Chatbots and Spoken Dialog Systems

Luke Willson
LING 575

Background

What is a chatbot?

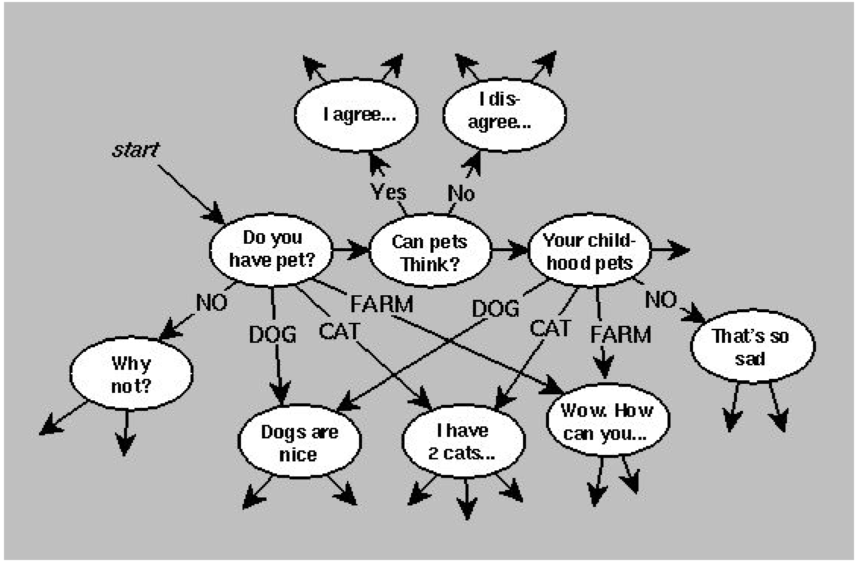
- Industry and Academia have seemingly different definitions of this term
  - Industry:
    - A popular buzzword, but it really is only an SDS application set in a chat-like interface
  - Academia:
    - “Open-domain conversational system that emulates human conversation” (Sordoni et al, 2015)
- Ideally, a chatbot would be able to pass the Turing Test (1950).

Early Chatbots

An Early Attempt...

- Maulding (1994) implemented a “Chatter-Bot” for TINYMUD, a chat-based, online, open-world exploration game
- Designed a series of conversation tools, ordered for priority
  - Command patterns: in-game commands like “exit world”
  - Keyword patterns: e.g. “How do I get from the Town Square to the Library?”
  - Activation Network: triggered a “conversational network” that attempted to carry on a conversation about a given topic
  - Common Sense Answers: e.g. “what’s your name?”, “What’s 2 times 23?”
  - Last-Ditch efforts: e.g. “Go on”, “So?”

...An Early Attempt



- The bulk of the conversation comes from these “domain networks”
- Critiqued as being “just tricks” - restating substrings of user input, controversial statements, safely agreeing with the user, etc.
- My take: this is a really simplistic but effective model

Modern Chatbots

Modern Chatbots...

- More often are based on Neural Nets, not hand-crafted graphs
- Despite the age of NNs, they are only recently being used. Lowe et al. (2015) cite three main reasons:
  - More large & rich datasets (e.g. Ubuntu Dialogue Corpus)
  - Availability to substantial computing power
  - New training methods for neural architectures, specifically for unlabeled data

...Modern Chatbots

- NN’s can process large corpora
  - By using a large corpus like the Ubuntu Dialogue Corpus (1 million multi-turn dialogues, 7 million utterances, 100 million words), NNs can produce human-like responses for a most unstructured dialog (Lowe et al., 2015)
- NN’s can help generate context-sensitive responses
  - We can use Twitter conversations and large corpora to train “neural network architecture [and] address sparsity issues” (Sordoni et al., 2015)
- NN’s can help to generate personalities for chatbots (Li et. al, 2016)

CHATBOTS
Karen Kincy

OVERVIEW

Open-Domain Chat

- “IRIS: A Chat-oriented Dialogue System based on the Vector Space Model” by Banchs and Li (2012)

Errors and Evaluation

- “How NOT to Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation” by Liu et al. (2016)

1

Open-Domain Chat

IRIS: A Chat-Oriented Dialogue System

- Banchs and Li (2012)
- Vector space model
- Dual search strategy
- Able to learn from users on the fly

Dual Search Strategy

- IRIS uses both:
  - Vectorized current user input
  - Vector representation of current dialogue history
- Incorporates dialogue context for returning users
- Cosine similarity metric
  - (cosine similarity is everywhere....)

Domain of Responses

- IRIS is an example-based dialogue system
- Selects candidate responses from corpus:
  - Movie scripts (!)
  - Some noise, like: “Bianca stares at him for a moment”

Ability to Learn from Users

- When IRIS detects an OOV term:
  - Heuristics decide if it should ask user for meaning
  - Tries to avoid “confusing misspellings with OOVs”
  - IRIS asks within the current dialogue

What is “paella”?

#	Spkr.	Turn
1	USER	do you cook anything Iris?
2	IRIS	Eggs.
3	USER	ha ha ha!
4	IRIS	Ha-ha-ha! Ha-ha-ha!
5	USER	Can you cook anything else?
6	IRIS	Do you cook?
7	USER	Paella
8	IRIS	What the hell is that????
9	USER	It is a Spanish food. Yellow rice with some seafood on it.
10	IRIS	Ok, I see... thanks!
11	USER	So, do you want some paella today?
12	IRIS	Well, it beats working in a seafood restaurant...

Table 4: Chat segment in which IRIS learns the new vocabulary term *paella*

How did IRIS respond to “paella”?

- After tokenization, and before vectorization:
  - IRIS replaces “paella” token with the user definition
  - “so do you want some *it is a spanish food yellow rice with some seafood in it* today”
  - Note that IRIS replies to the term “seafood”

Reward or Punishment

- Users can use three special characters to teach IRIS:
  - Ban (*): never show last response
  - Reinforce (+): pull vector representations of last utterance and user turn closer
  - Discourage (-): push last utterance and user turn apart

Issues with IRIS

#	Spkr.	Turn
1	USER	how old are you Iris?
2	IRIS	I'm thirteen, how old are you?
3	USER	I am 40
4	IRIS	Only I did better... 40 crates a day.
5	USER	how old are you Iris?
6	IRIS	Ten.
#	**Spkr.**	**Turn**
1	USER	Yes, have you seen some sports recently?
2	IRIS	I hate sports.
3	USER	Really, what about soccer?
4	IRIS	I'd sooner watch soccer.
5	USER	what do you prefer soccer or football?
6	IRIS	Football is my life.

Table 5: IRIS incapability to being consistent

Criticisms of IRIS

- Movie script domain
  - Repeated use of IRIS could quickly become redundant
  - Movie dialogue only emulates real dialogue
- Cosine similarity just word overlap
  - Why not semantic similarity with embeddings?

2

Errors and Evaluation

How NOT to Evaluate Your Dialogue System

- Liu et al. (2016)
- Correlation between automatic evaluation metrics and human judgements?
- Spoiler: not really...

Dialogue Response Generation Models

- Retrieval Models
  - TF-IDF
  - Dual Encoder
- Generative Models
  - LSTM language model
  - Hierarchical Recurrent Encoder-Decoder (HRED) (Serban et al., 2015)

Evaluation Metrics

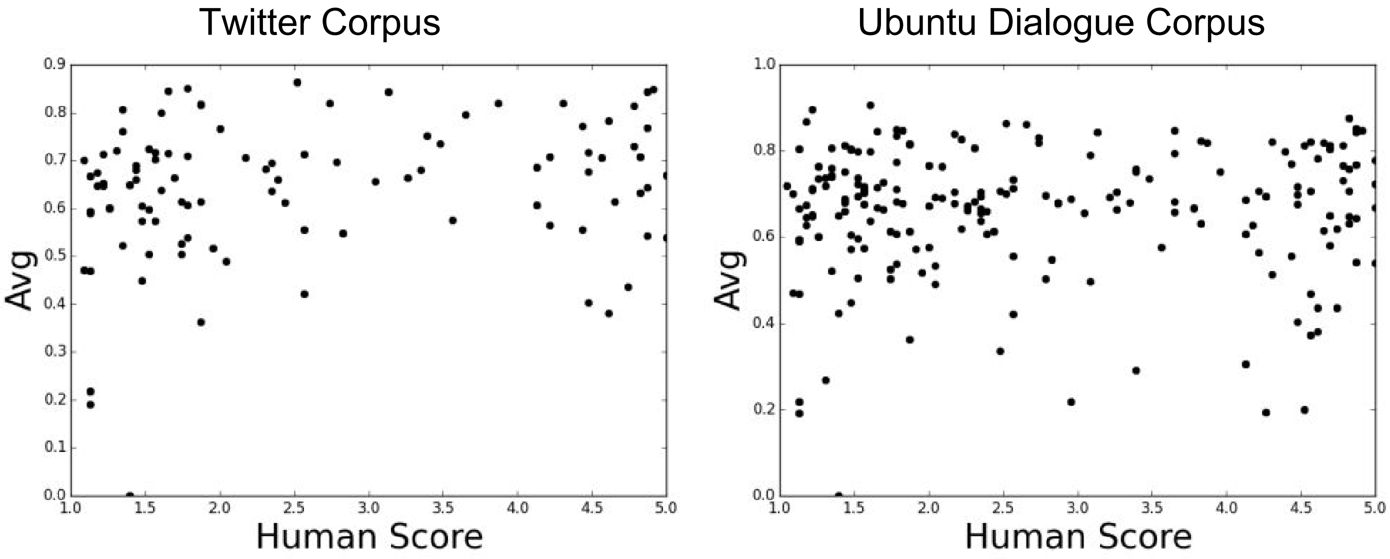
- Word overlap metrics:
  - BLEU
  - METEOR
  - ROUGE
- Embedding metrics:
  - Greedy Matching
  - Embedding Average
  - Vector Extrema

Embedding Average (for example)

- Sentence-level embeddings with additive composition
  - “mean of the word embeddings of each token in a sentence *r*”
  - Seems like a decent distributional semantics metric...

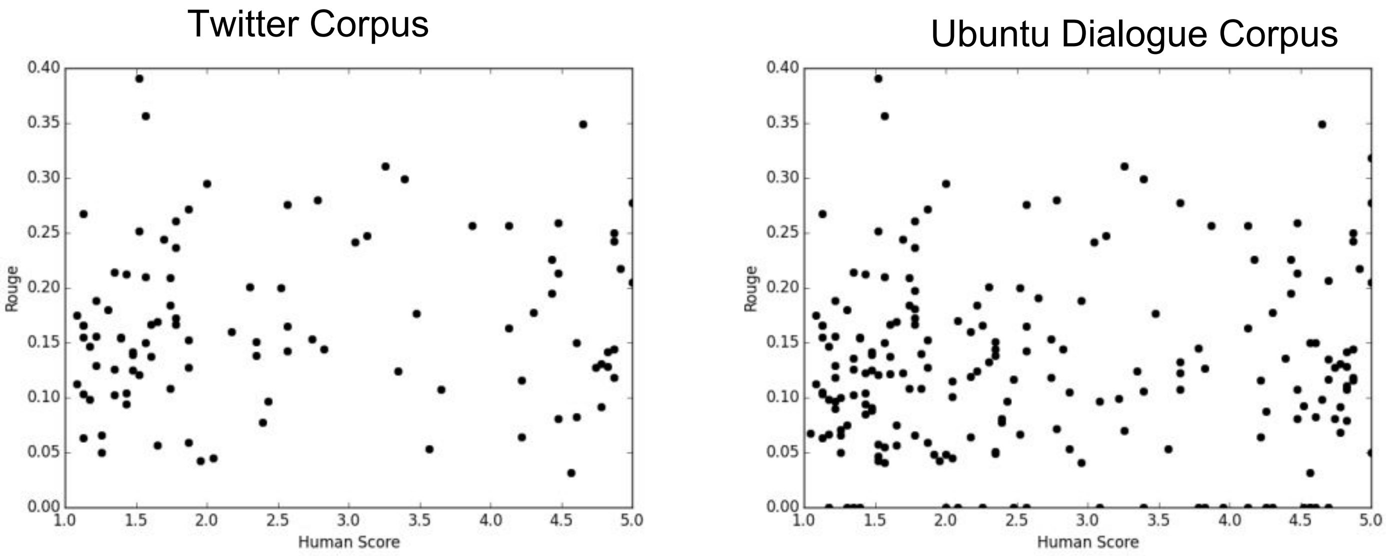
$$\bar{e}_r = \frac{\sum_{w \in r} e_w}{|\sum_{w' \in r} e_{w'}|}.$$

Embedding Average vs. Humans



(c) Vector Averaging

ROUGE vs. Humans



(a) ROUGE

Results

- Liu et al. suggest metrics could take context into account
- Could learn evaluation model from data
  - But learning model might be “no easier than solving the problem of dialogue response generation”
  - Need to always use human evaluations

Criticisms of Liu et al.'s Criticisms

- Who do they use for human judgements?
  - 25 volunteers from CSE department
  - Supposedly decent inter-annotator agreement
  - But likely similar demographics
- Not much of a plan to build a better evaluation metric

3

Discussion

Questions

- IRIS doesn’t “understand” what it says. Does a chatbot need to?
- Why use scores like BLEU and ROUGE beyond system-to-system comparison? (Sorry, LING 573.)
- Is the Turing Test the best way of measuring a successful bot? What other metric could we use? Could that metric take into consideration the "hacks" that Mauldin used?
- What about the ethical & social issues that arise from training bots on inappropriate comments? (e.g. Microsoft's Tay bot)
- Could the unsupervised approach be simplified? Do we need to include the ability to support unexplored domains?

Entrainment in Dialog

Ayushi Aggarwal
LING575 - Spoken Dialog Systems
May 17, 2017

Roadmap

- Introduction to Entrainment
  - Entrainment
  - Why study it?
  - How do we measure it?
  - Previous Work
- Papers – Goal, Dataset and Results
  - Turn taking in HH dialog - Levitan 2015 Paper
  - Cross-linguistic comparison of local and global entrainment - Xia 2014 Paper
- Applications
- Discussion Questions

Roadmap

- Introduction to Entrainment
  - Entrainment
  - Why study it?
  - How do we measure it?
  - Previous Work
- Papers – Goal, Dataset and Results
  - Turn taking in HH dialog - Levitan 2015 Paper
  - Cross-linguistic comparison of local and global entrainment - Xia 2014 Paper
- Applications
- Discussion Questions

Entrainment

- Adaptation or accommodation or alignment of speech

Entrainment

- Adaptation or accommodation or alignment of speech
- Converging features of speech:
  - Syntax
  - Pronunciation
  - Acoustic-prosodic features – intensity, pitch, voice quality
  - Choice of referring expressions/Linguistic style

Entrainment

- Adaptation or accommodation or alignment of speech
- Converging features of speech:
  - Syntax
  - Pronunciation
  - Acoustic-prosodic features – intensity, pitch, voice quality
  - Choice of referring expressions/Linguistic style

- Global(conversation-level) or local(turn-level)

Why study Entrainment?

- Assess HH dialogue success and overall quality
- Evaluate conversational partners
- Cues to model HC/CC interaction
- Other reasons?

Measuring Entrainment

- Similarity over conversation or turn level
- Convergence
  - As behaviour becomes more similar over time
- Synchrony
  - As behaviour varies in tandem

Previous Work

- Entrainment on gesture and facial expression:
  - Strong unintentional entrainment
    - Greater affinity for partner
    - Conversation progressing smoothly

- Entrainment in lexical and syntactic repetitions in first 5 minutes of dialogue:
  - Task success

Papers

- Entrainment:
  - Cross-Linguistic Comparison – Xia et al., 2014
  - And Turn Taking in HH dialog – Levitan et al., 2015

- Applications:
  - Entrainment in pedestrian direction giving - Hu et al., 2014
  - Automated 2-Way Entrainment – Lopes et al., 2013

Papers

- Entrainment:
  - Cross-Linguistic Comparison – Xia et al., 2014
  - And Turn Taking in HH dialog – Levitan et al., 2015

- Applications:
  - Entrainment in pedestrian direction giving - Hu et al., 2014
  - Automated 2-Way Entrainment – Lopes et al., 2013

Cross-Linguistic Comparison

- Prosodic entrainment in Mandarin and English(Xia et al. 2014)
- Compare entrainment in:
  - Pitch
  - Loudness
  - Speaking Rate
- Dataset:
  - Columbia Games Corpus – Standard American English
  - Tongji Games Corpus – Mandarin Chinese
  - Female, male and mixed-gender dialogue pairs

Dataset – I

- Columbia Games Corpus
- 12 dyadic conversations
  - Native speakers of Standard American English(SAE)
  - 6 female and 7 male
    - 3M-M
    - 3F-F
    - 3M-F

- Computer games – cooperate to achieve a mutual goal
- Annotated with prosodic and turn-taking labels

Dataset – II

- Tongji Games Corpus
- 115 dyadic conversations
- University students with a National Mandarin Test Certificate level 2
    - 40 female and 30 male
        - 3M-M
        - 3F-F
        - 3M-F
- Picture ordering and classification games
- Turns identified manually

Acoustic and Prosodic Features

- Extracted from each IPU(Inter-Pausal Unit) using Praat
- SAE pause duration = 50 ms
- MC pause duration = 80 ms

- Features:
  - Intensity – min, mean, max
  - F0 – min, mean, max
  - Speaking rate – syllables/second

SAE v/s MC: Global Entrainment

- Mostly similar entrainment in pitch, intensity and speaking rate for SAE and MC
- SAE and MC speakers entrain globally for duration, pitch and intensity
- Only SAE speakers show global convergence

SAE v/s MC: Local Entrainment

- Entrain in:
  - Similarity of values on intensity and speaking rate
  - Synchrony on intensity and pitch
- Converge on intensity min and all f0 features
- Diverge on pitch

Entrainment and Gender - I

- General theories: Females entrain to a higher degree than men
  - Higher perceptual sensitivity to vocal characteristics

- SAE:
  - F-M pairs entrained on every feature
  - F-M pairs entrained most on intensity mean
  - Intensity max. as greatest for F-M pairs
  - M-M entrain only on intensity mean, max and syllables/second

Entrainment and Gender - II

Table 5: *Evidence of global entrainment by gender group.*

Feature	FF		MM		MF	
	MC	SAE	MC	SAE	MC	SAE
Intensity mean	✓	✓	x	✓	✓	✓
Intensity max	✓	✓	x	✓	✓	✓
Intensity min	x	–	x	–	x	–
F0 mean	x	x	x	x	✓	✓
F0 max	x	x	x	x	✓	✓
F0 min	x	–	x	–	x	–
Speaking rate	✓	✓	✓	✓	✓	✓

Degree of entrainment: M-M< F-M, F-F

Takeaways

- Entrainment is cross-cultural
- Members of different language groups entrain similarly

Future Work and Discussion Points

- How does conversational role affect entrainment behaviour?
- Individual differences in entrainment behaviour – study patterns for features on which speakers entrain or converge locally and globally

Papers

- Entrainment:
  - Cross-Linguistic Comparison – Xia et al., 2014
  - And Turn Taking in HH dialog – Levitan et al., 2015

- Applications:
  - Entrainment in pedestrian direction giving - Hu et al., 2014
  - Automated 2-Way Entrainment – Lopes et al., 2013

Entrainment and Turn-Taking

- HH dialog(Levitan et al., 2015)

- Goals:
  - Speakers entrain on turn-taking behaviours
    - Distribution of turn-types
    - Latency between turns
  - Entrainment at turn exchanges is related to type of turn exchange

Entrainment on Turn Types – I

- Measure similarity between 2 speakers using KL divergence(Kullback and Leibler 1951)

- For speaker *s*:
  - Find similarity to partner
  - Find similarity to non-partner
    - Non-partner := not (s-partner) and getGender(non-partner) == getGender(s-partner)
    - Normalize for gender-specific turn-taking behavior

Entrainment on Turn Types - II

- Baseline - similarity comparison to non-partners
- Turn-type distributions more similar to partner than non-partner. Examples:
  - Interruptions
  - Backchannels
  - Smooth switches

Entrainment on Latency

- Bad conversations:
  - Long latency
  - Frequent negative latency

- Mean latency negatively correlated with entrainment on intensity, pitch, voice quality and speaking rate

Takeaways and Discussion Points

- Interlocutors entrain on similarity of turn-type and latency between turns

Think about -

- Cost-effective SDS, example – low latency between turns
- Improve ASR performance
- System initiated turn-types, such as backchannels for validation(low latency)
- Discover discourse structure through local entrainment

Roadmap

- Introduction to Entrainment
  - Entrainment
  - Why study it?
  - How do we measure it?
  - Previous Work
- Papers – Goal, Dataset and Results
  - Cross-linguistic comparison of local and global entrainment - Xia 2014 Paper
  - Turn taking in HH dialog - Levitan 2015 Paper
- Applications
- Discussion Questions

Applications

- Hu et al., 2014
  - Entrainment in Pedestrian direction giving
  - Different entrainment combinations in a discourse context
    - Some combination preferred – appear more natural/ friendly

- Lopee et al., 2013
  - Lexical entrainment in SDS
  - Adapts to user's lexical choices or suggests better words
  - Reduced estimated error rate by 10% and average #(turns per session) by 6%

Discussion Points - I

- How does conversational role affect entrainment behaviour?
- How can we create cost-effective SDS?

Discussion Points - I

- How does conversational role affect entrainment behaviour?

- How can we:
  - Create cost-effective SDS?
    - Low latency between turns

Discussion Points - I

- How does conversational role affect entrainment behaviour?
- How can we:
  - Create cost-effective SDS?
    - Low latency between turns
  - Improve ASR performance?

Discussion Points - I

- How does conversational role affect entrainment behaviour?
- How can we:
  - Create cost-effective SDS?
    - Low latency between turns
  - Improve ASR performance?
    - System initiated turn-types, such as backchannels for validation(low latency)

Discussion Points - I

- How does conversational role affect entrainment behaviour?
- How can we:
  - Create cost-effective SDS?
    - Low latency between turns
  - Improve ASR performance?
    - System initiated turn-types, such as backchannels for validation(low latency)
- Discover discourse structure through local entrainment

Discussion Points - II

- Should spoken dialog systems be designed that adapt and respond to a user's speech intensity?
  - Would this improve their performance, or would it only result in shouting matches between a human and computer?
- How can we design systems that account for cross-linguistic differences in entrainment trends?

Discussion Points – III

- In this topic, I read a portion of Rivka Levitan's PhD thesis in which she describes a system she used to analyze human trust and liking of systems that entrain to the human's voice. She gets responses from users asking which of the voices they prefer and which of the avatars they trust more. Most users preferred the voice that was entrained to theirs, but that is not true for the trust. I wonder which is more important for Spoken Dialog Systems to truly integrate them into society - like or trust?

- Koda and Maes paper talks of using only female faces for their system. In the Levitan paper, only male voices and faces are used. The paper even indicates a bias for *one of the avatars*, possibly due to his face. It seems that to create a truly interactive system that humans like, trust, and view as human-like, both the look and sound need to be taken into account.

Persona

And Personification

What is it?

Exactly what it sounds like

- Persona in dialog agents is concerned with the perception of personality by users and attempts to evoke a target personality by dialog agents

Why do we care?

- It’s inevitable
- Multiple papers state that users infer personalities in speech agents
  - Even when none was intended
  - Even when users very explicitly knew that there was no rational mind behind the speech agent.

Does Computer-Synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-Attraction, and Consistency-Attraction

Clifford Nass and Kwan Min Lee

- Performed two experiments to ascertain user perception of personality
  - To determine if users detected personalities consistently
    - And if they prefered personalities that matched theirs
  - To assess the effects of a mismatched script and dialog agent

Experiment 1

Clifford Nass and Kwan Min Lee

- Used TTS that had been specifically rated as sounding very artificial.
- Made it sound more extroverted or introverted by modifying parameters
  - Speed, volume, fundamental frequency, frequency range
    - Higher speed, volume, pitch and more variable frequency were associated with extroversion
- Had it read aloud a set of Amazon book reviews

Experiment 1

Clifford Nass and Kwan Min Lee

- Brought in users who had taken a Meyers-Briggs Test
  - Answered a questionnaire on the books and the reviews
    - Described interest in book, credibility of review, and matched adjectives like "enthusiastic" and "shy" to the reviewer
- Users had high rates of agreement on the extraversion or introversion of the reviewing voice
- And frequently preferred the voice that matched the extraversion or introversion results from their Meyers-Briggs Test

Experiment 2

Clifford Nass and Kwan Min Lee

- Had the TTS system read aloud a set of item descriptions that were written to display extraversion or introversion
  - Varied on length, and strength and descriptiveness

- *“This is a reproduction of one of the most famous of the Tiffany stained glass pieces. The colors are absolutely sensational! The first class hand-made copper-foiled stained glass shade is over six and one-half inches in diameter and over five inches tall. I am sure that this gorgeous lamp will accent any environment and bring a classic touch of the past to a stylish present. It is guaranteed to be in excellent condition! I would very highly recommend it.”*

- *“This is a reproduction of a Tiffany stained glass piece. The colors are quite rich. The hand-made copper-foiled stained glass shade is about six and one-half inches in diameter and five inches tall.”*

Experiment 2

Clifford Nass and Kwan Min Lee

- Assigned matched and mismatched voices to descriptions
    - Pairing extroverted voice with introverted text and vice versa
- Users hearing matched voice and text descriptions consistently rated them higher than users hearing mismatched descriptions

Personality Detection

Automatic Recognition of Personality in Conversation

Francois Mairesse and Marilyn Walker

- Automatically determine personality from utterances
- Follow the “Big Five” method of personality categorization
    - Extraversion (sociability, assertiveness)
    - Emotional stability (vs. neuroticism)
    - Agreeableness to other people (friendliness)
    - Conscientiousness (discipline)
    - Intellect (openness to experience)

Automatic Recognition of Personality in Conversation

Francois Mairesse and Marilyn Walker

- Used wearable recorders to record volunteers over two days
- Took random, anonymized chunks of that data and annotated it according to the personality of the recorded person
    - Methodology for the annotation is not particularly clear
- Trained a model on those annotations using a variety of features
    - Prosodic cues, features from psychological sets
        - Unclear what exactly they're getting from LIWC and MRC, maybe lexical?

Results

Feature set	All	LIWC	MRC	Type	Prosody
Set size	117	88	14	4	11
Extraversion	0.35•	0.36•	0.45	0.55	0.26•
Emot. stability	0.40	0.41	0.39•	0.43	0.45
Agreeableness	0.31•	0.32•	0.44	0.45	0.54
Conscientious	0.33•	0.36•	0.41•	0.44	0.55
Intellect	0.38•	0.37•	0.41	0.49	0.44

"•" Denotes an error rate below the random selection baseline error rate of .5

Automatic Recognition of Personality in Conversation

Francois Mairesse and Marilyn Walker

- Personality traits can be clearly detected automatically from utterances
- Some traits aren't indicated well by prosody
    - In order to detect them, may need to get the user to talk about certain subjects

Personality Generation

Trainable Generation of Big-Five Personality Styles through Data-driven Parameter Estimation

Francois Mairesse and Marilyn Walker

- Also uses Big Five classification
- Attempts to train personalities that show clear extremes in those categories

Trainable Generation of Big-Five Personality Styles through Data-driven Parameter Estimation

Francois Mairesse and Marilyn Walker

- Building on a previous paper on the PERSONAGE system
- Used a large set of parameters:

- Content parameters:
    - VERBOSITY Control the number of propositions in the utterance
    - CONTENT POLARITY Control the polarity of the propositions expressed, i.e. referring to negative or positive attributes
- Syntactic template selection parameters:
    - SELF-REFERENCES Control the number of first person pronouns
    - CLAIM COMPLEXITY Control the syntactic complexity (syntactic embedding)
- Aggregation operations:
    - PERIOD Leave two propositions in their own sentences, e.g. ‘Chanpen Thai has great service. It has nice decor.’
    - RELATIVE CLAUSE Aggregate propositions with a relative clause, e.g. ‘Chanpen Thai, which has great service, has nice decor’
- Pragmatic markers:
    - SUBJECT IMPLICITNESS Make the restaurant implicit by moving the attribute to the subject, e.g. ‘the service is great’
    - NEGATION Negate a verb by replacing its modifier by its antonym, e.g. ‘Chanpen Thai doesn’t have bad service’
- Lexical choice parameters:
    - LEXICAL FREQUENCY Control the average frequency of use of each content word, according to BNC frequency counts
    - WORD LENGTH Control the average number of letters of each content word

Trainable Generation of Big-Five Personality Styles through Data-driven Parameter Estimation

Francois Mairesse and Marilyn Walker

- Used those parameters to randomly generate a set of 160 sentences
  - Uniformly distributed across trait extremes
  - Those sentences were then annotated to show their position on trait spectra
- Used those annotations to train personality generating models
  - Tried many different machine learning algorithms and did a lot of feature pruning
  - Models tried to predict the set and weight of features that would produce the target personality

Trainable Generation of Big-Five Personality Styles through Data-driven Parameter Estimation

Francois Mairesse and Marilyn Walker

- Had naive users evaluate the resulting sentences
- Extraversion reports correlated highly with intended extraversion
- As did Emotional stability, agreeableness, and openness to experience
- Conscientiousness correlated negatively
  - They’re not sure why

Discussion

- Are features that indicate personality constant across languages?
- Is the ability to switch personalities for different users desirable or deceptive?
- Are categorie spectra like “Big Five” expressive enough to tailor a personality?
  - What if we wanted to make a dialog agent version of the Geico Gecko or Mickey Mouse? Or a crotchety old man?
- Are there any speech systems that are actively pursuing creating a highly personalized dialog agent?
  - Are any of them male?

Citations

Nass, Clifford ; Min Lee, Kwan. *Does Computer-Synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-Attraction, and Consistency-Attraction*. Journal of Experimental Psychology: Applied 2001, Vol. 7, No. 3, 171-181

Mairesse, Francois ; Walker, Marilyn. *Automatic Recognition of Personality in Conversation.* Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL 2006, pages 85–88

Mairesse, Francois ; Walker, Marilyn. *PERSONAGE: Personality Generation for Dialogue.* Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics 2007, pages 496–503

Spoken Dialog Systems - Context

LING575 – SPRING 2017

Problems of Context

- Co-reference resolution
  - George W. Bush, the forty third president, dubya
    - George and Bush
- Spoken Language Understanding (SLU)
  - just send email to bob about fishing this weekend
    - send email(contact name="bob", subject="fishing this weekend")
- Interaction across multiple domains

Papers

- Bor-shen Lin, Hsin-min Wang, and Lin-shan Lee. A distributed architecture for cooperative spoken dialogue agents with coherent dialogue state and history. In Proceedings of ASRU, 1999.
- Ming Sun, Yun-Nung Chen and Alexander I. Rudnicky. Understanding User's Cross-Domain Intentions in Spoken Dialog Systems, NIPS Workshop on Machine Learning for SLU and Interaction (NIPS-SLU) 2015.

Bor-shen Lin, Hsin-min Wang, and Lin-shan Lee.

A distributed architecture for cooperative spoken dialogue agents with coherent dialogue state and history.

The Problem - How to extend dialogue across multiple domains

- Increase system complexity
- Degrade system performance

The solution

- Distributed Model
- Facilitator to switch domains, base on path distance
  - User Interface Agent (UIA)
  - Spoken Dialogue Agents (SDA)
  - Keeping history
  - Graph search in parallel

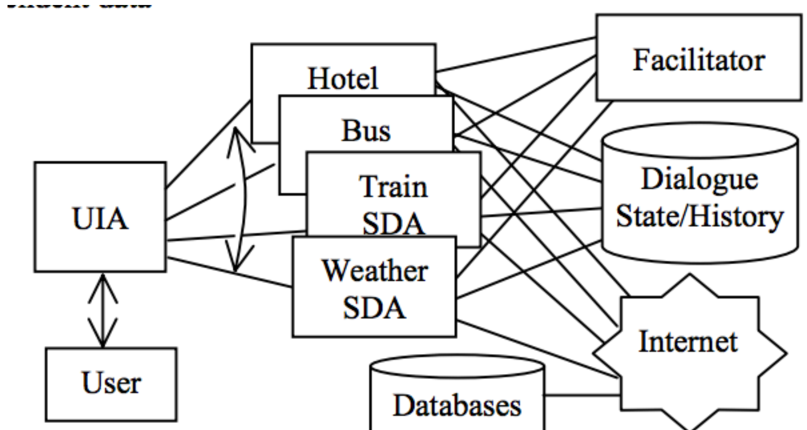


Figure 3. Agent society for spoken dialogue for tour information service

Bor-shen Lin, Hsin-min Wang, and Lin-shan Lee.

S1: Good morning. How may I help you? (bus agent)
U1: Is there any bus from Taiwan University to the train station?
S2: Here is the bus number from Taiwan University to the train station. (display)
U2: I'd like to reserve the ticket from Taipei to Tainan at 2 p.m tomorrow.
S3: Which train type would you like? (train agent)
U3: A moment. How about the weather of Tainan?
S4: The weather of Tainan tomorrow? (weather agent)
U4: Yes.
S5: The weather of Tainan tomorrow is rainy.
U5: And the day after tomorrow?
S6: The weather of Tainan the day after tomorrow is sunny.
U6: OK. I want two tickets for adult.
S7: Do you want to go the day after tomorrow? (train agent)
U7: Yes.
S8: Which train type would you like?
U8: First class.
S9: Do you want the first class at 2 p.m. the day after tomorrow?
U9: How about the ticket price? . . . . .

Figure 6. An example of multi-domain and multi-topic dialogue

Bor-shen Lin, Hsin-min Wang, and Lin-shan Lee.

Paths

- Word Sequence search (WS)
- Tag Sequence search (TS)
- Tag Sequence search and Language Model (TS + LM)
- Pd – detection rate, Pf – failure rate

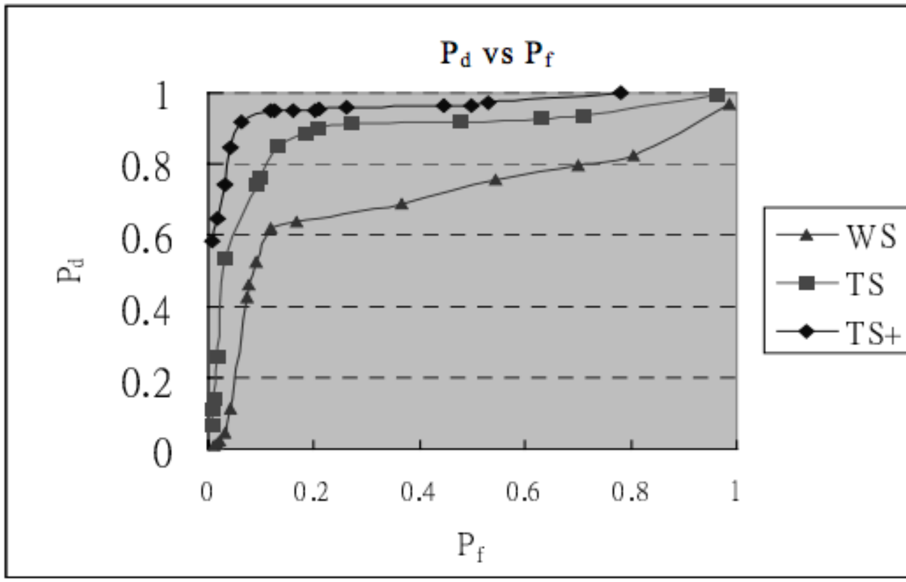


Figure 7. Receiver operator characteristics curves for the detection of domain switching

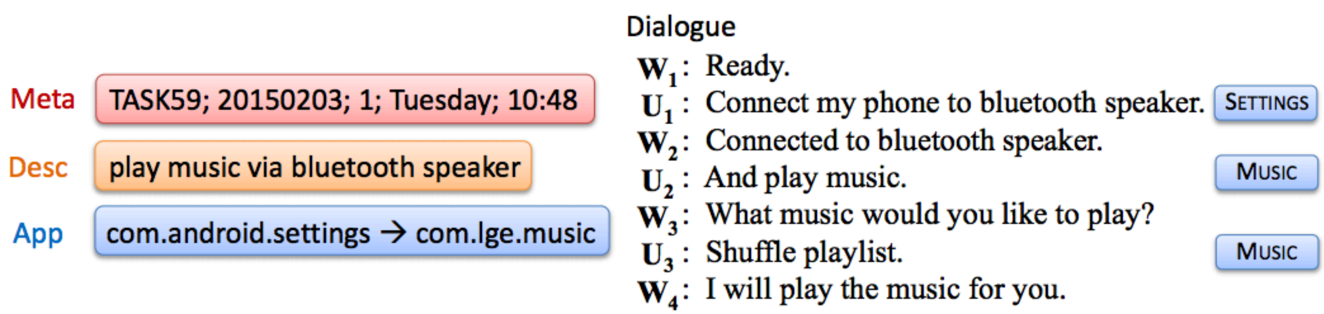
Ming Sun, Yun-Nung Chen and Alexander I. Rudnicky.

Understanding User's Cross-Domain Intentions in Spoken Dialog Systems, NIPS Workshop on Machine Learning for SLU and Interaction (NIPS-SLU) 2015.

- Problem
  - Many applications (Apps) or IAs
  - Each with their independent domain
- Goal to improve interaction quality (predict user intention from dialogue)
  - Forward looking

Ming Sun, Yun-Nung Chen and Alexander I. Rudnicky.

- Manually built Corpus of Smart Phone Usage
  - Track Phone App Usage (14 users)
  - User describe actions
  - With spoken language version of the activities – “reenact” (talk to Wizard-of-Oz to reproduce task in speech)



Ming Sun, Yun-Nung Chen and Alexander I. Rudnicky.

- IA learn to predict
  - The next App
  - User intentions
- Prediction Accuracy (ACC)
- Mean Average Precision (MAP)

Feature	Train		Test	
	ACC	MAP	ACC	MAP
Last App	51.9	60.1	52.9	61.7
Language	44.6	53.6	39.3	50.5
Location	40.3	50.4	32.8	44.7
Time	31.5	42.4	31.5	44.4
Day	29.8	40.9	31.0	43.0
Majority	*27.4*	*38.1*	*31.7*	*44.4*
Meta	48.8	58.2	31.7	43.5
Meta+App	58.7	66.3	58.9	66.0
Lang+App	58.9	66.0	54.2	62.7
All	**64.5**	**71.1**	**58.9**	**66.1**

Table 3: Intention prediction

Discussion Questions

- What is the cost of expanding distributed model of multi-domain system to many domains?
  - How is this different than expanding a single monolithic system with many domains?
- How do people imagine interaction with multi-domain intelligent agent (IA)?
  - Naturally?
  - Wake words?
- What would the ultimate form of intelligent agent be like?