Using Speech with Computers 2
Spoken Dialog Systems

- **Information giving**
  - Flights, buses, stocks weather
  - Driving directions
  - News

- **Information navigators**
  - Read your mail
  - Search the web
  - Answer questions

- **Provide personalities**
  - Game characters (NPC), toys, robots, chatbots

- **Speech-to-speech translation**
  - Cross-lingual interaction
Dialog Types

- **System initiative**
  - Form-filling paradigm
  - Can switch language models at each turn
  - Can “know” which is likely to be said

- **Mixed initiative**
  - Users can go where they like
  - System or user can lead the discussion

- **Classifying:**
  - Users can say what they like
  - But really only “N” operations possible
  - E.g. AT&T? “How may I help you?”

- **Non-task oriented**
Let’s Go Bus Information

- 412 268 3526
- Provides bus information for Pittsburgh

Tell Me

- Company getting others to build systems
- Stocks, weather, entertainment
- 1 800 555 8355
SDS Architecture

- Recognition
- Interpretation
- Dialog Manager
- Synthesis
- Generation
SDS Components

- **Interpretation**
  - Parsing and Information Extraction
  - *(Ignore politeness and find the departure stop)*

- **Generation**
  - From SQL table output from DB
  - Generate “nice” text to say
Siri-like Assistants

- **Advantages**
  - Hard to type/select things on phone
  - Can use context (location, contacts, calendar)

- **Target common tasks**
  - Calling, sending messages, calendar
  - Fall back on google lookup
“Call John”
“Call John, Bill and Mary and setup a meeting sometime next week about Plan B that’s fits my schedule”
“Make a reservation at a local Chinese restaurant for 4 at 8pm.”
“You should call your mom as it’s her birthday”
“I have sent flowers to your mom as it’s her birthday”
**CALO (DARPA)**

- **Cognitive Assistant that Learns Online**
  - DARPA project (2003-2008)
  - Led by SRI (involved many sites, including CMU)

- **Personal Assistant that Learns (Pal)**
  - Answers questions
  - Learn from experience
  - Take initiative

- **Spin-off company -> SIRI**
  - Acquired by Apple in April 2010
SPDA: Platform

- Desktop
  - Computational power

- Phone (non-smartphone)
  - General Magic
    - Was handheld, became phone based
  - Led into GM’s OnStar

- Smartphone
  - Local to device
  - With Cloud
Smartphone + Cloud

Smartphone

- Know about user
  - Contacts, Schedule etc
  - Same speaker
- Some computation possible on device

Cloud

- Learn from multiple examples
- Retrain acoustic/language/understanding models
Voice Search and User Feedback

- **Voice Search**
  - Google, Bing, Vlingo, Apple

- **Get users to help label the data**
  - Listen to user
  - Show best options
    - They select which one is correct

- **Find out how users actually speak**
  - Full sentences vs “search terms”
  - How do English speakers say ethnic names
Voice Search: Simplifications

- **Too many words** ...
- **Context**
  - Where you are (location: home/not home)
  - What is on your phone (contacts)
  - What you’ve said before
**Personality**

- **Have a character**
  - Calls you by name (you choose)
  - Pushy, helpful, nagging ...
  - Allow user choice

- **Personalize it**
  - May form better relationship with it

- **e.g. Siri**
  - US and UK are female/male
Make it do things well

- **Targeted apps**
  - Chose what it will do well

- **Say, 12 different apps**
  - Have target (hand written) interaction
  - Chose what fields you need, and how to interact with the back end data
  - If all else fails dump result in Google

- **Hardware aid**
  - Infra-red detector for VAD
Make sure people know it’s there

- (Voice search has been on PDA’s for years)
- Get a *lot* of people to use it
- Give “silly” examples
  - People will repeat them, you can adapt your system and expect them to say them
Know Your Users

- Young educated
- Standard English speakers
  - (Non-native too?)
- Can you train them to use it better
  - Get them to adapt
Will it work?

- **Will people talk in public**
  - Talking on the phone is now acceptable
  - Talking to the phone ...

- **Will people continue to use it**
  - Cool at first, but easier to use menus
  - Only use for setting alarms

- **Long term use ...**

- **But others may join in anyway**
Data-Driven Dialog Systems

- Question-Answering systems
- Finding answers from existing conversations
- Dialog vs QA
Neural Conversational Models

- Oriol Vinyals and Quoc Le [2015]
- Sequence to Sequence models
- Predicting reply from previous sentence
From trouble-shooting dataset

```
Machine: hi
Human: hello
Machine: could you please let me know what are the operating systems you are using?
Human: i am using linux
Machine: ok it was in the ssl vpn on which you stated it was broken
Machine: please check if you are logged in to corp network
Machine: what is the error that you are running please
Human: i am seeing an error related to vpn
Machine: what is the error message that you are getting when connecting to vpn using network connect?
Human: connection refused or something like that
Machine: may i know the version of network connect you connect?
```
Dialog Consistency

- Need more than local context (Li et al 16)

<table>
<thead>
<tr>
<th>message</th>
<th>Where do you live now?</th>
</tr>
</thead>
<tbody>
<tr>
<td>response</td>
<td>I live in Los Angeles.</td>
</tr>
<tr>
<td>message</td>
<td>In which city do you live now?</td>
</tr>
<tr>
<td>response</td>
<td>I live in Madrid.</td>
</tr>
<tr>
<td>message</td>
<td>In which country do you live now?</td>
</tr>
<tr>
<td>response</td>
<td>England, you?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>message</th>
<th>Where were you born?</th>
</tr>
</thead>
<tbody>
<tr>
<td>response</td>
<td>I was born in Canada.</td>
</tr>
<tr>
<td>message</td>
<td>Where are you from?</td>
</tr>
<tr>
<td>response</td>
<td>England, you?</td>
</tr>
<tr>
<td>message</td>
<td>Where did you grow up?</td>
</tr>
<tr>
<td>response</td>
<td>I grew up in Texas.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>message</th>
<th>How old are you?</th>
</tr>
</thead>
<tbody>
<tr>
<td>response</td>
<td>16 and you?</td>
</tr>
<tr>
<td>message</td>
<td>What’s your age?</td>
</tr>
<tr>
<td>response</td>
<td>18.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>message</th>
<th>What is your major?</th>
</tr>
</thead>
<tbody>
<tr>
<td>response</td>
<td>I’m majoring in psychology</td>
</tr>
<tr>
<td>message</td>
<td>What did you study in college?</td>
</tr>
<tr>
<td>response</td>
<td>English lit.</td>
</tr>
</tbody>
</table>

Table 1: Inconsistent responses generated by a 4-layer SEQ2SEQ model trained on 25 million Twitter conversation snippets.
Dialog Datasets

- Public datasets
  - Open (Movie) Subtitles
  - TV closed captions
  - Ubuntu forums
  - Reddit (and other forums)
  - Twitter/Weibo

- None are perfect
  - They are large, but not quite right
  - Actually not many long turn-by-turn conversations
Dialog Issues

- Generate dialog quality depends on data
- Need lots (and lots) of examples
  - Target style
  - Content
- Can you “factorize” the context/style
Let’s Talk about nothing …

Xiaobin (Xiaoice)
- Chatbot: Teenage girl persona
- Talk about anything
- Based on Microsoft Forum data (and rules)

But its hard to only talk about nothing
Evaluating non-task Dialog

- Engagement (Zhou 2017)
- Is a user engaged in the conversation
- Can keep them engaged
- When to change subject
- When to re-engage them
Amazon Alexa Prize

- Build a chatbot that engages in 20mins dialogs
- Funding 12 universities (US/Europe)
- Annual(ish) competition
- Systems are multicomponent
  - Combinations of task/non-task
  - Hard written vs statistical/deep learning
- “Let’s Chat”
Amazon Alexa Prize

- But its not about the competition ...
Amazon Alexa Prize

- But its not about the competition …

- Its about engaging researchers
  - Having more PhD students do dialog
  - Giving access for developers to users
  - Collecting data: what do users say
To Task or not to Task ...

- Task Oriented vs non-Task Oriented
- Two extremes
- You need something in the middle
- Let’s Talk:
  - In general
  - But also specifically about X
- Some kind of combinatorial system
Given a wikipedia entry, let’s talk about it
Still open dialog
But has specific information.
E.g. a movie
– Talk about if you’ve seen it
– What you like/dislike
– What other movies actors are in.
Document may be FAQ, product description, general info.
How to Evaluate

- Is it engaging
- Did the bot stay on topic
- (Did the user stay on topic)
- Did the user learn the information
Most languages are not written
- Literacy is often in a different language

Speech isn't the same as writing
- Less standardization (spelling checks)
- Different vocabulary
- Different grammar distribution
- “I'm gonna ….” vs
- “The Prime Minister, Tuesday, said …”
Oral Cultures

- Literacy changes the way you think
  - Literacy is often in a different language

- Oral culture:
  - Stories, proverbs, songs (easier to remember)
  - Rhyming, repetition, alliteration, repetition

- Why should we care?
  - Education, healthcare, (advertising)
  - Everyone can benefit from on-line information
Still part of our oral culture
- Long term spoken verse
- Passed down through the ages
- Rhymes, consistent
- Though sometimes archaic
Nursery Rhymes

Ring-a-ring o' roses,
A pocket full of posies,
A-tishoo! A-tishoo!
We all fall down.
Nursery Rhymes

Ring-a-round the rosie,
A pocket full of posies,
Ashes! Ashes!
We all fall down
Sing a song of sixpence,
A pocket full of rye.
Four and twenty blackbirds,
Baked in a pie.
Oral Rhymes

- **Archaic fixed forms**
  - “four and twenty”
  - “posies”
  - “treacle”
  - “daily bread” (Lord's Prayer)

- **Archaic Grammar**

- **Meaning can be obscure**
Songs, Rhymes, Mnemonics

- **Songs as**
  - Stories (Ballads)
  - Histories
  - Instructions Information
    - Maize planting song
    - Alphabet song
    - Washing hands
Songs, Rhymes, Mnemonics

- **Sayings/Mnemonics**
  - Red sky at night, shepherd's delight
  - Red sky in the morning, shepherd's warning

- **Facts [sic]**
  - In fourteen hundred and ninety-three
  - Columbus sailed the ocean sea

- Richard of York Gave Battle in Vain

- May I have a large container of coffee thank you

- Thirty days hath September ...
Culturally Appropriate

- Tamil planting songs
- Military marching songs
- Japanese Haiku warnings
- Sesame Street Songs
Speech and NLP

- **Same statistical methods**
  - Bayes, n-gram, classification trees

- **NLP in speech**
  - POS tagging (in new languages)
  - Parsing (Syntactic and Prosodic)
  - Information extraction
  - Dialog/Discourse analysis
  - “ASR output” as “noisy” text
Generating Poetry

- Healthcare messages for non-literate
- Appropriate rhyming and cultural references

Emotion ID

- Is this person angry when they are calling us

Singing
Fall Class

Covers

- Speech Recognition, Synthesis, Dialog systems
- Speech ID, evaluation
- Building real systems (ASR, TTS, SDS)
Language Technologies Minor

- 11-721 Grammars and Lexicons
- Plus 3 electives e.g.
  - 11-411 Natural Language Processing
  - 11-492 Speech Processing
  - 11-441 Search Engines and Web Mining
  - Or other LT (Masters) course
- Plus project
  - Often leading to a publication