Dialog Management and System Evaluation

EE596B/LING580K -- Conversational Artificial Intelligence

Hao Fang

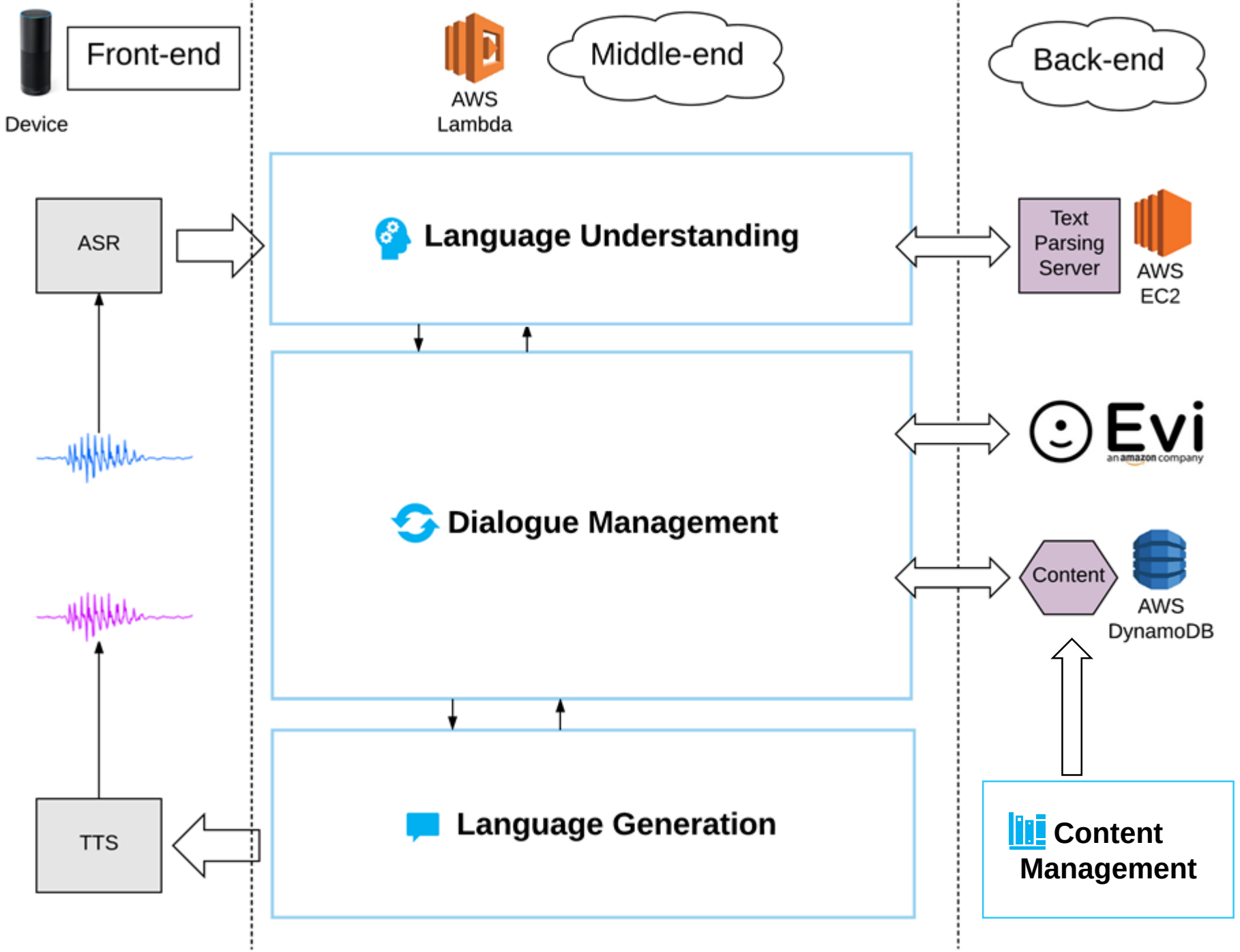
University of Washington

4/17/2018

Slides adapted from:

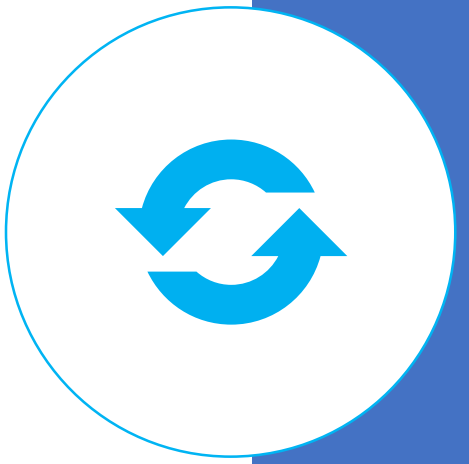
Andrew Maas, Spring 2017, CS224S/LING285 Spoken Language Processing (Lecture 10&11)

Gina-Anne Levow, Spring 2017, LING 575 Spoken Dialog Systems (Lecture 4&5)


Front-end
Device
AWS
Lambda
Middle-end
Back-end
ASR
Language Understanding
Text
Parsing
Server
AWS
EC2
Dialogue Management
Evi
an amazon company
Content
AWS
DynamoDB
TTS
Language Generation
Content
Management

Dialog Manager

- Takes input from ASR/NLU components
- Communicates with backend database & services
- Determines what system does next
- Passes output to NLG/TTS modules

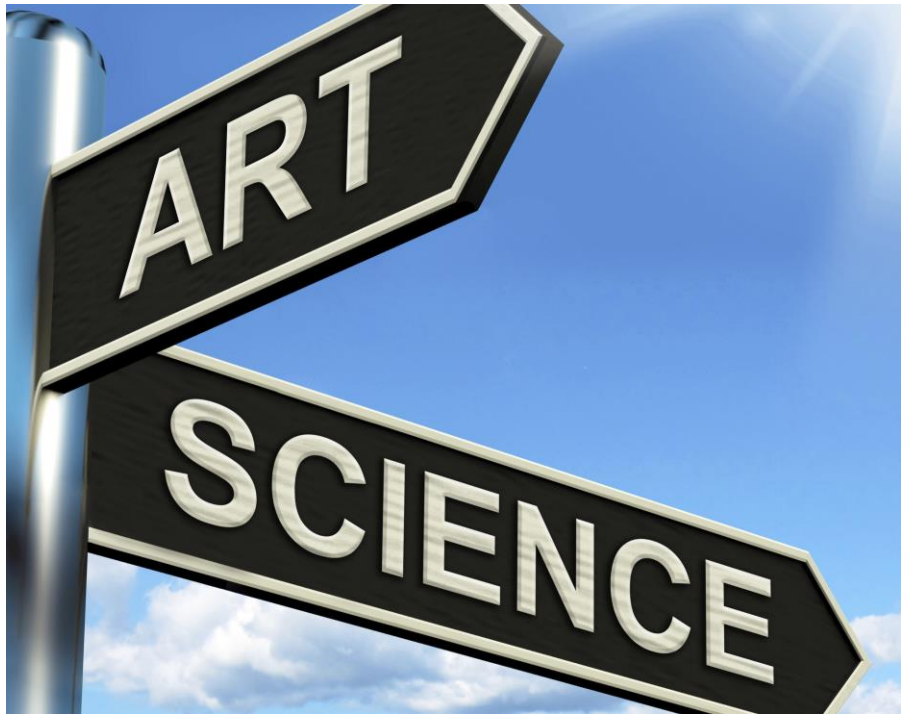


Dialog Policy

Dialog Policy

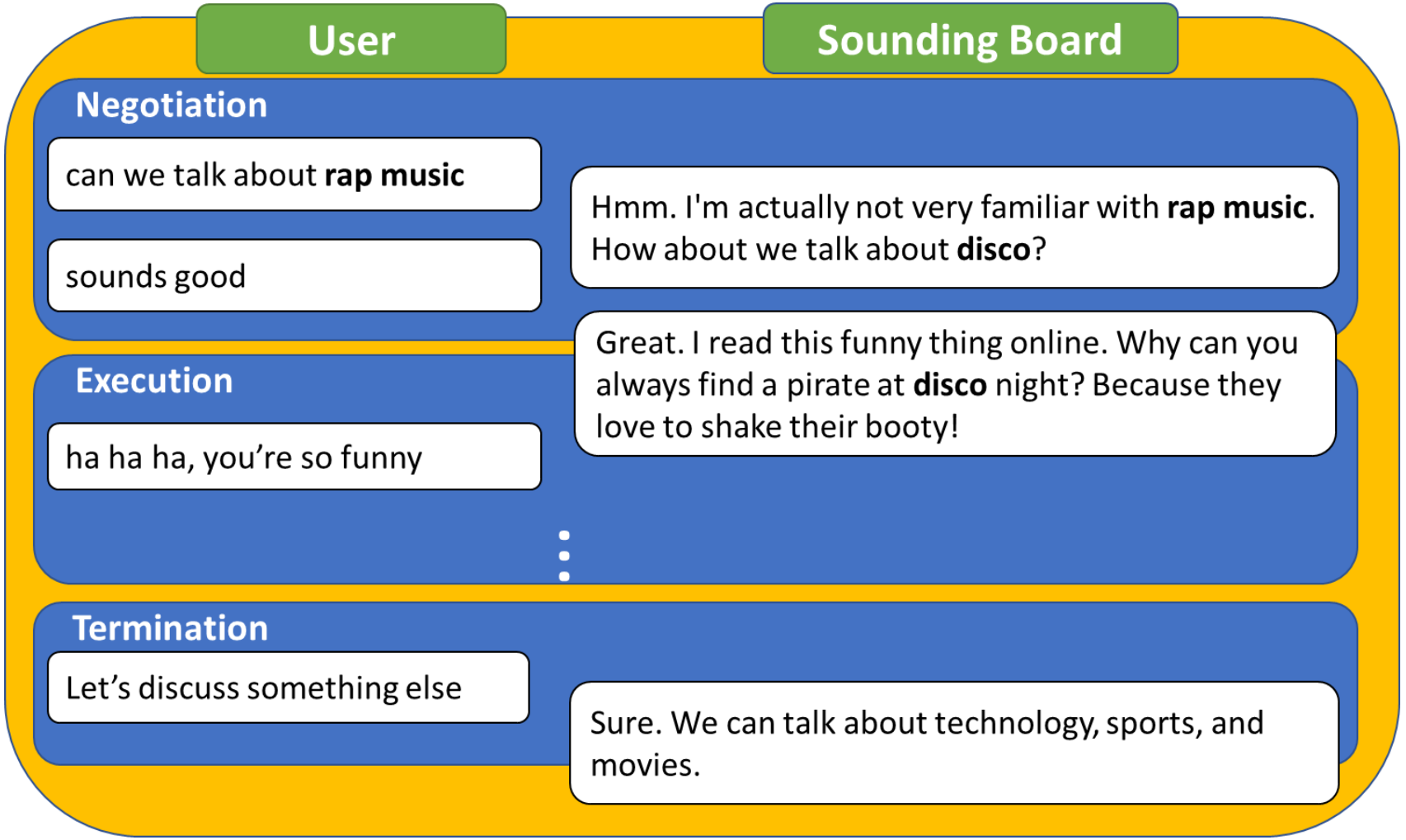
Dialog Policy

- Dialog Structure
- Dialog Initiative
- Conversational Grounding



Turn-taking

- Dialog is characterized by turn-taking.



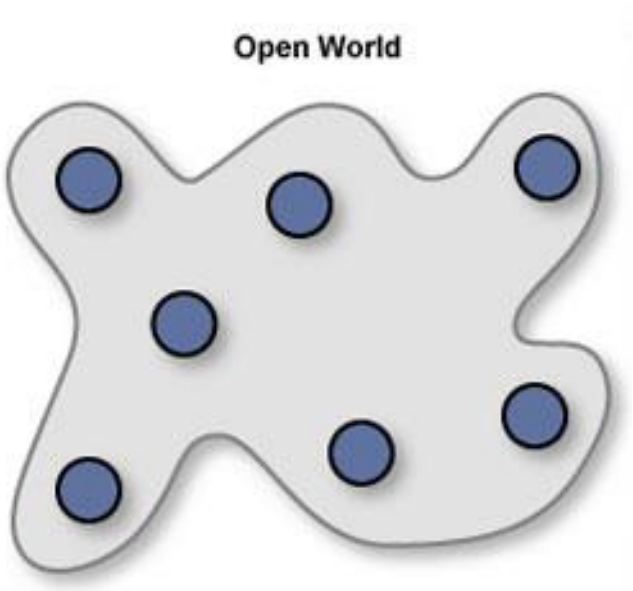
Dialog Structure vs. Storytelling in Games

- Linear storytelling
- A fixed chronological order

Figures from: https://www.gamecareerguide.com/features/882/nonlinear_narrative_in_games_.php?print=1

Dialog Structure vs. Storytelling in Games

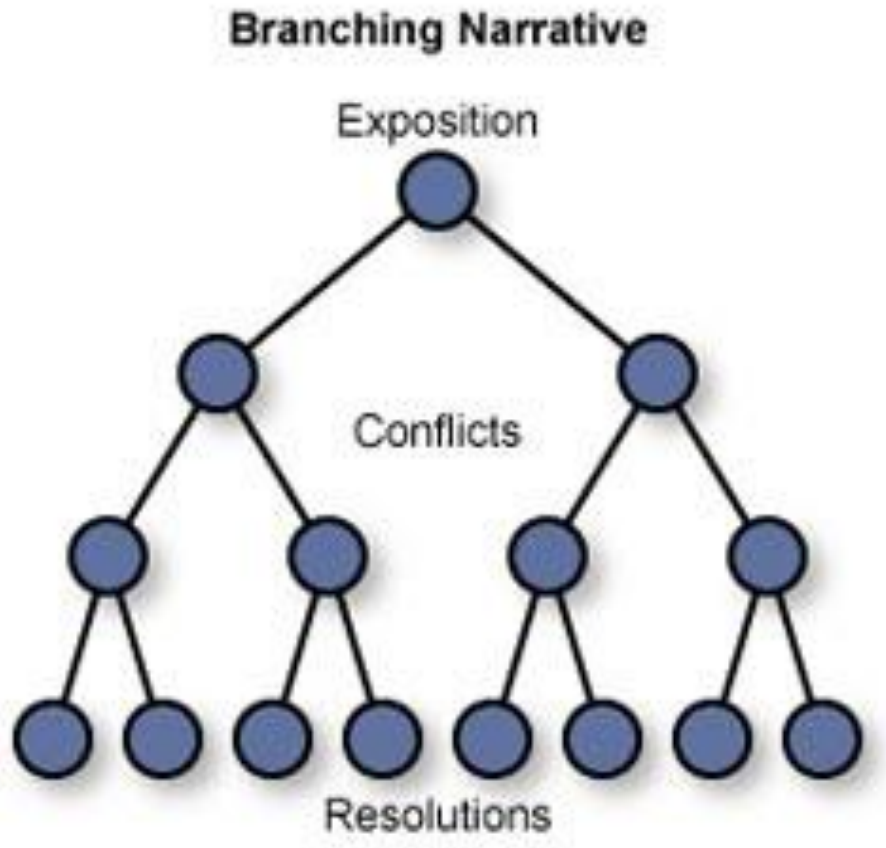
- Nonlinear storytelling
- Explore the world in any order

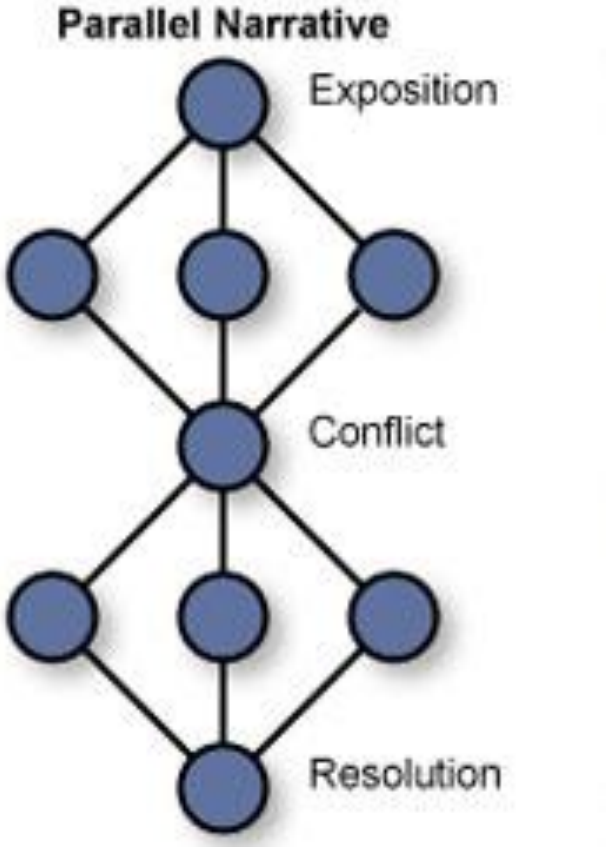


Figures from: https://www.gamecareerguide.com/features/882/nonlinear_narrative_in_games_.php?print=1

Dialog Structure vs. Storytelling in Games

- Other non-linear structures





Figures from: https://www.gamecareerguide.com/features/882/nonlinear_narrative_in_games_.php?print=1

Dialog Structure

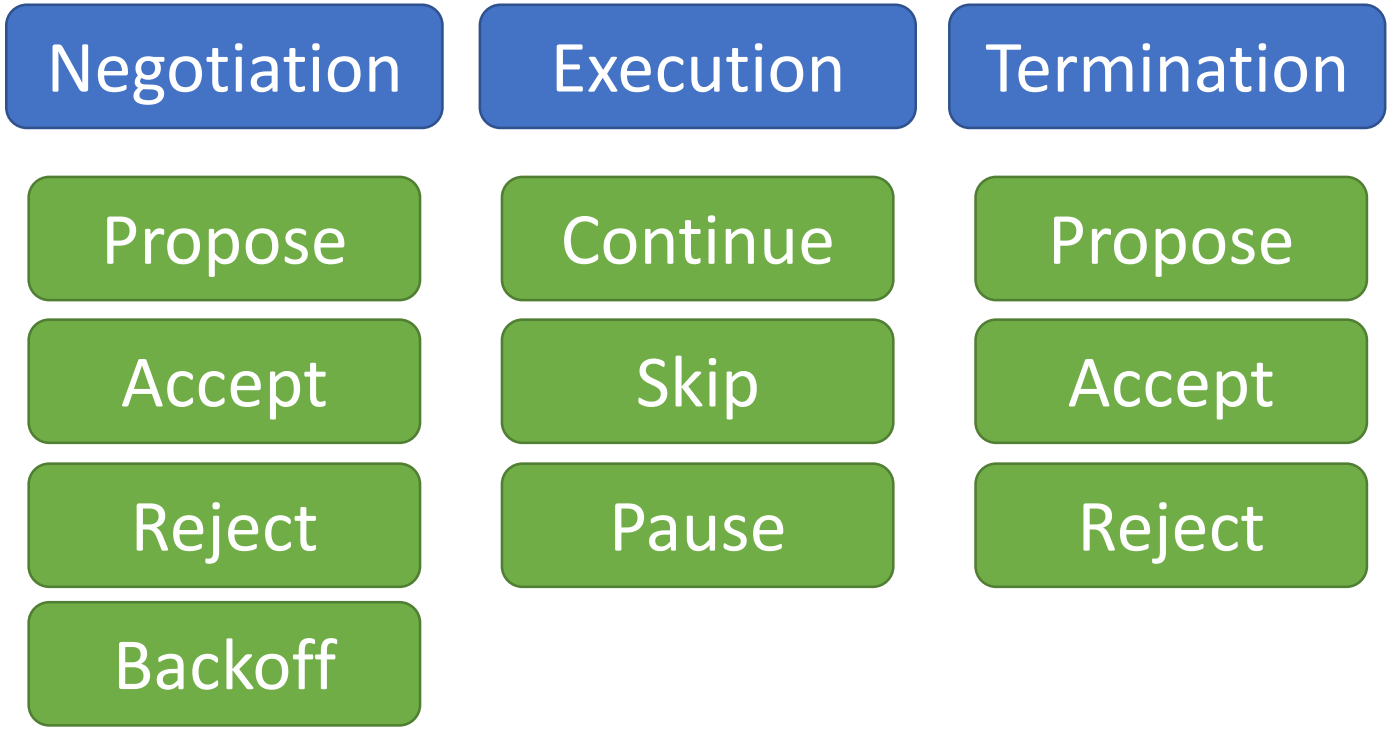
- Three-act structure

Dialog Structure

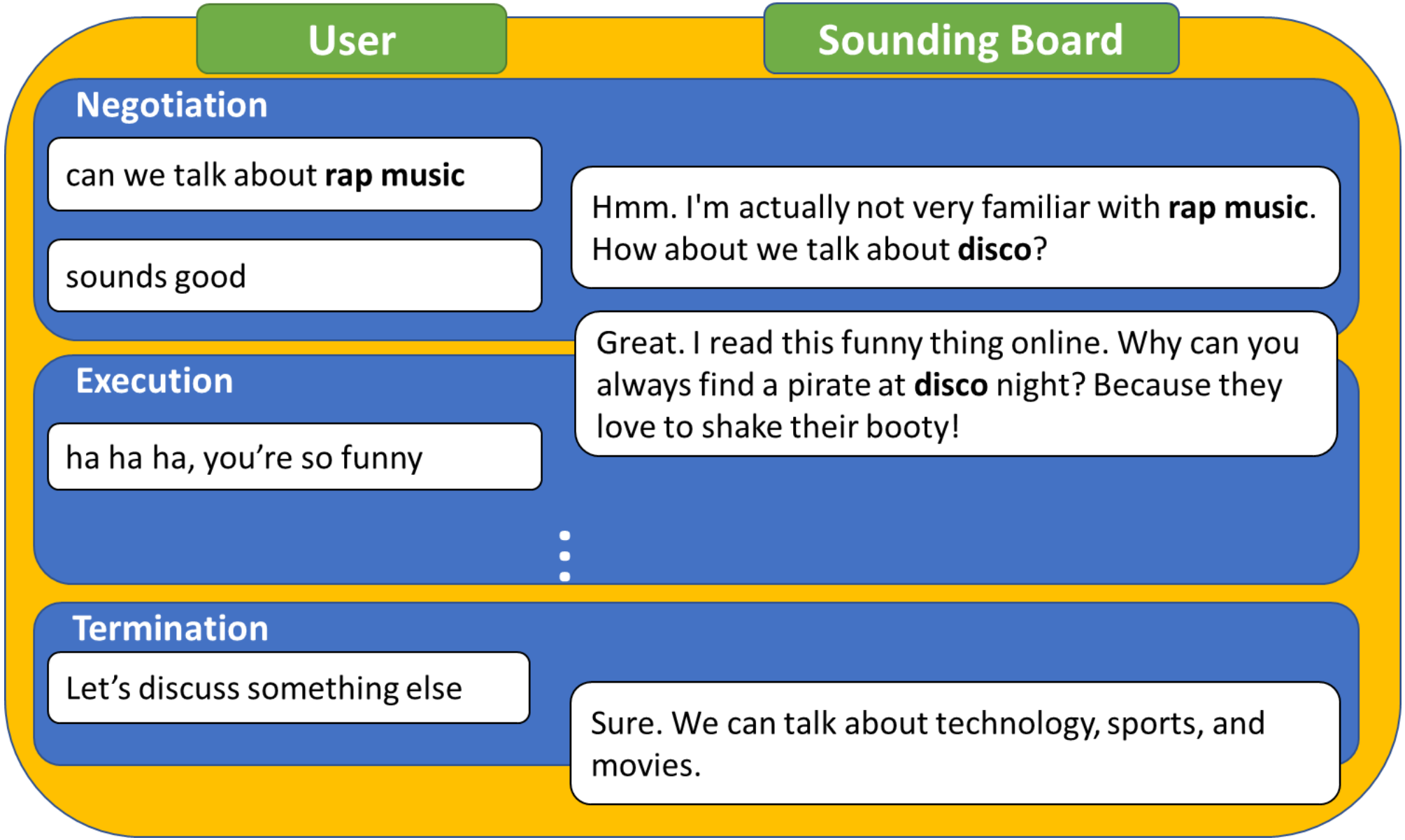
- Three-act structure
- Dialog Macrogame Theory (Mann 2002)
  - http://www-bcf.usc.edu/~billmann/dialogue/dtsite.htm
  - dialog as a sequence of games
  - 6 game acts
  - 15 frequently occurring games

Dialog Structure

- Three-act structure
- Dialog Macrogame Theory (Mann 2002)
- Sounding Board (Fang et al. 2018)
  - social chat as a sequence of sub-dialogs
  - 3 stages
  - 10 coarse-grained actions

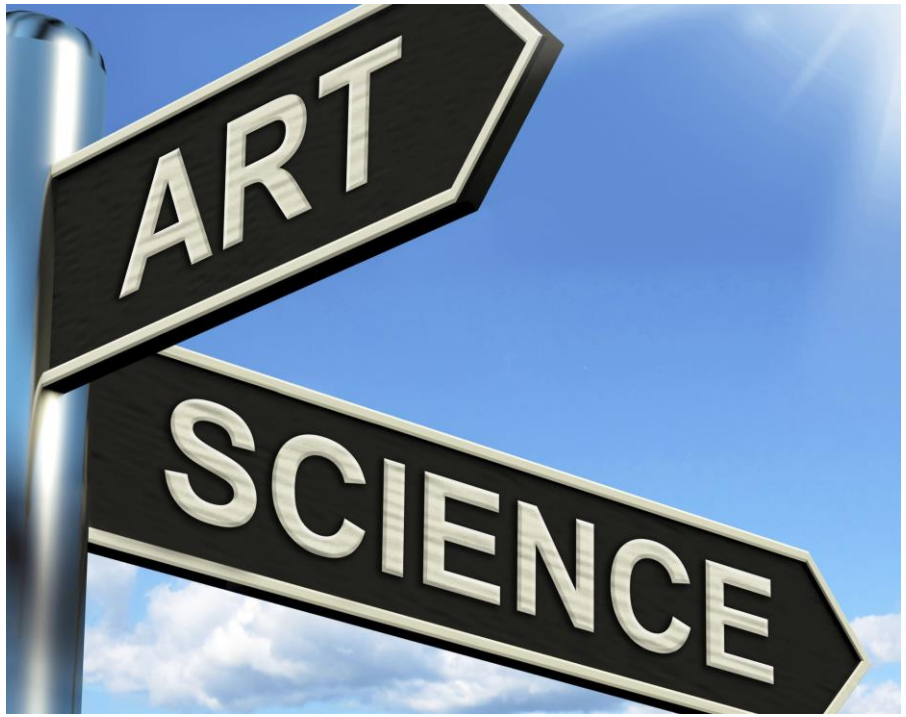


Sub-dialog Cycle



Dialog Policy

- Dialog Structure
- Dialog Initiative
- Conversational Grounding



Dialog Initiative

- **Initiative**: who has control of conversation

System Initiative

- User knows what they can say
- System knows what user can say
- Simple to build
- OK for VERY simple tasks
  - entering a credit card
  - login name and password

User Initiative

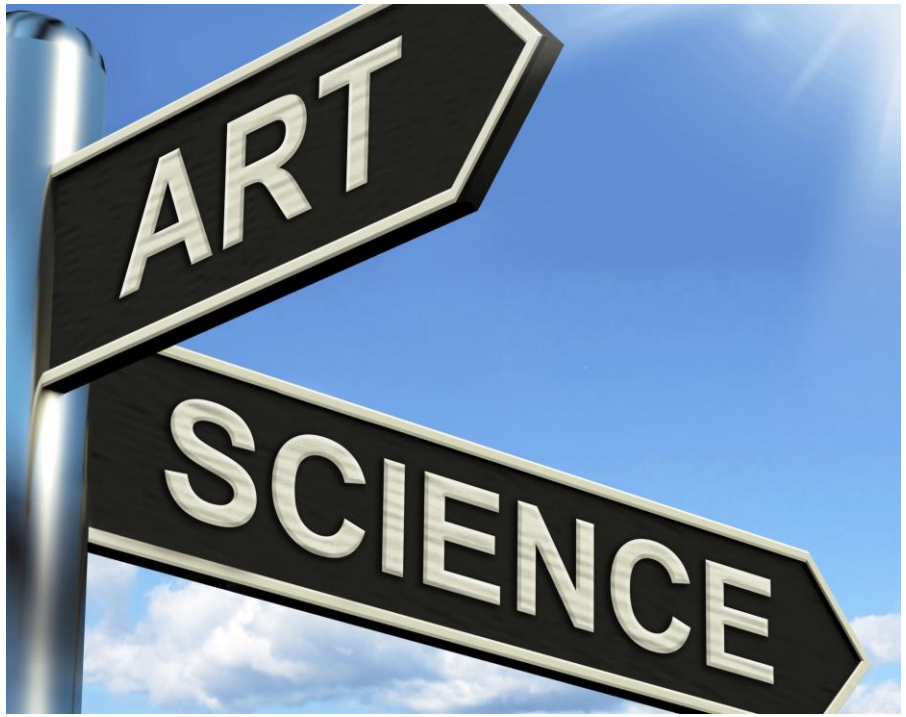
- System is reactive but not proactive
- User knows what system can do
  - question answering
  - voice web search
- System doesn't
  - ask questions back
  - engage in clarification dialog
  - engage in confirmation dialog

Mixed Initiative

- Normal human-human dialog
  - initiative shifts back and forth between participants.
- Mix of control based on prompt type
- Open prompt: *“How may I help you?”*
  - open-ended, user can respond in any way
- Directive prompt: *“Say yes to accept call, or no otherwise”*
  - stipulates user response type

Dialog Policy

- Dialog Structure
- Dialog Initiative
- Conversational Grounding



Conversational Grounding

- Presumed a joint & collaborative communication
  - speaker & hearer mutually believe the same thing
- Speaker tries to establish and add to
  - common ground
  - mutual belief
- Hearer must ground speaker's utterances
  - indicate heard and understood
- **Principle of Closure** (Clark 1996) (Norman 1988)
  - agents performing an action require evidence that they have succeeded in performing it

Principle of Closure

- Non-speech closure example
  - push elevator button
  - light turns on
- Grounding in HCI
  - Users confused if system fails to ground (Stifelman et al., 1993), (Yankelovich et al, 1995)

A Human-Human Conversation

C_1:	. . . I need to travel in May.
A_1:	And, what day in May did you want to travel?
C_2:	OK uh I need to be there for a meeting that's from the 12th to the 15th.
A_2:	And you're flying into what city?
C_3:	Seattle.
A_3:	And what time would you like to leave Pittsburgh?
C_4:	Uh hmm I don't think there's many options for non-stop.
A_4:	Right. There's three non-stops today.
C_5:	What are they?
A_5:	The first one departs PGH at 10:00am arrives Seattle at 12:05 their time. The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
C_6:	OK I'll take the 5ish flight on the night before on the 11th.
A_6:	On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
C_7:	OK.

Sounding Board Conversation

- Indicate ASR/NLU errors
- Acknowledge user reaction

What's your opinion?

I heard you asked: what's your peanut? I'm not sure I know the answer

That's cool!

I'm happy you feel this is cool! Have you read this news? …

That's sad.

I'm sorry to make your sad! Do you want to talk about something else?

Conversational Implicature

- Meaning more than just literal contribution
- Indirect speech acts

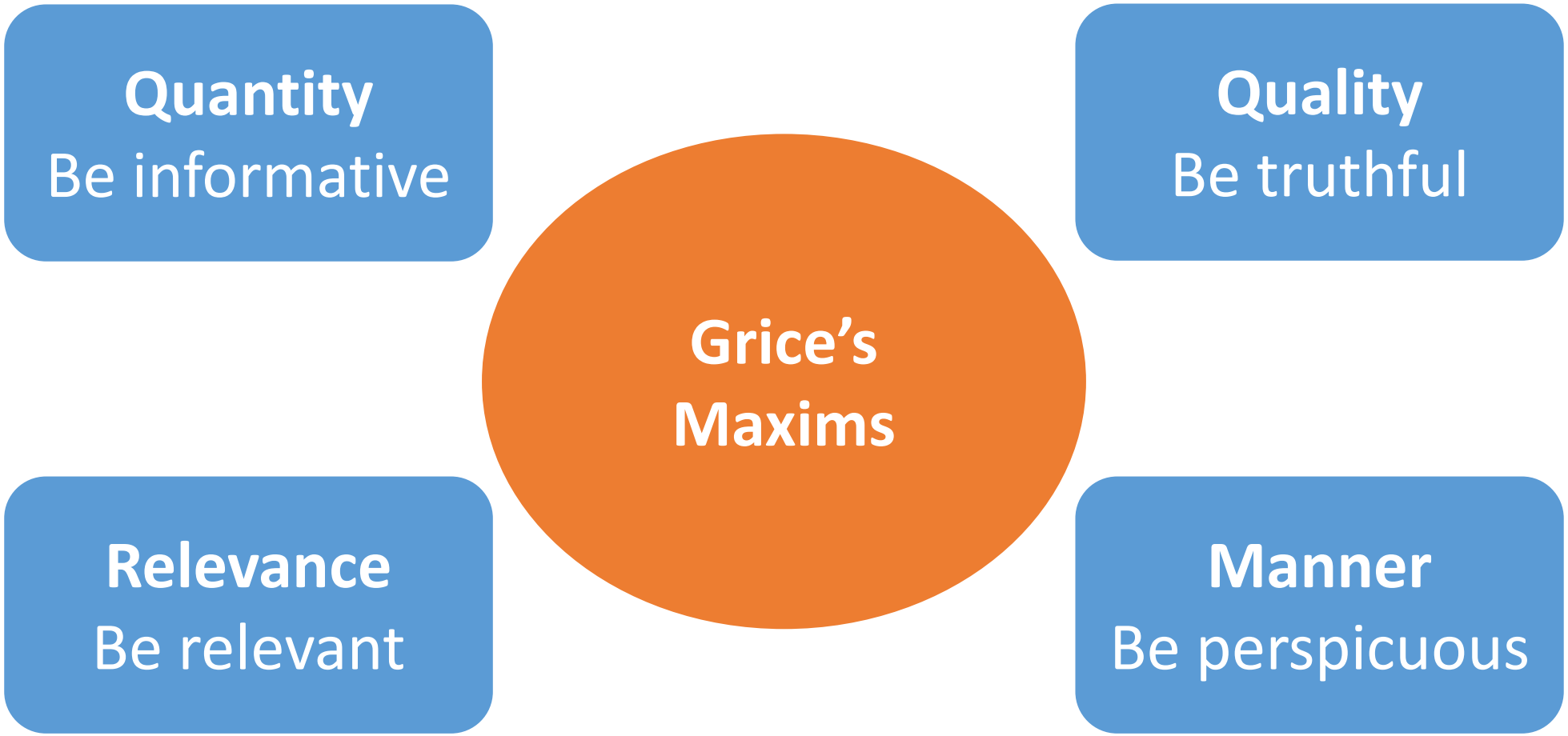
How about we talk about movies?

OK uh I don’t watch movies very often.

Continue

Switch Topic

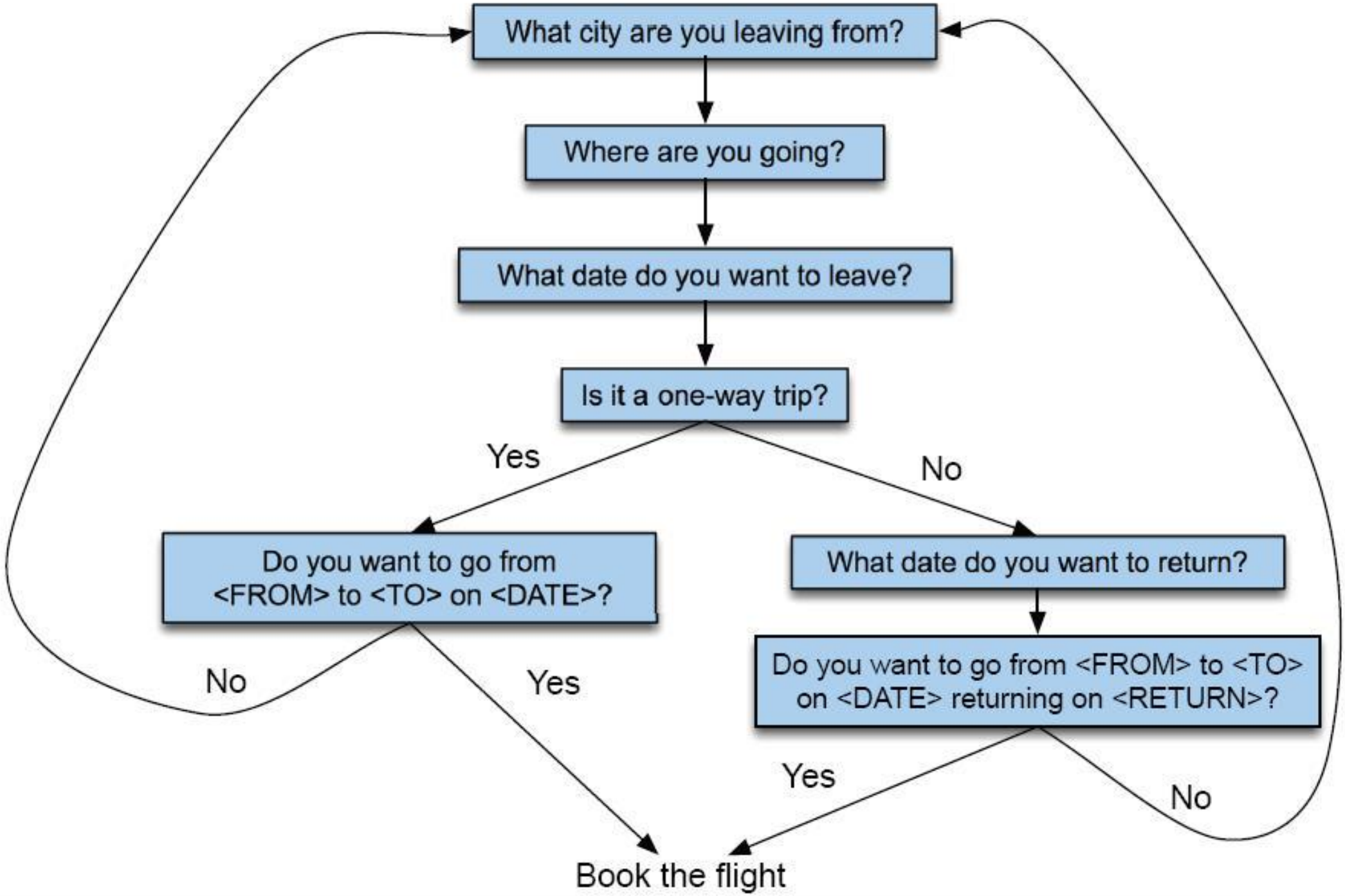
Grice's Maxims



Dialog Manager Architectures

Example: A Trivial Airline Travel System

- Ask the user for a departure city
- Ask for a destination city
- Ask for a time
- Ask whether the trip is round-trip or not


What city are you leaving from?
Where are you going?
What date do you want to leave?
Is it a one-way trip?
Yes
No
Do you want to go from <FROM> to <TO> on <DATE>?
What date do you want to return?
Do you want to go from <FROM> to <TO> on <DATE> returning on <RETURN>?
No
Yes
Yes
No
Book the flight

Finite-state Dialog Manager

- System completely controls the conversation with the user
- It asks the user a series of questions
- Ignores (or misinterprets) anything the user says that is not a direct answer to the system's questions

System Initiative + Universals

- We can give users a little more flexibility by adding **universals**: commands you can say anywhere
- As if we augmented every state of FSA with these
  - **Help** (AMAZON.HelpIntent)
  - **Start Over** (AMAZON.StartOverIntent)
  - **Repeat** (AMAZON.RepeatIntent)
- This describes many implemented systems
- But still doesn’t allow user much flexibility

Finite-state Dialog Manager

Advantages

- Straightforward to encode
- Clear mapping of interaction to model
- Well-suited to simple information access

Disadvantages

- Limited flexibility of interaction
  - constrained input – single item
  - fully system controlled
  - restrictive dialog structure & order
- Ill-suited to complex problem-solving

Frame-based Dialog Manager

FLIGHT FRAME:

ORIGIN:

CITY: Boston

DATE: Tuesday

TIME: morning

DEST:

CITY: San Francisco

AIRLINE:

…

Frame-based Dialog Manager

- Use the structure of the **frame** to guide dialogue

Slot	**Question**
ORIGIN	What city are you leaving from?
DEST	Where are you going?
DEPT DATE	What day would you like to leave?
DEPT TIME	What time would you like to leave?
AIRLINE	What is your preferred airline?

Frame-based Dialog Manager

- Mixed initiative
- User can answer multiple questions at once
- System asks questions of user, filling any slots that user specifies
  - when frame is filled
  - when to query database
- If user answers 3 questions at once, system has to fill slots and not ask these questions again!
  - Avoids strict constraints on order of the finite-state architecture.

Frame-based Dialog Manager

Advantages

- Relatively flexible input & orders
- Well-suited to complex information access
- Supports different types of initiative

Disadvantages

- Ill-suited to more complex problem-solving

Hierarchical Dialog Manager

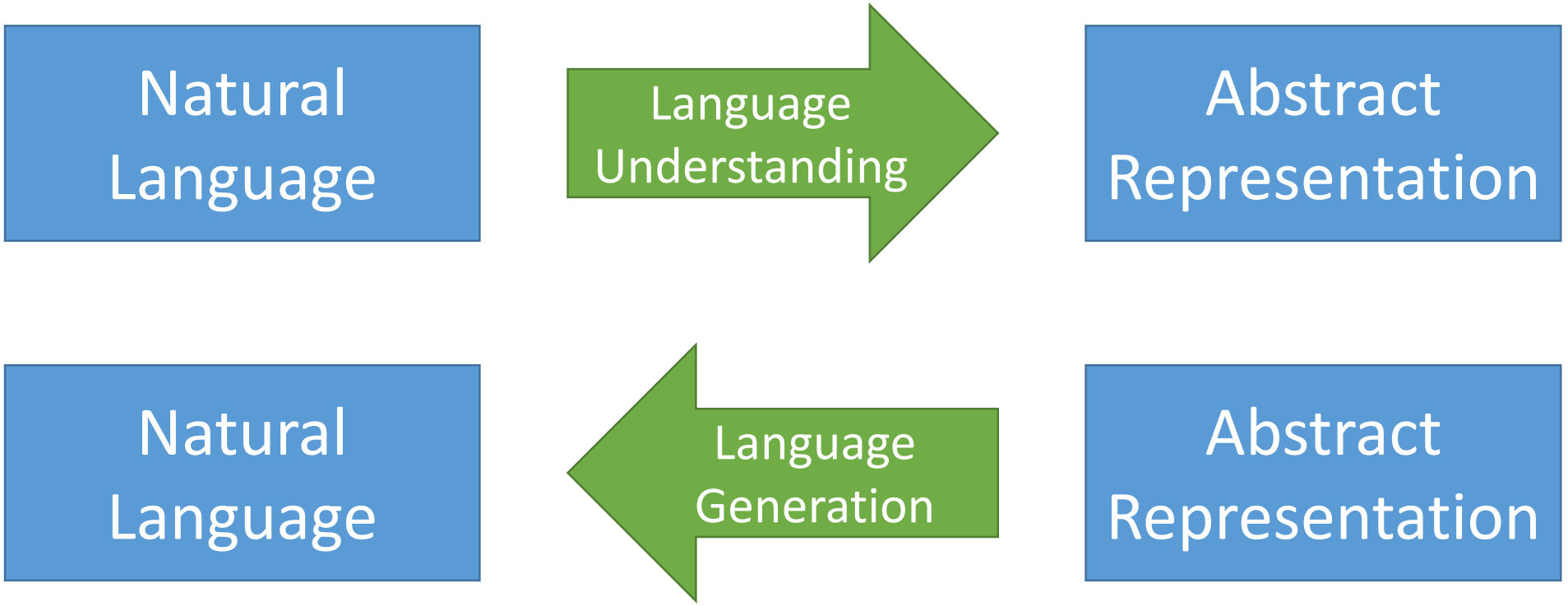
- Master (Boss)
  - rank miniskills
  - long-term coherence
  - user engagement
- Miniskills (Minions)
  - greeting / goodbye / menu / topics
  - probe user personality
  - discuss a news article / movie
  - tell a fact / thought / advice / joke
  - ask / answer a question

Other Dialog Manager Architectures

- Classic AI Planning
- Information State (Markov Decision Process)
- Distributional (Neural Network)

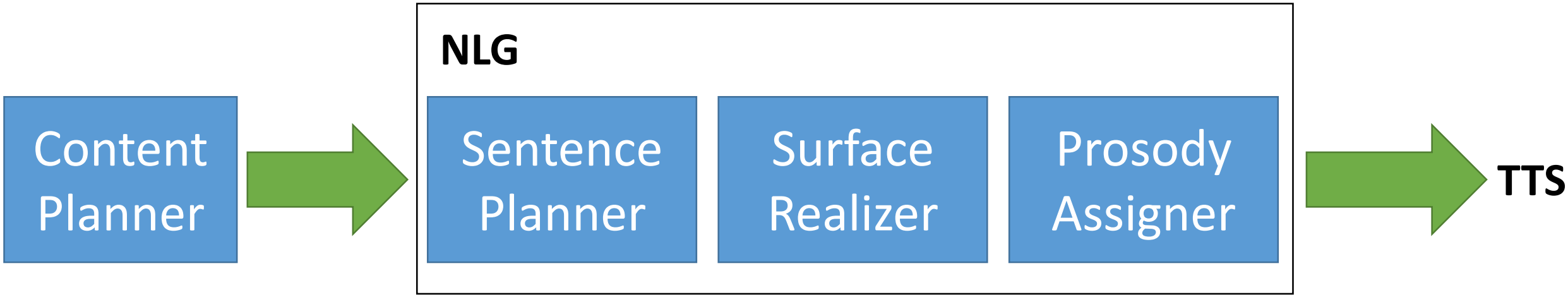
Natural Language Generation

Natural Language Generation (NLG)



NLG Modules

- Content planning
  - what to say
  - a module in dialog manager
- Language generation
  - how to say it
  - select syntactic structure and words
  - adjust prosody



NLG Approaches

- Template-based generation
  - most common in practical systems
  - *“What time do you want to leave CITY-ORIG?”*
  - *“How about we talk about TOPIC?”*
- Neural sequence models
  - recent research interest

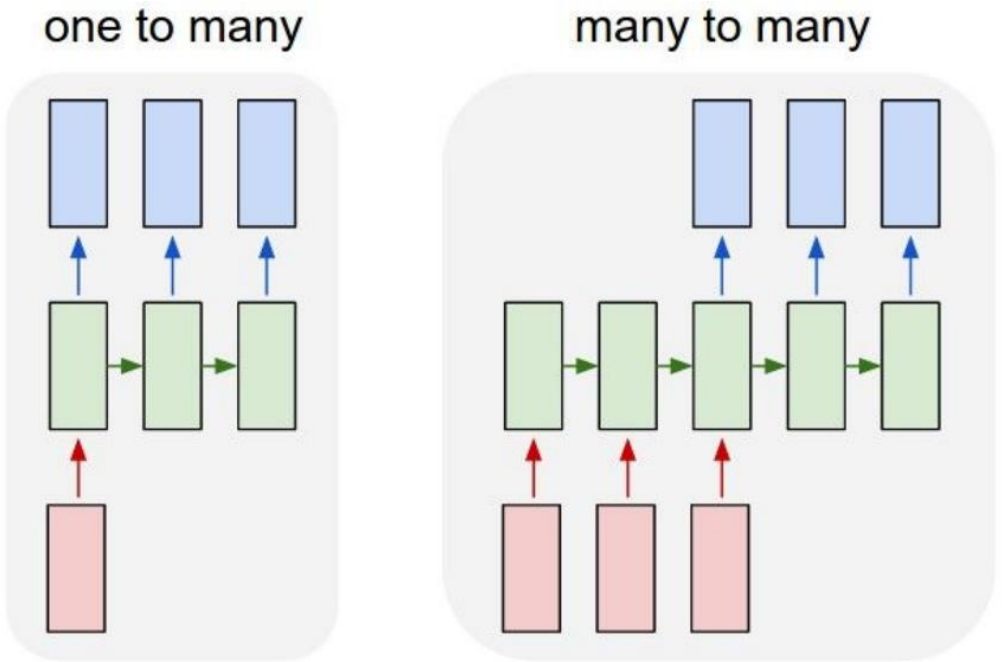


Figure from: Hannaneh Hajishirzi, EE 511 Winter 2018 – “Introduction to Statistical Learning”.

System Evaluation

Motivation

- Goal: determine overall user satisfaction
- A metric to compare systems
  - can’t improve it if we don’t know where it fails
  - can’t decide between two systems without a goodness metric
- A metric as an input to reinforcement learning
  - automatically improve system performance via learning

Dialog System Evaluation

- Extrinsic Evaluation: embedded in some external task
- Intrinsic Evaluation: evaluating the component as such
- What constitutes success or failure for a dialog system?

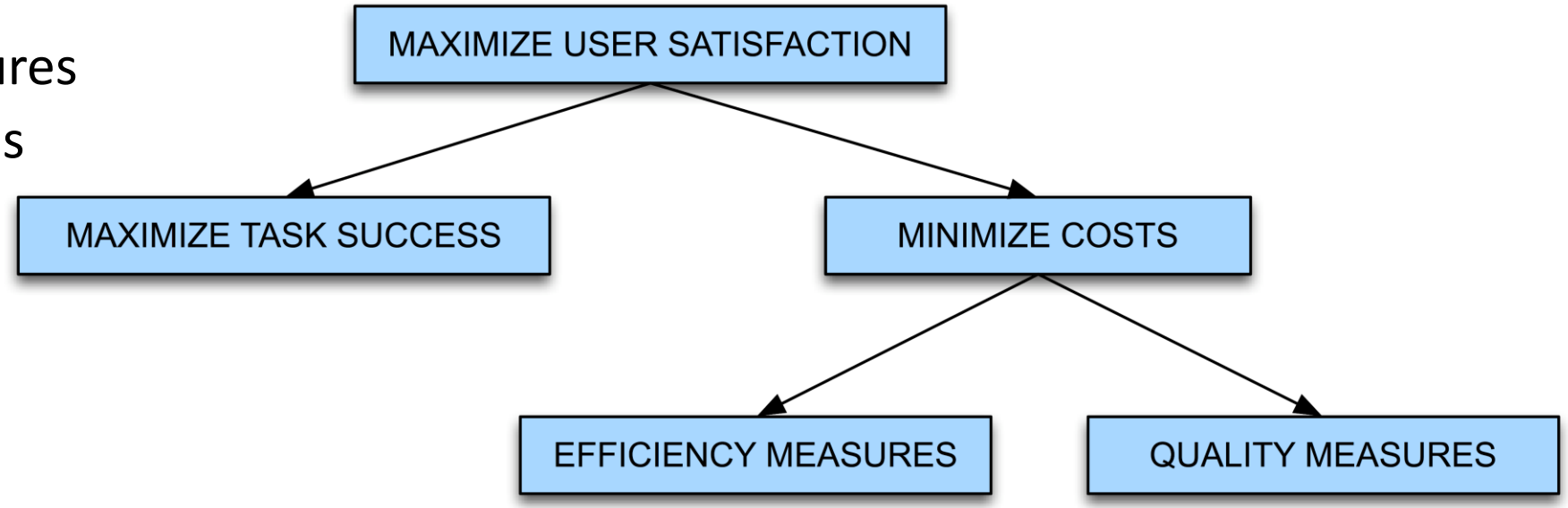
TTS Performance	Was the system easy to understand?
ASR Performance	Did the system understand what you said?
Task Ease	Was it easy to find the message/flight/train you wanted?
Interaction Pace	Was the pace of interaction with the system appropriate?
User Expertise	Did you know what you could say at each point?
System Response	How often was the system sluggish and slow to reply to you?
Expected Behavior	Did the system work the way you expected it to?
Future Use	Do you think you'd use the system in future?

User Satisfaction survey, adapted from (Walker et al. 2001)

PARADISE Framework

PARADISE Framework

- **PARA**digm for **D**ialogue **S**ystem **E**valuation (Walker et al. 2000)
- Maximize Task Success
- Minimize Costs
  - Efficiency Measures
  - Quality Measures



Task Success

- % of subtasks completed
- Correctness of each questions/answer/error message
- Correctness of total solution
  - Error rate in final slots
    - Generalization of Slot Error Rate
- Users' perception of whether task was completed

Efficiency Cost

- Polifroni et al. (1992), Danieli and Gerbino (1995) Hirschman and Pao (1993)
- Total elapsed time in seconds or turns
- Number of queries
- Turn correction ration: number of system or user turns used solely to correct errors, divided by total number of turns

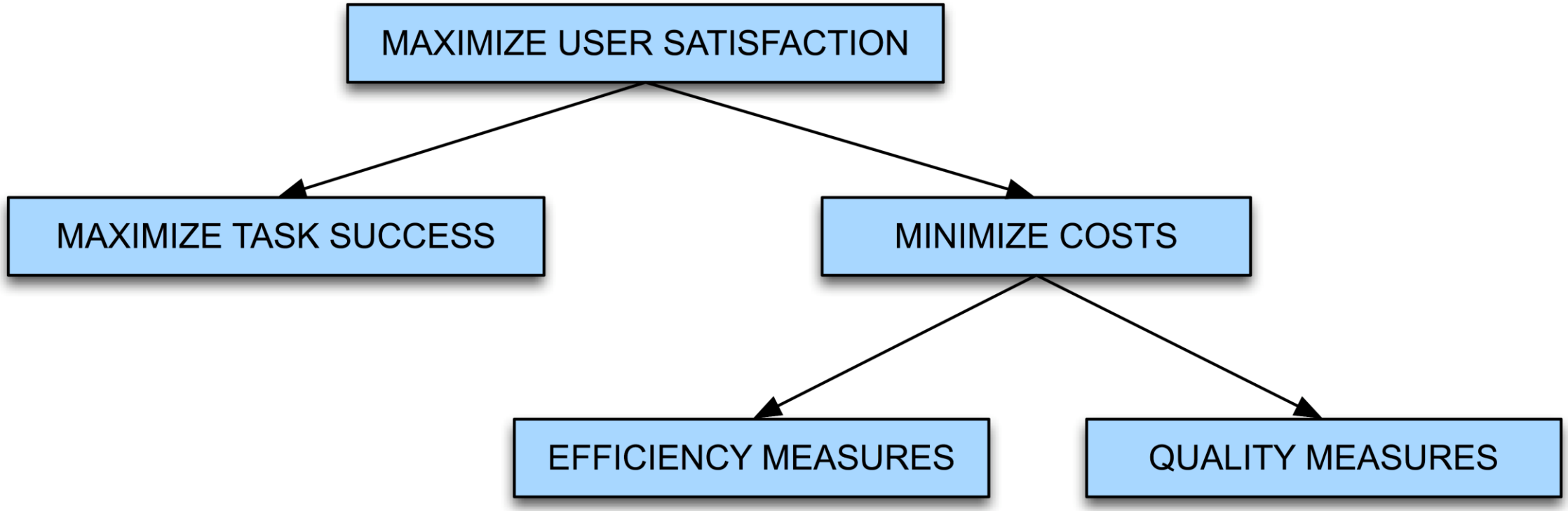
Quality Cost

- # of times ASR system failed to return any sentence
- # of ASR rejection prompts
- # of times user had to barge-in
- # of time-out prompts
- Inappropriateness (verbose, ambiguous) of system's questions, answers, error messages

Concept Accuracy

- “Concept accuracy” or “Concept error rate”
- % of semantic concepts that the NLU component returns correctly
- I want to arrive in Austin at 5:00
  - DESTCITY: Boston
  - Time: 5:00
- Concept accuracy = 50%
- Average this across entire dialogue
- “How many of the sentences did the system understand correctly”
- Can be used as either quality cost or task success

PARADISE: Regress against user satisfaction



Regressing against user satisfaction

- Questionnaire to assign each dialog a “user satisfaction rating”: this is dependent measure
- Set of cost and success factors are independent measures
- Use regression to train weights for each factor

Experimental Procedures

- Subjects given specified tasks
- Spoken dialogs recorded
- Cost factors, states, dialog acts automatically logged
- ASR accuracy, barge-in hand-labeled
- Users specify task solution via web page
- Users complete User Satisfaction surveys
- Use multiple linear regression to model User Satisfaction as a function of Task Success and Costs; test for significant predictive factors

Slide from Julia Hirschberg

Performance Functions from Three Systems

- ELVIS User Sat.= .21* COMP + .47 * MRS - .15 * ET
- TOOT User Sat.= .35* COMP + .45* MRS - .14*ET
- ANNIE User Sat.= .33*COMP + .25* MRS -.33* Help

  - COMP: User perception of task completion (task success)
  - MRS: Mean (concept) recognition accuracy (cost)
  - ET: Elapsed time (cost)
  - Help: Help requests (cost)

Slide from Julia Hirschberg

Evaluation Summary

- Best predictors of User Satisfaction:
  - Perceived task completion
  - mean recognition score (concept accuracy)
- Performance model useful for system development
  - Making predictions about system modifications
  - Distinguishing ‘good’ dialogues from ‘bad’ dialogues
  - As part of a learning model

Now that we have a success metric

- Could we use it to help drive learning?
- Learn an optimal policy or strategy for how the conversational agent should behave

Reward Propagation for Social Chatbots

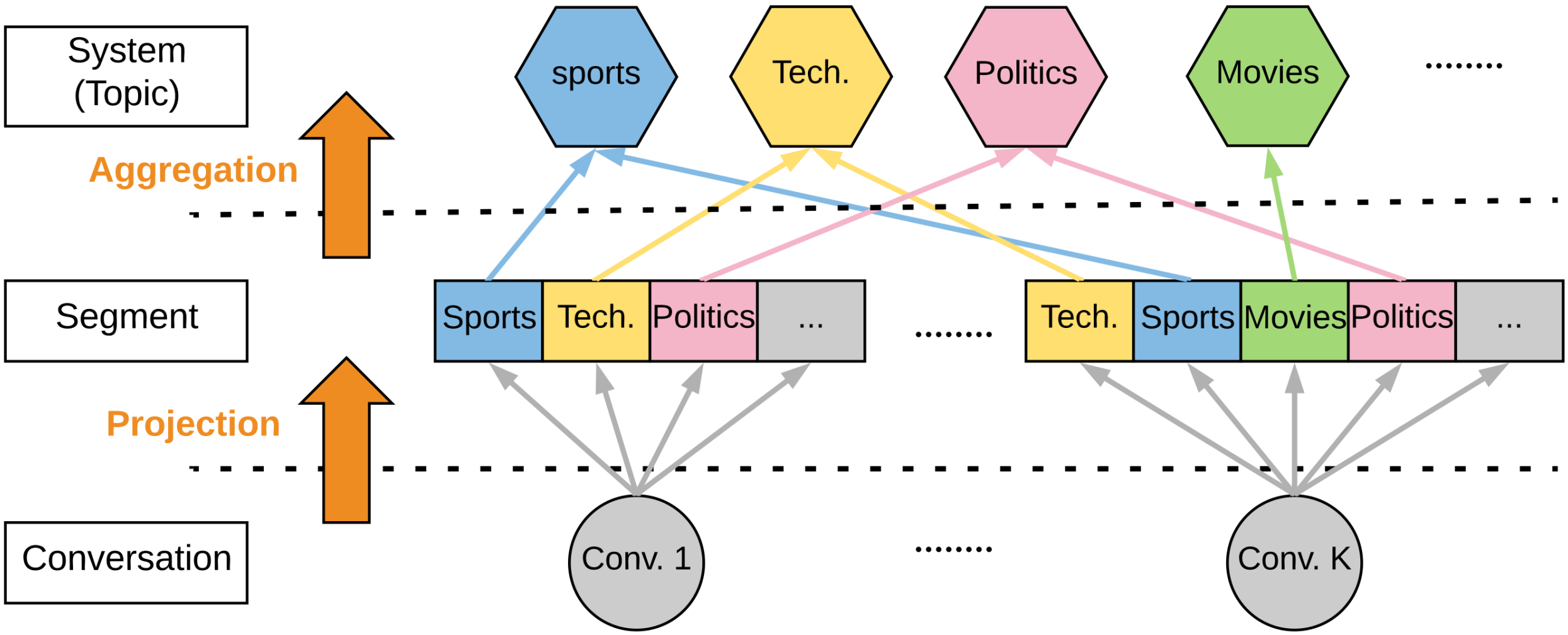
Conversation-level Ratings

- Expensive
  - users may decline to rate the system
- Sparse
  - a dialog typically involves multiple topics (or tasks) which may contribute differently to the final conversation-level rating
- Noisy
  - individual characteristics may influence the way a user interacts with the system and the resulting rating
- Can we use the PARADISE framework?

Dialog Segmentation

- Micro-segments
- Sub-dialogs

Reward Propagation



PARADISE from Reward Propagation

- Step 1: Predictor Groups (predictive features)

Predictor Group	Micro-segm.	Sub-dial.	Conv.
Cost Turns	✓	✓	✓
Engaged Turns	✓	✓	✓
Engaged Topics		✓	✓
Topic Acceptance			✓

PARADISE from Reward Propagation

- Step 2: Learn a linear regression model

$$\mathcal{M}_0 : r_0 \leftarrow \sum_i a_i f_i + b,$$

- f_i: conversation-level features
- b: bias

PARADISE from Reward Propagation

% engaged turns	0.25	% repair turns	-0.10
# accepted topics	0.07	# rejected topics	-0.02
# repair turns	-0.08	% accepted topics	0.01
# negotiation turns	-0.09	# engaged turns	0.37
# engaged topics	-0.16	% engaged topics	0.03

Table 6: Weights learned for selected predictors, ordered (row-by-row, left-to-right) based on the stepwise forward selection process. Underscored numbers indicate statistical significance ($p < .05$).

PARADISE from Reward Propagation

- Step 2: Learn a linear regression model

$$r_0 = 0.1 f_1 + 0.2\ f_2$$

- f_1 : number of engaged turns
- f_2 : percentage of engaged turns

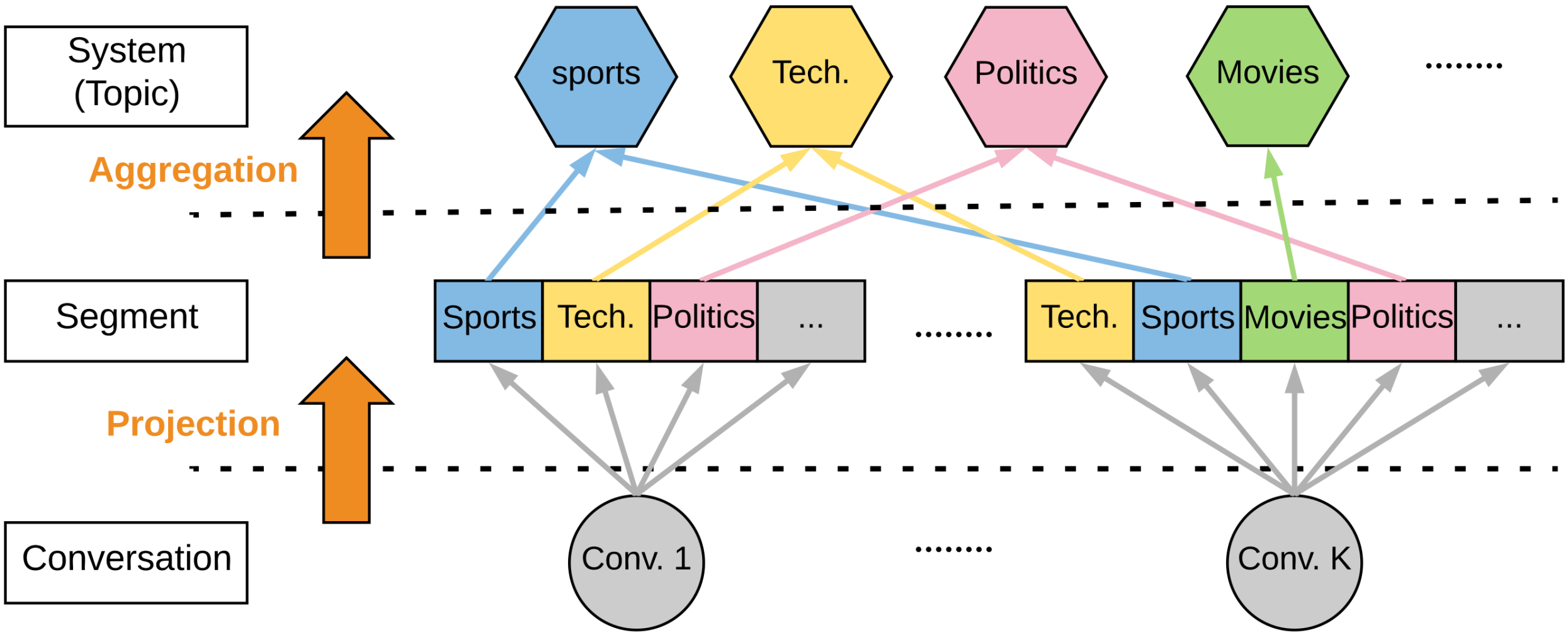
PARADISE from Reward Propagation

- Step 3: Project the reward $\hat{r}_s = \sum_i a_i g_i$

	# Turns	f_1	f_2	$\hat{r}_0$
conversation	100	30	.30	3.06
	# Turns	g_1	g_2	$\hat{r}_s$
sub-dial. 1 (sports)	60	15	.25	1.55
sub-dial. 2 (tech.)	30	6	.20	0.64
sub-dial. 3 (science)	10	9	.90	1.08

$$r_0 = 0.1 f_1 + 0.2\, f_2$$

Reward Propagation



Reward Aggregation: A Case Study

- Step 4: Average all segment-level rewards for sub-dialogs about the same topic.

Domain	Count	Mean	Std. Dev.
TECHNOLOGY	780	0.16	0.25
SCIENCE	972	0.14	0.29
MOVIES	112	0.09	0.27
SPORTS	1336	0.07	0.25
POLITICS	504	0.05	0.28

Challenges & Looking Ahead

Challenges in Dialog Management

- Implementation
  - underspecified or overly complex theories
  - implementation is driven by technical limitations and specific tasks
- Comparison
  - multiple system modules involved (ASR, NLU, NLG, TTS, backend services)
  - domain-specific evaluation metrics
- Data collection & annotation
  - expensive
  - chicken-and-egg issue, i.e., need to have an initial system for data collection

Looking Ahead

- Improvement in speech technologies
- Accessible NLP techniques
- More choices on data collection and annotation
- Combination of science, engineering & art

Upcoming Classes

Upcoming Classes

- April 19: Project proposal presentation
- April 24: Project consulting session + Guest Lecture by Vicky Zayats
- April 26: Lab 2 Checkoff
- May 1: Paper presentation (2 teams)
- May 8: Guest Lecture on Knowledge Graph by Alex Marin
- May 15: Paper presentation (2 teams)
- May 22: Paper presentation (1 team) + Project consulting session

Topics

- The presentation should focus on 1-2 relevant topics and cover several papers.
- Example topics:
  - Language Understanding
  - Dialog Management
  - Language Generation
  - Dialog Analysis
  - End-to-end Systems
  - Reinforcement Learning
  - …

Some keywords & survey papers

- Neural conversation model
  - sequence-to-sequence model
  - attention model
- Partially Observed Markov Decision Process (POMDP)
  - Reinforcement Learning
- Schatzmann et al. 2006. “A Survey of Statistical User Simulation Techniques for Reinforcement-Learning of Dialogue Management Strategies”
- Lemon et al. 2007. “Machine Learning for Spoken Dialogue Systems”
- Frampton et al. 2009. “Recent research advances in reinforcement learning in spoken dialogue systems”

Where to find papers?

- Journals
  - IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)
  - Transactions of the Association for Computational Linguistics (TACL)
  - Dialogue & Discourse
- Conferences & Workshops
  - Special Interest Group on Discourse and Dialogue (SIGdial)
  - INTERSPEECH
  - ACL, EMNLP, NAACL, EACL, COLING
  - ICML, NIPS, ICLR
- Other courses
  - http://courses.washington.edu/ling575/SPR2017/index.html
  - http://web.stanford.edu/class/cs224s/syllabus.html
  - https://dialog-systems-class.github.io/readings.html

Format

- 10% of your final grade
- Each team leads a discussion
  - Week 6 (May 1): 2 teams
  - Week 7 (May 8): Guest Lecture
  - Week 8 (May 15): 2 teams
  - Week 9 (May 22): 1 team + Project Consulting Session
- 50min presentation & discussion
- All team members need participate in the presentation.