Personalizing Spoken Dialogue Agents

Diane J. Litman
AT&T Labs - Research
Florham Park, NJ 07932
http://www.research.att.com/info/diane

Overview

- Problem
  - real-time spoken dialogue systems have many limitations, e.g., vocabulary, grammar, and application functionality
  - when users do not respect such limitations, performance typically degrades
- Hypothesis
  - performance can be improved by better matching a system's dialogue behaviors to a user's understanding of system limitations
- Results
  - performance can be improved by allowing users to dynamically adapt a system [Litman and Pan, 1999]
  - an approach for automating adaptation is to monitor dialogue level ASR performance [Litman, Walker, and Kearns, submitted]
  - with a simple tutorial, novice users can perform at expert level [Kamm, Litman, and Walker, 1998]

Dialogue with User Initiative, No Confirmation

Toot: Hi, this is AT&T Amtrak schedule system. This is TOOT. How may I help you?

User: I want the trains from New York City to Washington DC on Monday at 9:30 in the evening.

Toot: Do you want me to find the trains from New York City to Washington DC on Monday approximately at 9:30 in the evening now?

User: Yes.

Toot: I am going to get the train schedule for you ...

Another User Initiative, No Confirmation

U1: I'd like to go to Chicago from Baltimore on Saturday at 8 am.
T1: Do you want to leave in the morning or in the evening?
U2: Uh morning.
T2: Sorry, I can't understand you. Please repeat your answer.
U3: In the morning.
T3: Do you want me to find the trains from Baltimore to Chicago on Saturday around 10 morning now?
U4: No.
T4: How may I help you?
U5: I would like to go from Baltimore to Chicago on Saturday at 8 o'clock.
T5: Do you want me to find the trains from Baltimore to DC on Tuesday around 11:45 am now?
U6: No.
T6: How may I help you?
U7: *Change strategy.*
T7: *You are using the user initiative. Which initiative do you want to change to? System –*
U8: *System initiative please.*
T8: *You are using the no confirmation strategy. Which confirmation strategy do you want to change to? No confirmation, implicit confirmation, or explicit confirmation?*
U9: *Implicit please.*
T9: I heard you say go to DC on Tuesday around 11:45 am. Do you want me to find the trains from Baltimore to DC on Tuesday around 11:45 am now?
U10: No.
T10: Which city do you want to go to?
U11: Chicago.
T11: I heard you say go to Chicago. Which city do you want to leave from?

Dialogue w/ System Initiative, Explicit Confirmation

Toot:	Welcome to AT&T Amtrak schedule system. This is TOOT. Which city do you want to go to?
User:	Washington DC.
Toot:	Do you want to go to Washington DC?
User:	*Change strategy.*
	...
	(user changes to user initiative, implicit confirmation)
	...
Toot:	I heard you say go to Washington DC. Which city do you want to leave from?
User:	From New York City on Monday at 9:30 pm.
Toot:	I heard you say leave from New York City on Monday approximately at 9:30 pm. Do you want me to find the trains from New York City to Washington DC on -
User:	Yes.
Toot:	I am going to get the train schedule for you ...

An Empirical Evaluation of an Adaptable Spoken Dialogue System

(joint work with Shimei Pan)

- 4 versions of TOOT
  - adaptable and non-adaptable
  - "System Initiative with Explicit Confirmation" and "User Initiative with No Confirmation"
- 5 subjects per TOOT
- 4 tasks (similar difficulty) per subject
- 80 total dialogues
  - 2633 turns
  - 5.4 hours of speech

Sample Task

- **Scenario Description**
  - Try to find a train going to **Philadelphia** from **New York City** on the **weekend** at **4:00 pm.** If you cannot find an exact match, find the one with the **closest** departure time. Please write down the **exact departure time** of the train you found as well as the **total travel time.** ("weekend" means the train departure date includes either Saturday or Sunday)
- **Hints**
  - If you don't know what to say or what the system is doing, say **Help** to hear a help message.
  - ...

Dialogue Efficiency Measures (automatically logged)

- User Turns
- System Turns
- Elapsed Time

Dialogue Quality Measures (automatically logged, except where noted)

- Speech Recognition Rejections
- Timeouts
- Help Requests
- Barge Ins (manually labeled)
- ASR Rate (manually labeled)

Task Success Measure
(computed from logs and user surveys)

- Please enter the answers you found below.
  - Exact Departure Time:
  - Total Travel Time:

User Satisfaction: Sum of Many Measures (computed from user surveys)

- **TTS Performance**
  - Was the system easy to understand in this conversation?
    - VeryEasy SomewhatEasy NeitherEasyNorDifficult SomewhatDifficult VeryDifficult
- **Speech Recognition Performance**
- **Task Ease**
- **Interaction Pace**
- **User Expertise**
- **System Response**
- **Expected Behavior**
- **Future Use**

Results

- Main effects of adaptability

Measure	Non-Adaptable	Adaptable
Task Success (%)	55.00	*80.00*
User Satisfaction	26.68	*31.60*

- Interaction effects of adaptability and initial dialogue strategy

	Non-Adaptable		Adaptable	
Measure	SystemExplicit	UserNo	SystemExplicit	UserNo
ASR Rate (%)	*88.44*	57.94	82.55	*75.85*
User Expertise	*4.69*	3.01	4.45	*3.85*
Future Use	3.50	1.70	*3.60*	*3.80*

- Performance function derived using PARADISE evaluation framework (accounts for 55% of the variance)

$$UserSat. = .45\mathcal{N}(ASR) + .33\mathcal{N}(Success) - .14\mathcal{N}(Time)$$

- Summary
  - adaptable TOOT generally outperforms non-adaptable TOOT
  - the utility of adaptation depends on the initial dialogue strategies

Detecting Poor ASR at the Dialogue Level: A Machine Learning Approach

(joint work with Marilyn Walker and Michael Kearns)

- Automate the adaptation process, by detecting and reacting to poor speech recognition
  - poor speech recognition is most predictive of performance
  - classifiers for poor performance can be learned from corpora
  - recognition performance can be improved by changing dialogue strategies
- 2 Classifications
  - if (% of misrecognized utterances > threshold) then (dialogue = "bad"); else (dialogue = "good")
  - objective definition, but requires hand-labeling
- 23 Features
  - acoustic, efficiency, quality, lexical, experimental
  - automatically computable from system logs

Dialogue Systems and Experimental Corpora

- ANNIE: An agent for voice dialing, employee directory lookup and messaging [Kamm, Litman, and Walker, 1998]
  - novice, novice + tutorial, and expert users
- ELVIS: An agent for accessing email [Walker et al., 1997, 1998]
  - various initiative and presentation dialogue strategies
  - novice, novice + tutorial users
- TOOT: An agent for accessing online AMTRAK schedules [Litman, Pan, and Walker 1998; Litman and Pan 1999]
  - various presentation, initiative, and confirmation strategies
  - novice users
- 544 dialogues, over 40 hours of speech

An Example ELVIS Dialogue

Elvis: Hi, Elvis here. I've got your mail.
User: Read me my message from Kim about a meeting.
Elvis: In the messages from Kim, there's 1 message about "Meeting Today."
It says "The meeting this morning is at 10:30 in D102. Kim."

An Example ANNIE Dialogue

	Actual Dialogue	*ASR*	*Confidence*
A1:	Annie here.		
U1:	Hello Annie, could you give me the information on a meeting I have tomorrow?	REJECT	-4.84
A2:	What was that?		
U2:	I need to find out uh, a meeting time and place. Kim sent me a message on it.	REJECT	-4.20
A3:	For all options say help me out.		
U3:	Help me out.	REJECT	-6.35
A4:	Annie here.		
U4:	Help.	help	-1.74
A5:	Help. Say call or say send a message, followed by a label or a number. Or say get my messages...		
U5:	Get messages.	get me sips	-3.33
A6:	Speech and Image Processing Lab.		
U6:	Help	help	-1.78
A7:	Help. Say the first and last name ...		
U7:	Kim.	annie	-3.81
	...		

Input Features for Classifying Dialogue Performance

- **Acoustic Features (continuous)**
  - mean confidence, pmisrecs%1, pmisrecs%2, pmisrecs%3, pmisrecs%4
- **Dialogue Efficiency Features (continuous)**
  - elapsed time, system turns, user turns
- **Dialogue Quality Features (continuous)**
  - rejections, timeouts, helps, cancels, bargeins (raw)
  - rejection%, timeout%, help%, cancel%, bargein% (normalized)
- **Experimental Variable Features (symbolic)**
  - system, user, task, condition
- **Lexical Features (text)**
  - ASR text

Class and Feature Encoding for Annie Dialogue

class	bad
mean confidence	-2.7
pmisrecs%1	29
pmisrecs%2	29
pmisrecs%3	0
pmisrecs%4	0
elapsed time	300
system turns	7
user turns	7
rejections	3
timeouts	0
helps	2
cancels	0
bargeins	1
rejection%	43
timeout%	0
help%	29
cancel%	0
bargein%	14
system	annie
user	mike
task	day1
condition	novices without tutorial
ASR text	"REJECT REJECT REJECT help get me sips help annie"

Results

- Best performing RIPPER rules, 25-fold cross-validation

Features Used	Accuracy (SE)
MAJORITY CLASS BASELINE	52%
EFFICIENCY, NORMALIZED QUALITY	69.7 % (1.9)
LEXICAL	72.0 % (1.7)
ACOUSTIC	72.6 % (2.0)
EFFICIENCY, QUALITY, EXPERIMENTAL	73.4 % (1.9)
ALL FEATURES	77.4 % (2.2)

- Summary
  - learned classifiers can significantly outperform the baseline
  - apparent redundancy between "orthogonal" feature sets
  - "inherently" acoustic problems can (best?) be solved using "higher-level" information
  - comparable results for BOOSTEXTER and other classifiers

Efficiency + Quality + Experimental Vars Rules

---

if (cancel% $\geq$ 4) $\wedge$ (system = toot) **then** *bad*
if (system turns $\geq$ 26) $\wedge$ (rejection% $\geq$ 5) **then** *bad*
if (condition = mixed) $\wedge$ (user turns $\geq$ 12) **then** *bad*
if (system = toot) $\wedge$ (user turns $\geq$ 14) **then** *bad*
if (cancels $\geq$ 1) $\wedge$ (timeout% $\geq$ 11) **then** *bad*
if (elapsed time $\leq$ 87 secs) **then** *bad*
default is *good*

Lexical Rules

---

if (ASR text contains `cancel`) **then** *bad*
if (ASR text contains `the`) $\wedge$ (ASR text contains `get`) $\wedge$ (ASR text contains `TIMEOUT`) **then** *bad*
if (ASR text contains `today`) $\wedge$ (ASR text contains `on`) **then** *bad*
if (ASR text contains `the`) $\wedge$ (ASR text contains `p-m`) **then** *bad*
if (ASR text contains `to`) **then** *bad*
if (ASR text contains `help`) $\wedge$ (ASR text contains `the`) $\wedge$ (ASR text contains `read`) **then** *bad*
if (ASR text contains `help`) $\wedge$ (ASR text contains `previous`) **then** *bad*
if (ASR text contains `about`) **then** *bad*
if (ASR text contains `change-strategy`) **then** *bad*
default is *good*

All Features Rules

---

if (mean confidence $\leq$ -2.2) $\wedge$ (pmisrecs%4 $\geq$ 6) **then** *bad*
if (pmisrecs%3 $\geq$ 7) $\wedge$ (ASR text contains `yes`) $\wedge$ (mean confidence $\leq$ -1.9) **then** *bad*
if (cancel% $\geq$ 4) **then** *bad*
if (system turns $\geq$ 29) $\wedge$ (ASR text contains `message`) **then** *bad*
if (elapsed time $\leq$ 90) **then** *bad*
default is *good*

Experimental Variables Rules (65.5% Accuracy)

if (condition = mixed) **then** *bad*
if (system = toot) **then** *bad*
if (condition = novices without tutorial) **then** *bad*
default is *good*

Future Directions

- Generalization of TOOT results to other dialogue strategies
- Incorporation of learned classifiers into TOOT and ELVIS
- Use of additional features, classes and learning methods

Summary

- Builders of spoken dialogue systems face fundamental design choices that strongly influence system performance
- Users can improve system performance by *adapting* the system to the circumstances at hand
- Machine learning is a promising approach for *automating* adaptation, e.g., by detecting poor speech recognition

ANNIE and ELVIS Task Scenarios

- 1.1: You are working at home in the morning and plan to go directly to a meeting when you go into work. Kim said she would send you a message telling you where and when the meeting is. Find out **the Meeting Time** and **the Meeting Place**.
- 1.2: The second task involves finding information in a different message. Yesterday evening, you had told Lee you might want to call him this morning. Lee said he would send you a message telling you where to reach him. Find out **Lee's Phone Number**.

RIPPER Results

Features Used	Accuracy (SE)
MAJORITY CLASS BASELINE	52%
REJECTION%	54.5 % (2.0)
EFFICIENCY	61.0 % (2.2)
EXPERIMENTAL	65.5 % (2.2)
DIALOGUE QUALITY (NORMALIZED)	65.9 % (1.9)
MEAN CONFIDENCE	68.4 % (2.0)
EFFICIENCY, NORMALIZED QUALITY	69.7 % (1.9)
LEXICAL	72.0 % (1.7)
ACOUSTIC	72.6 % (2.0)
EFFICIENCY, QUALITY, EXPERIMENTAL	73.4 % (1.9)
ALL FEATURES	77.4 % (2.2)