.39) **then** T
then F

Conclusion

- We have found that aware sites have prosodic properties which distinguish them from other turns, such as corrections and normal turns
- Subsequent machine learning experiments show that these aware sites can be classified as such automatically, with a considerable degree of accuracy (best performance: 87.8%).
- Future:
  explore usefulness of predicted aware sites for detecting prior error and for signaling upcoming corrections or restarts