Speech recognition and synthesis

Introduction

Successful *Automatic Dialog Systems* must

- Handle numerous different users
- Incite effective user expectations
- Fail gracefully (eg, with human back-up)
- Allow multimodal interaction, if at all possible
- Allow user initiative
- *Automatic Dialog Systems* are as much an *ergonomic* as a *speech* technology problem

Many pictures (and their copyrights) are from [Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

Introduction

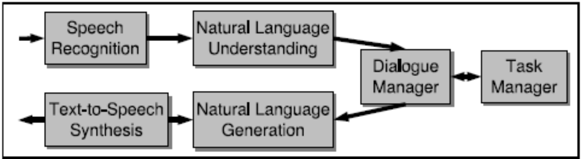


Figure 19.5 Simplified architecture of the components of a conversational agent.

Automatic Dialog Systems have the combined limitations of:

- **ASR + NLP**: The real bottleneck
- **NLG + TTS**: Normally not a problem
- **Dialog management + database**: A bottleneck in complex tasks

[Jurafsky and Martin(2000)]

Conversational Human-Computer Interaction: Practical dialogs

General conversations are much too complex. Limit *Automatic Dialog Systems* to practical dialogues

Dialogues that are focused on a concrete task, eg,

- Task-oriented
- Information seeking
- Advice and tutoring
- Command and control

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

Conversational Human-Computer Interaction

The Practical Dialogue Hypothesis

The conversational competence required for practical dialogues, while still complex, is significantly simpler to achieve than general human conversational competence

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

Conversational Human-Computer Interaction

The Domain-Independence Hypothesis

Within the genre of practical dialogue, the bulk of the complexity in the language interpretation and dialogue management is independent of the task being performed

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

Conversational Human-Computer Interaction

Technique Used	Example Task	Task Complexity	Dialogue Phenomena handled
Finite-state Script	Long-distance dialing	least complex	User answers questions
Frame-based	Getting train arrival and departure information		User asks questions, simple clarifications by system
Sets of Contexts	Travel booking agent		Shifts between predetermined topics
Plan-based Models	Kitchen design consultant		Dynamically generated topic structures, collaborative negotiation subdialogues
Agent-based Models	Disaster relief management	most complex	Different modalities (e.g., planned world and actual world)

Dialogue and task complexity

- Practical Dialogues
- Frame based (form-filling) is currently most used
- Set of frames complex due to switch (going back)
- Plan and Agent based require model-of-the-world

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

Conversational Human-Computer Interaction

Parameter	Possible Values
The train ID?	BN101, ...
The event?	Departure, arrival
The location?	Avon, Bath, Corning, ...
The date/time range?	Monday, Aug 3, afternoon, ...

Context for a train information task

- Frame based dialogue system
- Fill in forms, send query when ready
- Simple and robust
- Simplifies ASR+NLP tasks (pattern matching)

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

Spoken Dialogue Systems

Challenges for Dialogue Systems

- Parsing Language in Practical Dialogues
- Integrating Dialogue and Task Performance
- Intention Recognition
- Mixed Initiative Dialogue

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

Spoken Dialogue Systems: Challenges

Parsing Language in Practical Dialogues

- Detailed semantic, “deep”, representation
- Broad coverage NL grammars fail due to ambiguity
- Semantic restrictions could work
- Add domain-specific restrictions for tasks
- Apply *Grice's Maxims*
- Parsing based on *Speech Acts*

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

Spoken Dialogue Systems: Challenges

Integrating Dialogue and Task Performance

- Complex tasks based on Agents
- Abstract **problem-solving** model:
- **Objectives**: The way we want the world to be
- **Solutions**: Courses of action to achieve objectives
- **Resources**: Objects and abstractions available
- **Situations**: The way the world currently might be

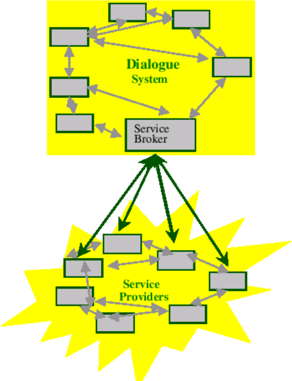


Figure 4: The Agent-based Architecture

Agent based architecture

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

Spoken Dialogue Systems: Challenges

Intention Recognition

- Determine the goal of the user
- Can switch with every utterance
- Use implicatures
- Extrapolate from preceding actions
- Interpolate from "parent" (sub-)goals
- Is a probabilistic framework possible?

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

Spoken Dialogue Systems: Challenges

Mixed Initiative Dialogue

- Finite-state: typically fixed system-initiative
- Frame based: Fixed user/system-initiative (eg, Query-Answer)
- Fixed user-initiative: User does not know what information is needed
- Solution: Limited mixed initiative
- Full mixed-initiative: Both user and system can *barge-in*
- Mixed-initiative: Weight costs of interruption and system goals

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

Spoken Dialogue Systems

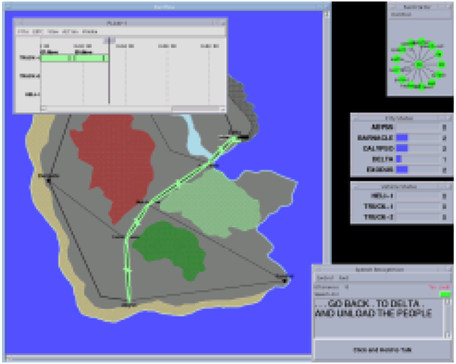
1. *USR: We need to get the woman in Penfield to Strong*
2. *SYS: OK*
3. *USR: What vehicles are available?*
4. *SYS: There are ambulances in Pittsford and Webster*
5. *USR: OK. Use one from Pittsford*
6. *SYS: Do you know that Route 96 is blocked due to construction?*
7. *USR: Oh*
8. *USR: Let's use the interstate instead*
9. *SYS: OK. I'll dispatch the crew*

A short example of a practical dialog

- Both speakers use acknowledgements (*OK*, *Oh*)
- Semantic representation
- Intention recognition
- Goal driven
- System interrupts dialog (barge-in)

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

TRIPS

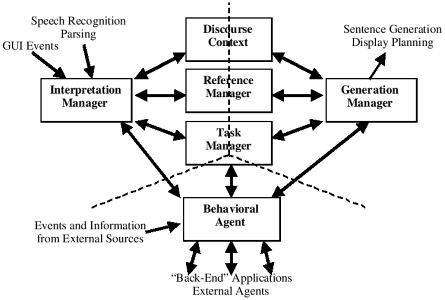


Interacting with TRIPS

- Multi modal interaction with current state shown
- Emergency Response System

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

TRIPS



TRIPS system architecture

- Interpretation
- Generation
- Behavior

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

TRIPS

1. USR: *We* need to get *the woman in Penfield* to *Strong*

Reference resolution

- SS1: The set consisting of *USR* and *SYS* (general dialogue setting)
- WOM1: The Injured woman in Penfield previously discussed (discourse history)
- Strong Memorial Hospital (general world knowledge)

[Allen et al.(2001)Allen, Byron, Dzikovska, Ferguson, Galescu, and Stent]

OVIS

Public Transport Information System

- Deliver train travel information (station-to-station)
- Telephone based application
- Speech only
- Replaced existing human based service
- Based on an existing German system (Philips Aachen)
- Has been in active service (still is)
- Frame-based

[Strik et al.(1997)Strik, Russel, van den Heuvel, Cucchiarini, and Boves]

OVIS

Spoken Dialogue System (SDS) components

1. Continuous HMM based Speech Recognition (CSR)
2. Natural Language Processing (NLP)
3. Dialogue Management (DM)
4. Text-To-Speech (TTS)

OVIS

Skip *Wizard-of-Oz* or *Green-curtain* scenarios and build a working system from scratch.

Stages to build and train SDS

1. Make a first version of the SDS with available data (which need not be application-specific)
2. Ask a limited group of people to use this system, and store the dialogues
3. Use the recorded data (which are application-specific) to improve the SDS
4. Gradually increase the data and the number of users
5. Repeat steps [2], [3], and [4] until the system works satisfactorily

OVIS: Continuous Speech Recognition

Start training with the Polyphone multi-speaker corpus

- 2500 utterances
- Read speech
- Semi-spontaneous (read) speech
- Recorded over the phone
- For each speaker, 5 out 50 *Polyphone* sentences selected
- Phonetically rich sentences (*all* Dutch phonemes)
- 50 Dutch phone models (2 for each of /r/ and /l/)

OVIS: Pronunciation lexicon

Phoneme representations

- Names of stations from the ONOMASTICA database
- Lemma forms of other words from the CELEX database
- Remaining generated by a grapheme-to-phoneme converter
- Pronunciation variation initially not modelled

OVIS: NLP and DM

NLP and DM taken from German original

- Date and time conventions adapted
- Interface with different train table format (eg, start of *tomorrow*)
- Adaptations for user preferences, eg, train numbers
- Collect volunteer queries from keyboard simulation
- Form based database query system with feed-back
- Allows user to correct the system

OVIS: TTS

Speech generation (TTS)

- German original could not be used
- Concatenate utterance fragments
- Female voice

OVIS: Training

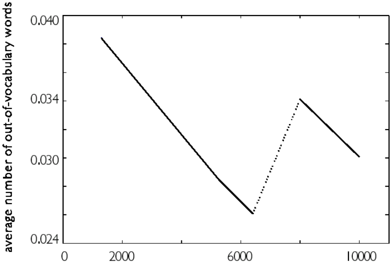
Database	utterances	source	duration (hours:min)
DB0	2500	Polyphone	4:42
DB1	1301	application	0:41
DB2	5496	application	3:47
DB3	6401	application	4:35
DB4	8000	application	5:55
DB5	10003	application	7:20

Databases used during development of the SDS

- Start with the *Polyphone* database (DB0)
- Collect volunteer responses from this system
- Retrain the system with the new speech and repeat
- DB1-5 are incremental, i.e. DB5 contains all of DB4 etc.

[Strik et al.(1997)Strik, Russel, van den Heuvel, Cucchiarini, and Boves]

OVIS: Training



number of utterances in database

Out-of-vocabulary words per utterance vs. corpus size

- Number of OOV words is small
- DB0-DB3 small number of users
- After DB3 (6401 utterances) new users recruited

[Strik et al.(1997)Strik, Russel, van den Heuvel, Cucchiarini, and Boves]

OVIS: Training

System	P0 + L0	P02 + L0	P02 + L2
WG - WER	20.59	18.36	6.72
WG - SER	40.00	36.60	16.00
BS - WER	39.87	31.45	14.73
BS - SER	65.00	54.20	28.00

Performance level for different phoneme models (Pi) and language models (Lj). Evaluation is done with test database 1

- Training phoneme models on both DB0 (polyphone) and DB2 (application) reduced error rates
- Training language model on DB2 (application) reduced errors more
- Application specific data is more important for language modelling than phoneme modelling

[Strik et al.(1997)Strik, Russel, van den Heuvel, Cucchiarini, and Boves]

OVIS: Training

System	P02 + L2	P03 + L2	P03 + L3	P3 + L2	P3 + L3
WG - WER	6.72	6.94	6.94	6.94	6.94
WG - SER	16.00	15.20	15.60	16.20	15.40
BS - WER	14.73	15.43	15.70	16.41	14.84
BS - SER	28.00	29.00	28.60	26.00	26.40

Performance level for different phoneme models (P02/3 vs P3) and language models (L2 vs L3). Evaluation is done with test database 1

- Increasing DB size from 5496 to 6401 utterances had little effect
- Leaving out Polyphone data (DB0) hardly had an effect
- Leaving out DB0 even decreased WER a little

WG: word-graph, BS: best sentence, [Strik et al.(1997)Strik, Russel, van den Heuvel, Cucchiarini, and Boves]

OVIS: Training

testDB	old	new		
System	P3 + L3	P3 + L3	P4 + L4	P5 + L5
WG - WER	6.94	8.87	6.81	6.69
WG - SER	15.40	17.80	14.40	13.80
BS - WER	14.84	15.27	12.93	14.02
BS - SER	26.40	25.40	24.20	24.60

Performance levels for different phoneme models (Pi) and language models (Lj). Evaluation is done with test database 1 (column 2: old) and 2 (columns 3-5: new)

- Test database 2 induced more errors
- DB4 (8,000 utterances) had lower WER again
- Increase to 10,000 utterances (DB5) had little effect

WG: word-graph, BS: best sentence, [Strik et al.(1997)Strik, Russel, van den Heuvel, Cucchiarini, and Boves]

OVIS

Pronunciation variation and non-speech sounds

- A single pronunciation per word gives problems
- Eg, /ɣɛldərɔp/ vs. /ɣɛldrɔp/ and /ɑmsədɑm/ vs. /ɑmstərdɑm/
- Different sources causes inconsistencies
- People use several different variants
- Variant in lexicon not the “best” one

OVIS: Conclusions

It actually worked!

- Adapt an existing frame-based system
- Bootstrap on actual usage
- Collect and train more
- Use robust DM
- Use human fall-back

Further Reading I

James F. Allen, Donna K. Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent.
Toward conversational human-computer interaction.
AI Magazine, Winter, 2001.
URL http://www.cs.rochester.edu/research/cisd/pubs/2001/allen-et-al-aimag2001.pdf.

P. Boersma.
Praat, a system for doing phonetics by computer.
Glot International, 5:341–345, 2001.
URL http://www.Praat.org/.

P. Boersma and D. Weenink.
Praat 4.2: doing phonetics by computer.
Computer program: http://www.Praat.org/, 2004.
URL http://www.Praat.org/.

FSF.
GNU General Public License.
Web, June 1991.
URL http://www.gnu.org/licenses/gpl.html.

Daniel Jurafsky and James H. Martin.
Speech and Language Processing.
Prentice-Hall, 2000.
ISBN 0-13-095069-6.
URL http://www.cs.colorado.edu/~martin/slp.html.
Updates at http://www.cs.colorado.edu/

Further Reading II

Helmer Strik, Albert Russel, Henk van den Heuvel, Catia Cucchiarini, and Lou Boves.
A spoken dialog system for the dutch public transport information service.
Int. Journal of Speech Technology, 2:121–131, 1997.
URL `http://lands.let.ru.nl/literature/strik.1996.4.ps`.
Link is to an older version.

W. Wesseling and R. J. J. H. van Son.
Timing of experimentally elicited minimal responses as quantitative evidence for the use of intonation in projecting TRPs.
In *Proceedings of Interspeech2005*, Lisbon, 2005.

Wieneke Wesseling and R.J.J.H. Van Son.
Early Preparation of Experimentally Elicited Minimal Responses.
In *Proceedings of SIGdial 2005*, September 2005.
URL `http://www.fon.hum.uva.nl/rob/Publications/ArtikelSIGdial2005.pdf`.

Appendix A: Implicatures

Conversations contain rules of inference

Conversational Maxims of Grice

- **Quantity**: Be *exactly* as informative as required
  - Not *less* informative
  - Not *more* informative
- **Quality**: Speak the*truth*
  - Do not say what you believe is *false*
  - Do not say that for which you lack *evidence*
- **Relevance**: Be relevant
- **Manner**: Be *perspicuous*
  - Avoid *obscurity*
  - Avoid *ambiguity*
  - Be *brief*
  - Be *orderly*

Back to Challenges

Copyright License