

Grounded natural language understanding

Chris Potts
Stanford Linguistics

CS 224U: Natural language understanding
May 21



Overview

- 1 Natural language is situated
- 2 Reasoning about other minds
- 3 Natural language as social
- 4 Examples of grounded NLU systems
- 5 Decision theoretic NLU agents
- 6 Conclusion

HAL

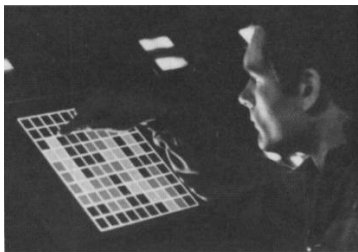
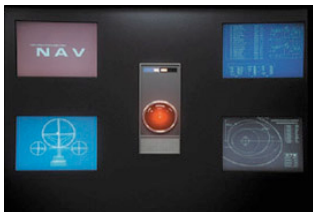
- In the 1967 Stanley Kubrick movie *2001: A Space Odyssey*, the spaceship's computer HAL can
 - display graphics;
 - play chess; and
 - conduct natural, open-domain conversations with humans.
- How well did the filmmakers do at predicting what computers would be capable in 2001?

(Slide idea from Andrew McCallum)

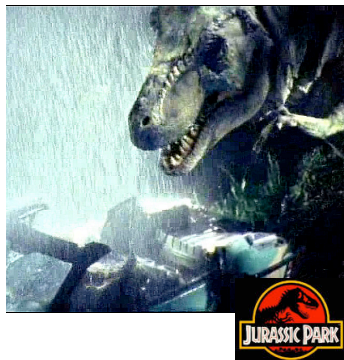
HAL

Graphics

HAL



Jurassic Park (1993)

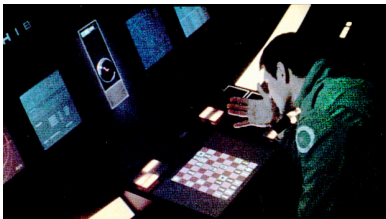


(Slide idea from Andrew McCallum)

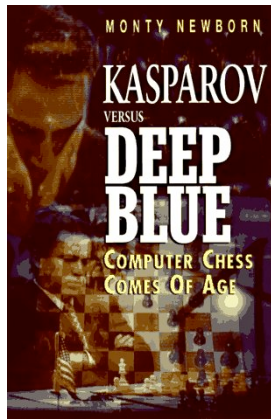
HAL

Chess

HAL



Deep Blue (1997)



(Slide idea from Andrew McCallum)

HAL

Dialogue

HAL

2014

David Bowman: Open the pod bay doors, HAL.

HAL: I'm sorry, Dave, I'm afraid I can't do that.

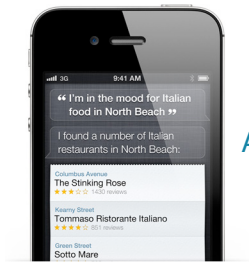
David: What are you talking about, HAL?

HAL: I know that you and Frank were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.



(Slide idea from Andrew McCallum)

Siri



You: Any good burger joints around here?

Siri: I found a number of burger restaurants near you.

You: Hmm. How about tacos?

Apple: [Siri remembers that you asked about restaurants. so it will look for Mexican restaurants in the neighborhood. And Siri is proactive, so it will question you until it finds what you're looking for.]

(Slide from Marie de Marneffe)

Siri

Colbert: For the love of God, the cameras are on, give me something?

Siri: What kind of place are you looking for? Camera stores or churches?
[...]

Colbert: I don't want to search for anything! I want to write the show!

Siri: Searching the Web for "search for anything. I want to write the shuffle."



(Slide from Marie de Marneffe)

Language is action

Winograd (1986:170):

“all language use can be thought of as a way of activating procedures within the hearer. We can think of an utterance as a program – one that indirectly causes a set of operations to be carried out within the hearer’s cognitive system.”

Levinson's (2000) analogy



Figure 0.1

Rembrandt sketch

Levinson's (2000) analogy

“We interpret this sketch instantly and effortlessly as a gathering of people before a structure, probably a gateway; the people are listening to a single declaiming figure in the center. [...] But all this is a miracle, for there is little detailed information in the lines or shading (such as there is). Every line is a mere suggestion [...]. So here is the miracle: from a merest, sketchiest squiggle of lines, you and I converge to find adumbration of a coherent scene [...].



Figure 0.1
Rembrandt sketch

Levinson's (2000) analogy

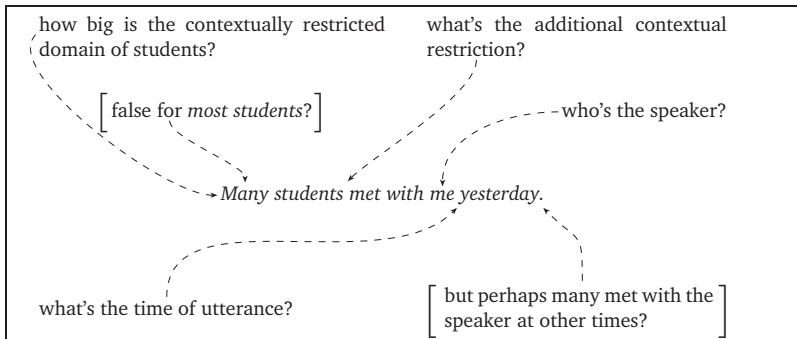


Figure 0.1
Rembrandt sketch

“We interpret this sketch instantly and effortlessly as a gathering of people before a structure, probably a gateway; the people are listening to a single declaiming figure in the center. [...] But all this is a miracle, for there is little detailed information in the lines or shading (such as there is). Every line is a mere suggestion [...]. So here is the miracle: from a merest, sketchiest squiggle of lines, you and I converge to find adumbration of a coherent scene [...].

“The problem of utterance interpretation is not dissimilar to this visual miracle. An utterance is not, as it were, a veridical model or “snapshot” of the scene it describes [...]. Rather, an utterance is just as sketchy as the Rembrandt drawing.”

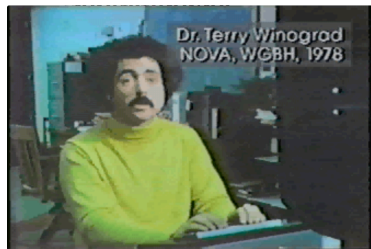
Routine pragmatic enrichment



SHRDLU (Winograd 1972)

Full-grounded system that parses the user's input, maps it to logical form, interprets that logical form in its world, tries to take appropriate action, and generates responses.

<http://hci.stanford.edu/winograd/shrdlu/>



One project did succeed. Terry Winograd's program SHRDLU could use English intelligently, but there was a catch: the only subject you could discuss was a micro-world of simulated blocks.

SHRDLU (Winograd 1972)

- 1 Person: WHAT DOES THE BOX CONTAIN?
- 2 Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
- 3 Person: WHAT IS THE PYRAMID SUPPORTED BY?
- 4 Computer: THE BOX.
- 5 Person: HOW MANY BLOCKS ARE NOT IN THE BOX?
- 6 Computer: FOUR OF THEM.
- 7 Person: IS AT LEAST ONE OF THEM NARROWER THAN THE ONE WHICH I TOLD YOU TO PICK UP?
- 8 Computer: YES, THE RED CUBE.
- 9 Person: IS IT SUPPORTED?
- 10 Computer: YES, BY THE TABLE.

Natural language is situated

- 1 **Natural language is situated**
- 2 Reasoning about other minds
- 3 Natural language as social
- 4 Examples of grounded NLU systems
- 5 Decision theoretic NLU agents
- 6 Conclusion

Indexicality

Indexicality

① I am speaking.

Indexicality

- 1 I am speaking.
- 2 We won.

[A team I'm on; a team I support; ...]

Indexicality

- ① I am speaking.
- ② We won.
- ③ I am here

[A team I'm on; a team I support; ...]

[classroom; Stanford; ... planet earth; ...]

Indexicality

- ① I am speaking.
- ② We won. [A team I'm on; a team I support; ...]
- ③ I am here [classroom; Stanford; ... planet earth; ...]
- ④ We are here. [pointing at a map]

Indexicality

- 1 I am speaking.
- 2 We won. [A team I'm on; a team I support; ...]
- 3 I am here [classroom; Stanford; ... planet earth; ...]
- 4 We are here. [pointing at a map]
- 5 I'm not here now. [old-fashioned answering machine]

Indexicality

- 1 I am speaking.
- 2 We won. [A team I'm on; a team I support; ...]
- 3 I am here [classroom; Stanford; ... planet earth; ...]
- 4 We are here. [pointing at a map]
- 5 I'm not here now. [old-fashioned answering machine]
- 6 We went to a local bar after work.

Indexicality

- 1 I am speaking.
- 2 We won. [A team I'm on; a team I support; ...]
- 3 I am here [classroom; Stanford; ... planet earth; ...]
- 4 We are here. [pointing at a map]
- 5 I'm not here now. [old-fashioned answering machine]
- 6 We went to a local bar after work.
- 7 three days ago, tomorrow, now

Context dependence

Where are you from?

Context dependence

Where are you from?

- *Connecticut.*

(Issue: birthplaces)

Context dependence

Where are you from?

- *Connecticut.* (Issue: birthplaces)
- *The U.S.* (Issue: nationalities)

Context dependence

Where are you from?

- *Connecticut.* (Issue: birthplaces)
- *The U.S.* (Issue: nationalities)
- *Stanford.* (Issue: affiliations)

Context dependence

Where are you from?

- *Connecticut.* (Issue: birthplaces)
- *The U.S.* (Issue: nationalities)
- *Stanford.* (Issue: affiliations)
- *Planet earth.* (Issue: intergalactic meetings)

Context dependence

I didn't see any.

Context dependence

- Are there typos in my slides?

I didn't see any.

Context dependence

- Are there typos in my slides?
- Are there bookstores downtown?

I didn't see any.

Context dependence

- Are there typos in my slides?
- Are there bookstores downtown?
- Are there cookies in the cupboard?

I didn't see any.

Context dependence

- Are there typos in my slides?
- Are there bookstores downtown?
- Are there cookies in the cupboard?
- ...

I didn't see any.

Context dependence

- 1 The light is on. Chris must be in his office.
- 2 The Dean passed a new rule. Chris must be in his office.

Context dependence

If kangaroos had no tails, they would fall over.

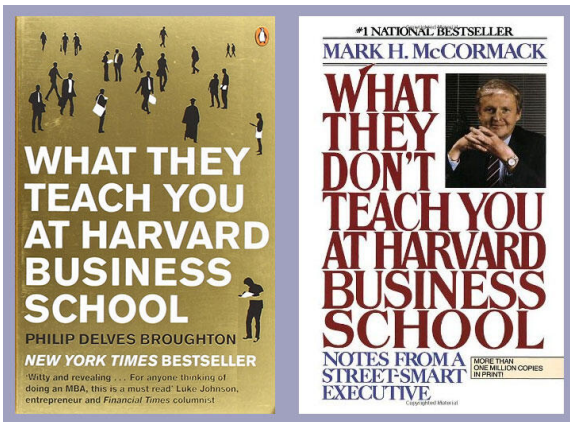
Seems true

Context dependence

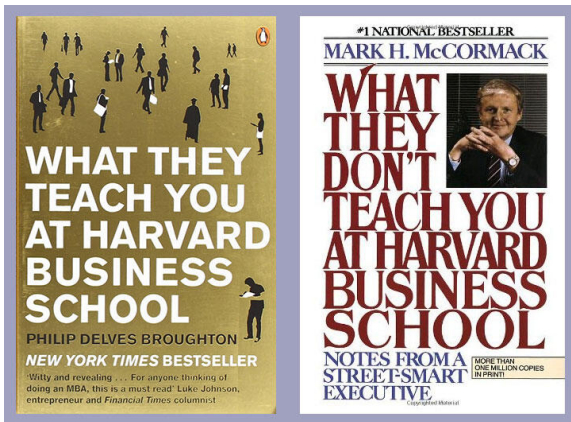
If kangaroos had no tails, they would fall over.

Seems true, but suppose they had jetpacks.

Context dependence



Context dependence



“These two books contain the sum total of all human knowledge”
 (@James_Kpatrick)

Perspectival expressions



Colors in context




Context			Utterance
			blue

Table: Example from the Colors in Context corpus from the Stanford Computation & Cognition Lab

Colors in context

	Context		Utterance
			blue
			The darker blue one

Table: Example from the Colors in Context corpus from the Stanford Computation & Cognition Lab

Colors in context








	Context		Utterance
			blue
			The darker blue one
			dull pink not the super bright one

Table: Example from the Colors in Context corpus from the Stanford Computation & Cognition Lab

Colors in context













	Context		Utterance
			blue
			The darker blue one
			dull pink not the super bright one
			Purple

Table: Example from the Colors in Context corpus from the Stanford Computation & Cognition Lab

Colors in context
















	Context		Utterance
			blue
			The darker blue one
			dull pink not the super bright one
			Purple
			blue

Table: Example from the Colors in Context corpus from the Stanford Computation & Cognition Lab

Situated word learning

Children learn word meanings

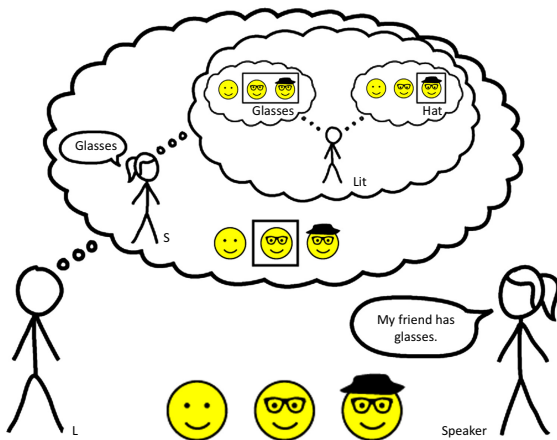
- 1 with incredible speed
- 2 despite relatively few inputs
- 3 by using cues from
 - contrast inherent in the forms they hear
 - social cues
 - assumptions about the speaker's goals
 - regularities in the physical environment.

Frank et al. (2012); Frank and Goodman (2014)

Reasoning about other minds

- 1 Natural language is situated
- 2 Reasoning about other minds**
- 3 Natural language as social
- 4 Examples of grounded NLU systems
- 5 Decision theoretic NLU agents
- 6 Conclusion

Reference resolution under uncertainty



From Goodman and Frank 2016

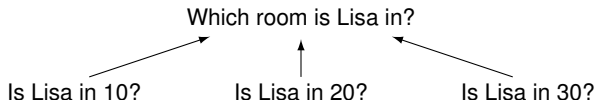
Attending to the questions under discussion

Context: Homer calls a hotel.

Homer: Is Lisa Simpson in Room 10?

Clerk A: She's in room 20.

Clerk B: #No.



Other phenomena involving reasoning about other minds

Other phenomena involving reasoning about other minds

1 I think this is the way to the library.

[politeness]

Other phenomena involving reasoning about other minds

- 1 I think this is the way to the library. [politeness]
- 2 Would you mind if I stole your pen for a second? [politeness]

Other phenomena involving reasoning about other minds

- 1 I think this is the way to the library. [politeness]
- 2 Would you mind if I stole your pen for a second? [politeness]
- 3 He's not exactly a genius/idiot. [irony]

Other phenomena involving reasoning about other minds

- 1 I think this is the way to the library. [politeness]
- 2 Would you mind if I stole your pen for a second? [politeness]
- 3 He's not exactly a genius/idiot. [irony]
- 4 Great idea! [sarcasm(?)]

Other phenomena involving reasoning about other minds

- 1 I think this is the way to the library. [politeness]
- 2 Would you mind if I stole your pen for a second? [politeness]
- 3 He's not exactly a genius/idiot. [irony]
- 4 Great idea! [sarcasm(?)]
- 5 Any chance we can sort this out here, officer? [bribery(?)]

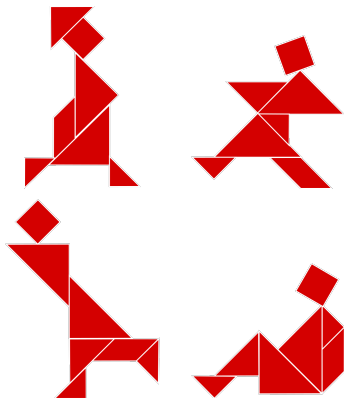
Other phenomena involving reasoning about other minds

- 1 I think this is the way to the library. [politeness]
- 2 Would you mind if I stole your pen for a second? [politeness]
- 3 He's not exactly a genius/idiot. [irony]
- 4 Great idea! [sarcasm(?)]
- 5 Any chance we can sort this out here, officer? [bribery(?)]
- 6 It'd be a shame if something happened to your dog. [threat(?)]

Natural language use is social

- 1 Natural language is situated
- 2 Reasoning about other minds
- 3 Natural language as social**
- 4 Examples of grounded NLU systems
- 5 Decision theoretic NLU agents
- 6 Conclusion

Lexical pacts



Round 1: All right, the next one looks like a person who's ice skating, except they're sticking their arms out in front.

Round 2: Um, the next one's the person ice skating that has arms out.

[...]

Round 6: The ice skater.

(Clark and Wilkes-Gibbs 1986)

Style matching (alignment)

When interacting, people unconsciously align their communicative behaviors at many levels:

- Posture
- Head nodding
- Speech rate
- Pause length
- Backchannel
- Self-disclosure
- Function word rates
- Concept naming



(Niederhoffer and Pennebaker 2002; Danescu-Niculescu-Mizil et al. 2013; Doyle et al. 2016; Srivastava et al. 2016)

Shared assumptions and Winograd sentences

(Winograd 1972; Levesque 2013)

Shared assumptions and Winograd sentences

- 1 The council refused the demonstrators a permit because they **feared** violence. Who **feared** violence?
The council/The demonstrators

(Winograd 1972; Levesque 2013)

Shared assumptions and Winograd sentences

- 1 The council refused the demonstrators a permit because they **feared** violence. Who **feared** violence?
The council/The demonstrators
- 2 The council refused the demonstrators a permit because they **advocated** violence. Who **advocated** violence?
The council/**The demonstrators**

(Winograd 1972; Levesque 2013)

Shared assumptions and Winograd sentences

- 1 The council refused the demonstrators a permit because they **feared** violence. Who **feared** violence?
The council/The demonstrators
- 2 The council refused the demonstrators a permit because they **advocated** violence. Who **advocated** violence?
The council/**The demonstrators**
- 3 Sandy told Kim how to fix the air conditioner.

(Winograd 1972; Levesque 2013)

Shared assumptions and Winograd sentences

- 1 The council refused the demonstrators a permit because they **feared** violence. Who **feared** violence?
The council/The demonstrators
- 2 The council refused the demonstrators a permit because they **advocated** violence. Who **advocated** violence?
The council/**The demonstrators**
- 3 Sandy told Kim how to fix the air conditioner.
 - Sandy is a master plumber; Kim is an apprentice.

(Winograd 1972; Levesque 2013)

Shared assumptions and Winograd sentences

- 1 The council refused the demonstrators a permit because they **feared** violence. Who **feared** violence?
The council/The demonstrators
- 2 The council refused the demonstrators a permit because they **advocated** violence. Who **advocated** violence?
The council/**The demonstrators**
- 3 Sandy told Kim how to fix the air conditioner.
 - Sandy is a master plumber; Kim is an apprentice.
 - Kim felt **grateful**/annoyed.

(Winograd 1972; Levesque 2013)

Shared assumptions and Winograd sentences

- 1 The council refused the demonstrators a permit because they **feared** violence. Who **feared** violence?
The council/The demonstrators
- 2 The council refused the demonstrators a permit because they **advocated** violence. Who **advocated** violence?
The council/**The demonstrators**
- 3 Sandy told Kim how to fix the air conditioner.
 - Sandy is a master plumber; Kim is an apprentice.
 - Kim felt **grateful**/annoyed.
- 4 Sandy told Kim how to fix the air conditioner.

(Winograd 1972; Levesque 2013)

Shared assumptions and Winograd sentences

- 1 The council refused the demonstrators a permit because they **feared** violence. Who **feared** violence?
The council/The demonstrators
- 2 The council refused the demonstrators a permit because they **advocated** violence. Who **advocated** violence?
The council/**The demonstrators**
- 3 Sandy told Kim how to fix the air conditioner.
 - Sandy is a master plumber; Kim is an apprentice.
 - Kim felt **grateful**/annoyed.
- 4 Sandy told Kim how to fix the air conditioner.
 - Kim is a master plumber; Sandy is an apprentice.

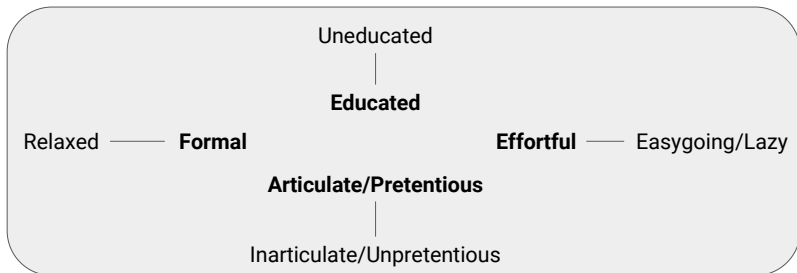
(Winograd 1972; Levesque 2013)

Shared assumptions and Winograd sentences

- 1 The council refused the demonstrators a permit because they **feared** violence. Who **feared** violence?
The council/The demonstrators
- 2 The council refused the demonstrators a permit because they **advocated** violence. Who **advocated** violence?
The council/**The demonstrators**
- 3 Sandy told Kim how to fix the air conditioner.
 - Sandy is a master plumber; Kim is an apprentice.
 - Kim felt **grateful**/annoyed.
- 4 Sandy told Kim how to fix the air conditioner.
 - Kim is a master plumber; Sandy is an apprentice.
 - Kim felt grateful/**annoyed**.

(Winograd 1972; Levesque 2013)

Sociolinguistic variation



(Campbell-Kibler 2007; Eckert 2008)

Consequences for NLU

- Human children are the best agents in the universe at learning language, and they depend heavily on grounding.
- Problems that are intractable without grounding are solvable with the right kinds of grounding.
- Deep learning is a flexible toolkit for reasoning about different kinds of information in a single model, so it's lead to conceptual and empirical improvements in this area.
- We should seek out (and develop) data sets that include the right kind of grounding.

Some grounded NLU systems

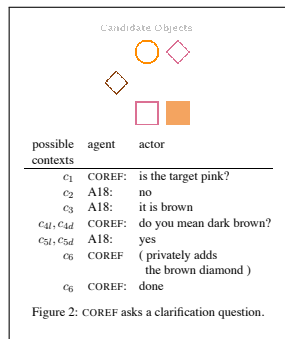
- 1 Natural language is situated
- 2 Reasoning about other minds
- 3 Natural language as social
- 4 Examples of grounded NLU systems**
- 5 Decision theoretic NLU agents
- 6 Conclusion

Hard-wired indexical assumptions

- *I* picks out the current user.
- *you* is the agent.
- *Kathryn* is a distribution over names in the address book.
- *now* includes the current time.
- *here* includes the current location (size set by current task?)
- *Chicago* is a distribution over music, movies, or locations, biased by the current location.
- ...

The COREF system (DeVault and Stone)

COREF and its human interlocutor collaborate on a simple referential task, improving forms and resolving ambiguities using contextual and linguistic information.



possible contexts	agent	actor
c_1	COREF:	is the target pink?
c_2	A18:	no
c_3	A18:	it is brown
c_{4l}, c_{4d}	COREF:	do you mean dark brown?
c_{5l}, c_{5d}	A18:	yes
c_6	COREF	(privately adds the brown diamond)
c_6	COREF:	done

Figure 2: COREF asks a clarification question.

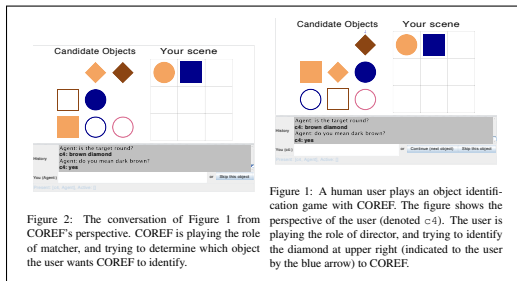


Figure 1: A human user plays an object identification game with COREF. The figure shows the perspective of the user (denoted c_4). The user is playing the role of director, and trying to identify the diamond at upper right (indicated to the user by the blue arrow) to COREF.

Figure 2: The conversation of Figure 1 from COREF's perspective. COREF is playing the role of matcher, and trying to determine which object the user wants COREF to identify.

(DeVault and Stone 2007, 2009)

The Rational Speech Acts Model

The Rational Speech Acts Model

Literal listener

$$I_0(w \mid msg, Lex) \propto Lex(msg, w)P(w)$$

The Rational Speech Acts Model

Pragmatic speaker

$$s_1(msg | w, Lex) \propto \exp \lambda (\log I_0(w | msg, Lex) - C(msg))$$

Literal listener

$$I_0(w | msg, Lex) \propto Lex(msg, w)P(w)$$

The Rational Speech Acts Model

Pragmatic listener

$$l_1(w | msg, Lex) \propto s_1(msg | w, Lex)P(w)$$

Pragmatic speaker

$$s_1(msg | w, Lex) \propto \exp \lambda (\log l_0(w | msg, Lex) - C(msg))$$

Literal listener

$$l_0(w | msg, Lex) \propto Lex(msg, w)P(w)$$

The Rational Speech Acts Model

Pragmatic listener

$$l_1(w | msg, Lex) = \text{pragmatic speaker} \times \text{state prior}$$

Pragmatic speaker

$$s_1(msg | w, Lex) = \text{literal listener} - \text{message costs}$$

Literal listener

$$l_0(w | msg, Lex) = \text{lexicon} \times \text{state prior}$$

RSA listener example



beard

T

F

glasses

T

T

l_1

s_1

l_0

Lex

RSA listener example



beard

1

0

glasses

.5

.5



l_1

s_1

l_0

Lex

RSA listener example

	<i>beard</i>	<i>glasses</i>
	.67	.33
	0	1

 l_1 **s_1** l_0

Lex

RSA listener example



beard

1

0

glasses

.25

.75

l_1

s_1

l_0

Lex

Limitations

Limitations

- Hand-specified lexicon

Limitations

- Hand-specified lexicon
- Reasoning about *all* possible utterances?



$$s_1(msg | w, Lex) = \frac{l_0(w | msg, Lex)}{\sum_{msg'} l_0(w | msg', Lex)}$$

Limitations

- Hand-specified lexicon
- Reasoning about *all* possible utterances?

$$s_1(msg | w, Lex) = \frac{I_0(w | msg, Lex)}{\sum_{msg'} I_0(w | msg', Lex)}$$

- High-bias model; few chances to learn from data

		
<i>beard</i>	1	0
<i>glasses</i>	.25	.75

Golland et al. (2010)

Pioneering pragmatic speaker and listener agents:

- The speaker observes a referent w and chooses a message msg by reasoning pragmatically about the lexicon.
- The listener observes the speaker's msg and chooses a referent w' .
- Their shared utility is based on w and w' .
- Complex utterances are interpreted compositionally, in terms of distributions over possible referents.

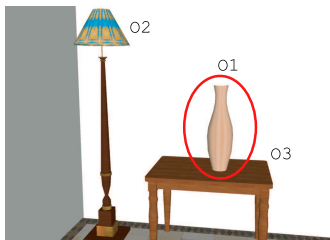


Figure 1: An example of a 3D model of a room. The *speaker's* goal is to reference the target object 01 by describing its spatial relationship to other object(s). The *listener's* goal is to guess the object given the speaker's description.

Tellex et al.'s (2014) Inverse Semantics

Robots and humans collaborate to assemble IKEA furniture:

- “Our approach views the language generation problem as inverse language understanding”
- “By modeling the probability of a human misinterpreting the request, the robot is able to generate targeted requests that humans follow more quickly and accurately [. . .]”
- Robot utterances are scored by models similar to RSA’s pragmatic speakers.



“Help me” (S_0) “Help me.”
 Templates “Please hand me part 2.”
 $G^3 S_1$ “Give me the white leg.”
 $G^3 S_2$ “Give me the white leg that is on the black table.”
 Hand-written “Take the table leg that is on the table and place it in the robot’s hand.”

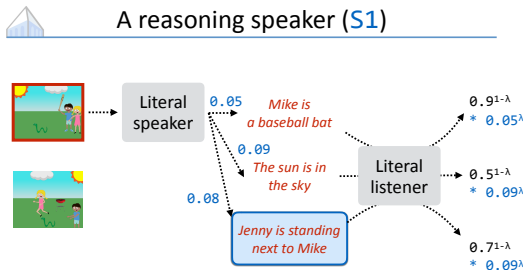
Fig. 5. Scene from our dataset and the requests generated by each approach.

TABLE II
 FRACTION OF CORRECTLY FOLLOWED REQUESTS

Metric	% Success	95% Confidence
Chance	20.0	
“Help me” Baseline (S_0)	21.0	±8.0
Template Baseline	47.0	±5.7
G^3 Inverse Semantics with S_1	52.3	±5.7
G^3 Inverse Semantics with S_2	64.3	±5.4
Hand-Written Requests	94.0	±4.7

Andreas and Klein (2016)

The speaker observes w and seeks to choose an utterance that maximizes the probability that a listener would pick out w :



33

(Diagram from Jacob Andreas)

It is intractable to reason about *all* utterances, so the pragmatic speaker samples utterances from the literal speaker.

Colors in context
















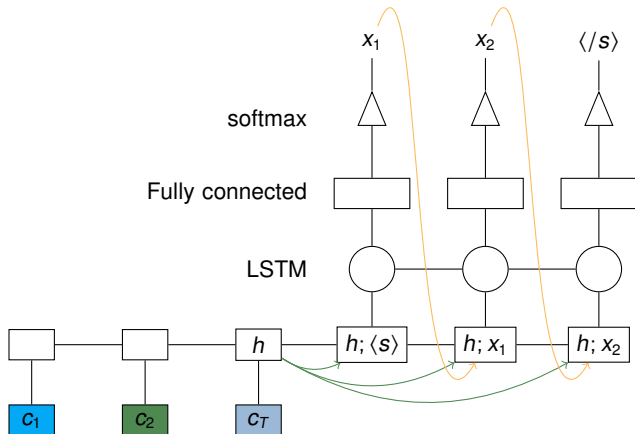
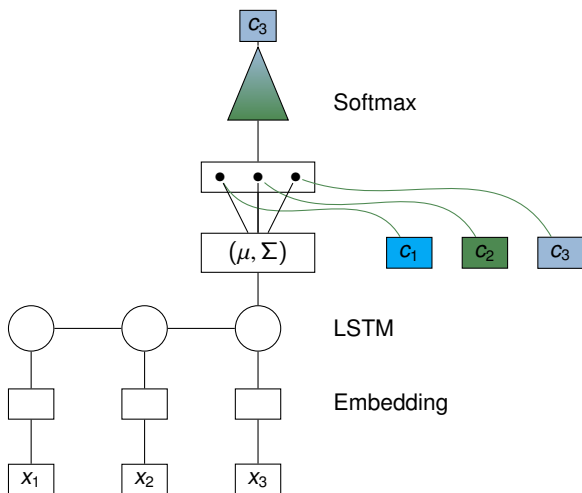
	Context		Utterance
			blue
			The darker blue one
			dull pink not the super bright one
			Purple
			blue

Table: Example from the Colors in Context corpus from the Stanford Computation & Cognition Lab

Literal neural speaker S_0



Neural literal listener \mathcal{L}_0



Neural pragmatic agents

Neural pragmatic agents

Neural pragmatic speaker (Andreas and Klein 2016)

$$\mathcal{S}_1(msg | c, C; \theta) = \frac{\mathcal{L}_0(c | msg, C; \theta)}{\sum_{msg' \in X} \mathcal{L}_0(c | msg', C; \theta)}$$

Neural pragmatic agents

Neural pragmatic speaker (Andreas and Klein 2016)

$$\mathcal{S}_1(msg | c, C; \theta) = \frac{\mathcal{L}_0(c | msg, C; \theta)}{\sum_{msg' \in X} \mathcal{L}_0(c | msg', C; \theta)}$$

where X is a sample from $\mathcal{S}_0(msg | c, C; \theta)$ such that $msg^* \in X$.

Neural pragmatic agents

Neural pragmatic speaker (Andreas and Klein 2016)

$$\mathcal{S}_1(msg | c, C; \theta) = \frac{\mathcal{L}_0(c | msg, C; \theta)}{\sum_{msg' \in X} \mathcal{L}_0(c | msg', C; \theta)}$$

where X is a sample from $\mathcal{S}_0(msg | c, C; \theta)$ such that $msg^* \in X$.

Neural pragmatic listener

$$\mathcal{L}_1(c | msg, C; \theta) \propto \mathcal{S}_1(msg | c, C; \theta)$$

Neural pragmatic agents

Neural pragmatic speaker (Andreas and Klein 2016)

$$\mathcal{S}_1(msg | c, C; \theta) = \frac{\mathcal{L}_0(c | msg, C; \theta)}{\sum_{msg' \in X} \mathcal{L}_0(c | msg', C; \theta)}$$

where X is a sample from $\mathcal{S}_0(msg | c, C; \theta)$ such that $msg^* \in X$.

Neural pragmatic listener

$$\mathcal{L}_1(c | msg, C; \theta) \propto \mathcal{S}_1(msg | c, C; \theta)$$

Blended neural pragmatic listener

Weighted combination of \mathcal{L}_0 and \mathcal{L}_1 .

Pragmatic image captioning

Mao et al. (2016); Vedantam et al. (2017): Captions that are true *and distinguish their images from related ones*.



S_0 caption: the dog is brown

S_1 caption: the head of a dog

Reasoning about *all* possible utterances/captions?

(Cohn-Gordon et al. 2018)

Pragmatic image captioning

Mao et al. (2016); Vedantam et al. (2017): Captions that are true *and distinguish their images from related ones*.



S_0 caption: the dog is brown

S_1 caption: the head of a dog

Reasoning about *all* possible utterances/captions?

⇒ Sample from \mathcal{S}_0

(Cohn-Gordon et al. 2018)

Pragmatic image captioning

Mao et al. (2016); Vedantam et al. (2017): Captions that are true *and distinguish their images from related ones*.



S_0 caption: the dog is brown

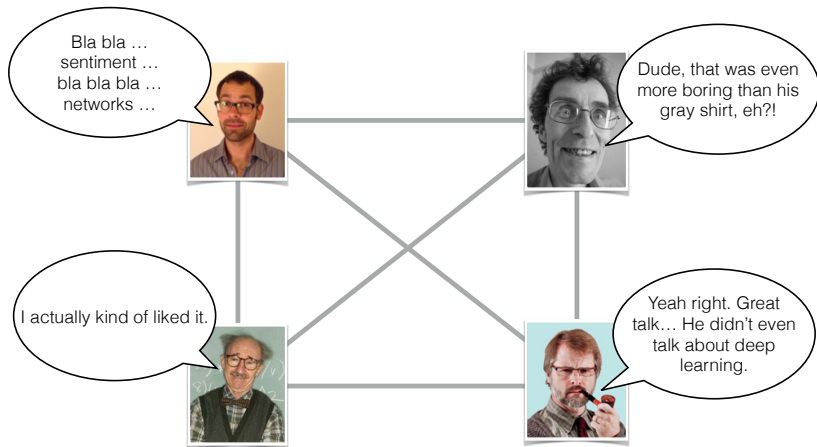
S_1 caption: the head of a dog

Reasoning about *all* possible utterances/captions?

⇒ **Full RSA reasoning about *characters***

(Cohn-Gordon et al. 2018)

Sentiment and social networks

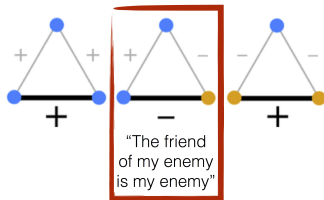


(West et al. 2014)

Sentiment and social networks



Social balance theory



(West et al. 2014)

PLOW: webpage structure as context

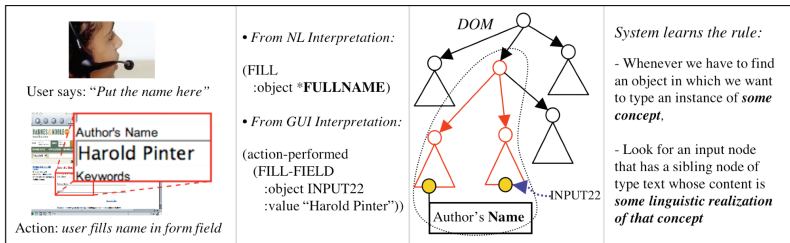


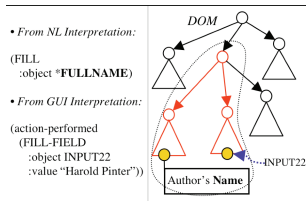
Figure 4: Learning to find and fill a text field

- Learning rules of the form 'If A, then B, else C' is a challenge because the latent variable A is generally not observed. Rather, one sees only B or C.
- In an interactive, instructional setting, one needn't rely entirely on abduction or probabilistic inference: users generally state the needed rules during their interactions.

(Allen et al. 2007)

PLOW: webpage structure as context

- 1 The user's actions ground the parsed language.



- 2 The DOM structure grounds the user's indexicals: referential devices.
 - Put the name here. (user clicks on the DOM element)
 - This is the ISBN number. (user highlights some text)
 - Find another tab. (user has selected a tab)
- 3 Indefinites mark new info; definites refer to established info:
 - *A man walked in. He/The man looked tired.*
 - *an address* ⇒ new input parameter
 - *the address* ⇒ existing input parameter

Common themes

- These systems draw on both speaker and listener perspectives, drawing on the insight that most humans play both roles as well.
- They mix linguistic and non-linguistic information.
- They seek to learn context-dependent meanings.
- They draw insights from linguistics and cognitive psychology, but they confront the scalability issues of NLP.

Decision theoretic NLU agents

- 1 Natural language is situated
- 2 Reasoning about other minds
- 3 Natural language as social
- 4 Examples of grounded NLU systems
- 5 Decision theoretic NLU agents**
- 6 Conclusion

A decision-theoretic framework for dialogue agents

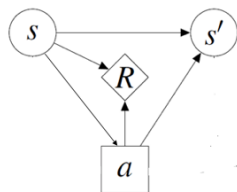


Figure: MDP

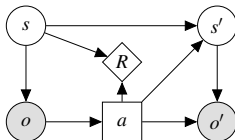


Figure: POMDP

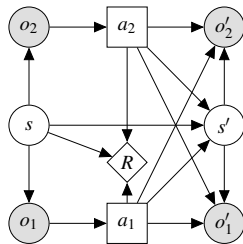
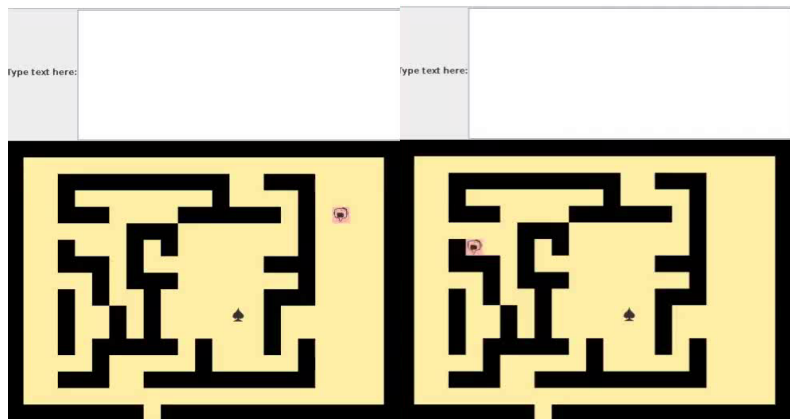


Figure: Dec-POMDP

Scenario

Both players must find the ace of spades.

DialogBot:



(Adapted from the Cards Corpus of Potts 2012)

POMDPs and approximate Dec-POMDPs

We want our agent to:

- Make moves that are likely to lead it to the card.
- Change its behavior based on observations it receives.
- Respond to locative advice from the other player.
- Give locative advice to the other player.

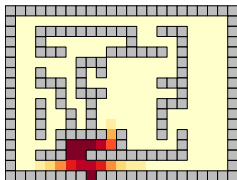
Modeling the problem as a POMDP allows us to train agents that have these properties.

Grounded language interpretation

“in the bottom you see the opening on the bottom row”



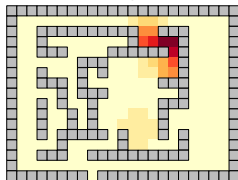
BOARD(entrance & bottom); $H: 5.48$



“in the top right of the middle part of the board”



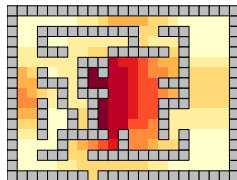
middle(top & right); $H: 5.27$



“i'm in the center”



BOARD(middle); $H: 7.37$



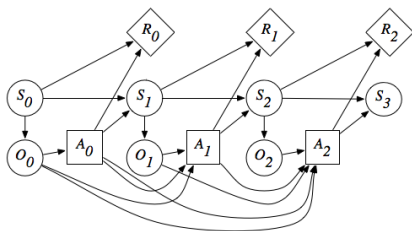
POMDPs

The agent has only probabilistic information about its current state (and the effects of its actions are non-deterministic, as in MDPs).

Definition (POMDP)

A POMDP is a structure (S, A, R, T, Ω, O) :

- (S, A, R, T) is an MDP.
- Ω is a finite set of observations.
- $O : (A \times S \times \Omega) \mapsto [0, 1]$ is the observation function.



ListenerBot (a POMDP agent)

- S : all combinations of the player's region and the card's region
- b_0 : initial belief state (distribution over S)
- A : *travel* actions for each region, and a single *search* action
- Ω : {AS seen, AS not seen}
- Σ : a set of messages, treated as observations; each message σ denotes a distribution $P(s | \sigma)$ over states s . We apply Bayes rule to incorporate these into the POMDP observations.
- T : distributions $P(s' | s, a)$, except *travel* actions fail between nonadjacent regions
- O : distributions $P(o | s, a)$; *travel* actions never return positive observations; *search* actions return positive observations only if the player's current region contains the AS
- R : small negative for not being on the card, large positive for being on it. No sensitivity to the other player.

Optimization

A belief state for (S, A, R, T, Ω, O) is a probability distribution b over S .

$$P(s, a, o, b) = O(s, a, o) \sum_{s' \in S} T(s', a, s) b(s') \quad (1)$$

$$b_o^a(s) = \frac{P(s, a, o, b)}{\sum_{s' \in S} P(s', a, o, b)} \quad (2)$$

Definition (Bellman operator for POMDPs)

Let b be a belief state for (S, A, R, T, Ω, O) . Set $\mathcal{P}_0(b') = 0$ for all belief states b' . Then for all $t > 0$:

$$\mathcal{P}_t(b, a) = \left(\sum_{s \in S} b(s) R(s, a) \right) + \gamma \sum_{o \in \Omega} \left(\sum_{s \in S} P(s, a, o, b) \right) \mathcal{P}_{t-1}(b_o^a)$$

where $0 < \gamma \leq 1$ is a discounting term.

Approximate solutions take us (only) part of the way

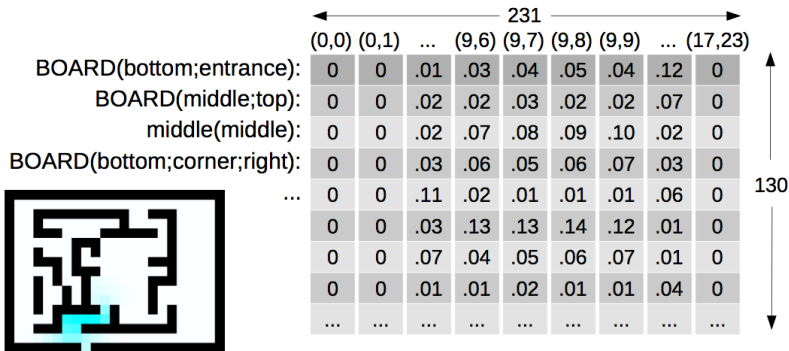
- An exact solution specifies the value of every action at any reachable belief state.
- In practice, only approximate solutions are tractable. We used the PERSEUS solution algorithm (Spaan and Vlassis 2005).
- Even approximate solutions are generally only possible for problems with $< 10K$ states.

Card location		Agent location		Partner location		Partner's card beliefs
231	×	231	×	231	×	231
		≈ 50K		≈ 12M		≈ 3B

Table: Size of the state-space for the one-card game.

Language as a representation for planning

- Divide the board up into n regions, for some tractable n
- Generate this partition using our locative phrase distributions.
- k -means clustering in locative phrase space.



Clusters induced

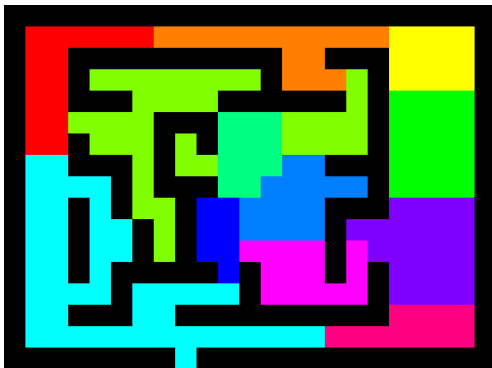


Figure: 12-cell clustering.

Clusters induced

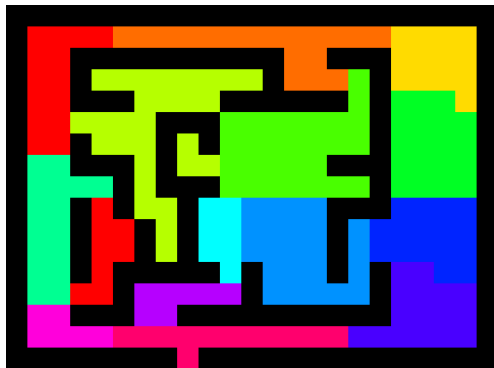


Figure: 14-cell clustering.

Clusters induced

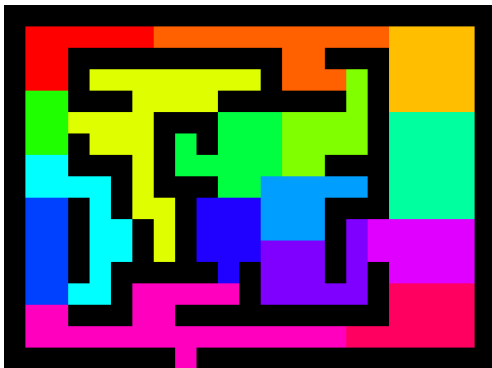


Figure: 16-cell clustering.

Clusters induced

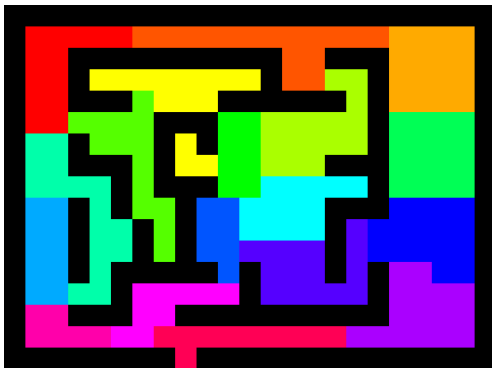
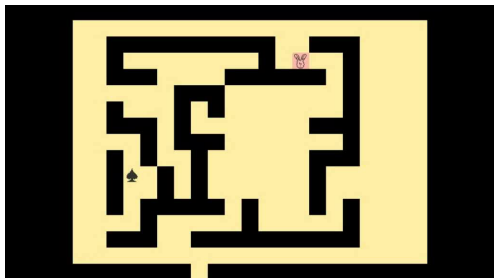


Figure: 18-cell clustering.

ListenerBot example



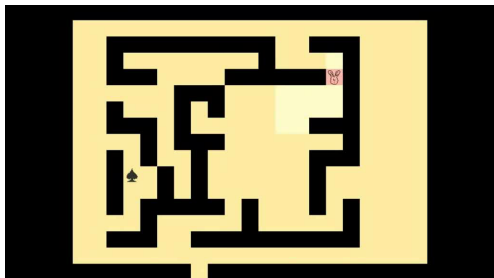
ListenerBot:



ListenerBot example



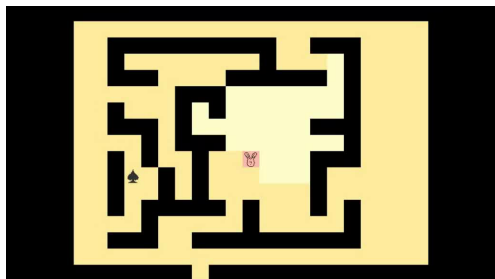
ListenerBot:



ListenerBot example



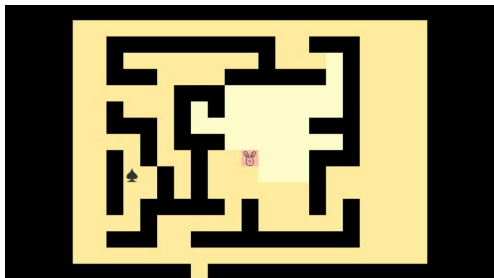
ListenerBot:



ListenerBot example



ListenerBot:



“it’s on the left side”



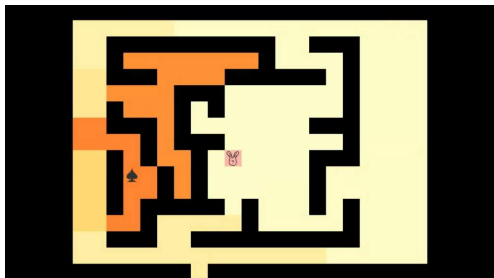
BOARD(left)



ListenerBot example



ListenerBot:

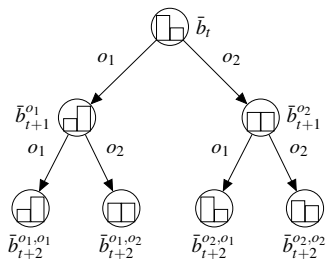


DialogBot (an approximate Dec-POMDP)

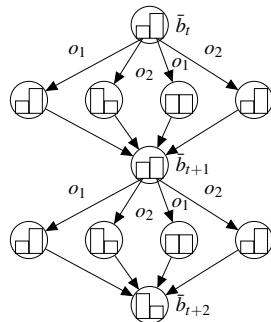
DialogBot is a strict extension of ListenerBot:

- The set of states is now all combinations of
 - both players' positions
 - the card's region
 - the region the other player believes the card to be in
- The set of actions now includes dialog actions.
- (The player assumes that) a dialog action U alters the other player's beliefs in the same way that U would impact his own beliefs.
- Same basic reward structure as for Listenerbot, except now also sensitive to whether the other player has found the card.

Belief-state approximation

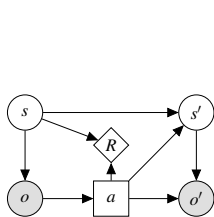


(a) Exact multi-agent belief tracking

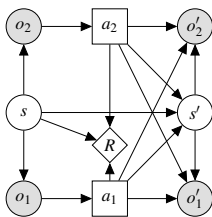


(b) Approximate multi-agent belief tracking

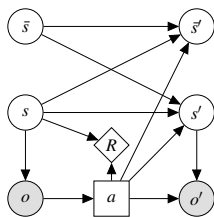
How the agents relate to each other



(a) ListenerBot POMDP



(b) Full Dec-POMDP

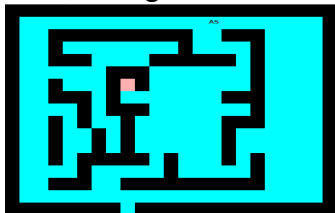


(c) DialogBot POMDP

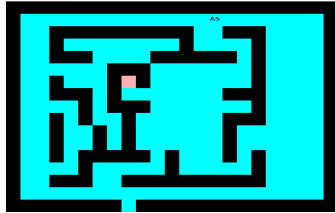
Figure: In the full Dec-POMDP (b), both agents receive individual observations and choose actions independently. Optimal decision making requires tracking all possible histories of beliefs of the other agent. DialogBot approximates the full Dec-POMDP as single-agent POMDP. At each time step, DialogBot marginalizes out the possible observations \bar{o} that ListenerBot received, yielding an *expected belief state* \bar{s} .

DialogBot and ListenerBot play together

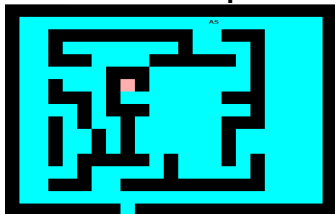
DialogBot beliefs



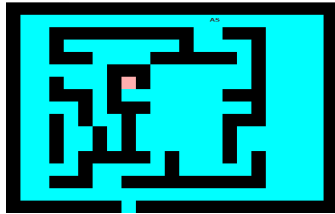
ListenerBot beliefs



DialogBot beliefs: ListenerBot's position

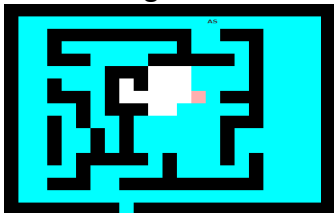


DialogBot beliefs: ListenerBot's beliefs

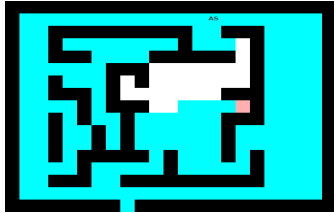


DialogBot and ListenerBot play together

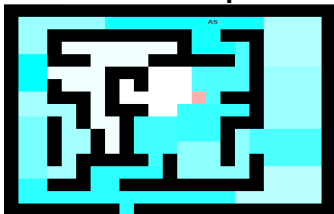
DialogBot beliefs



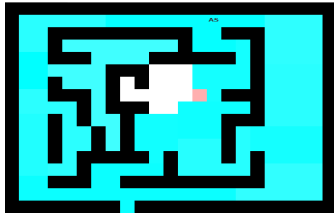
ListenerBot beliefs



DialogBot beliefs: ListenerBot's position



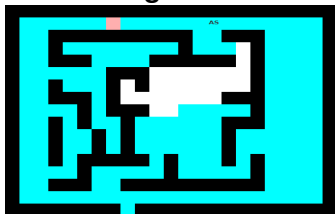
DialogBot beliefs: ListenerBot's beliefs



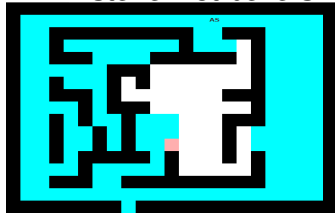
DialogBot and ListenerBot play together

Dialogbot: "Top"

DialogBot beliefs



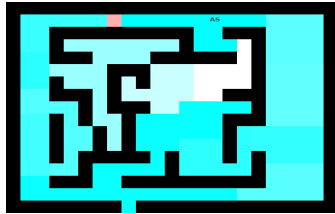
ListenerBot beliefs



**DialogBot beliefs:
ListenerBot's position**



**DialogBot beliefs:
ListenerBot's beliefs**



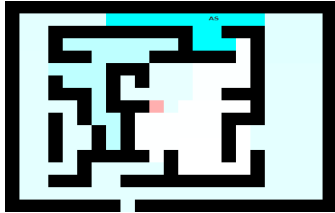
DialogBot and ListenerBot play together

Dialogbot: "Top"

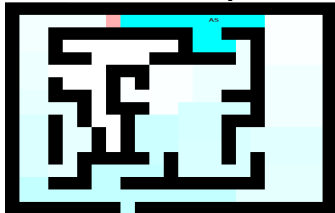
DialogBot beliefs



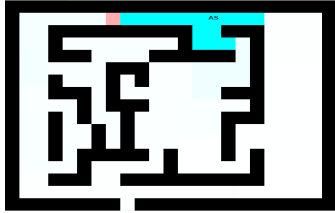
ListenerBot beliefs



**DialogBot beliefs:
ListenerBot's position**



**DialogBot beliefs:
ListenerBot's beliefs**



DialogBot and ListenerBot play together

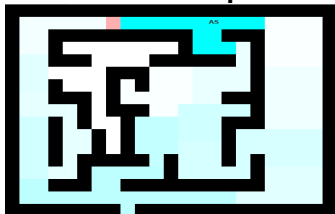
DialogBot beliefs



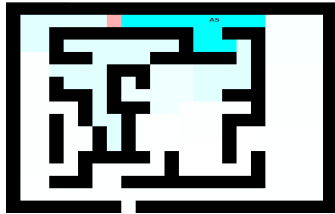
ListenerBot beliefs



DialogBot beliefs: ListenerBot's position

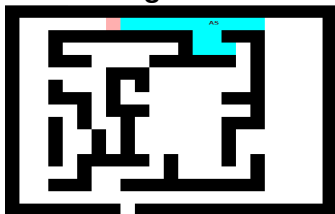


DialogBot beliefs: ListenerBot's beliefs

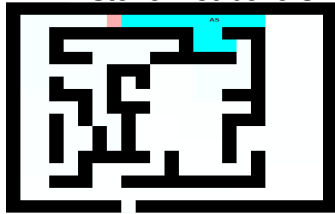


DialogBot and ListenerBot play together

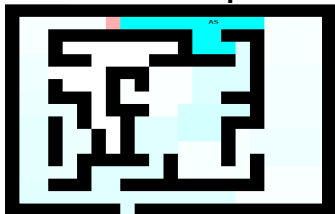
DialogBot beliefs



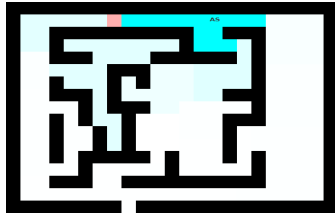
ListenerBot beliefs



DialogBot beliefs: ListenerBot's position

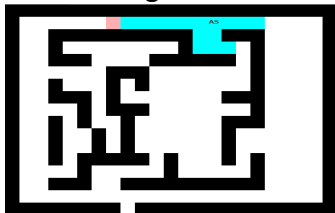


DialogBot beliefs: ListenerBot's beliefs

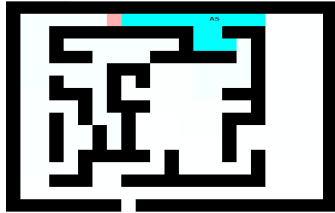


DialogBot and ListenerBot play together

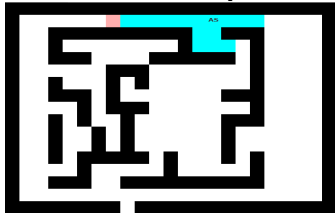
DialogBot beliefs



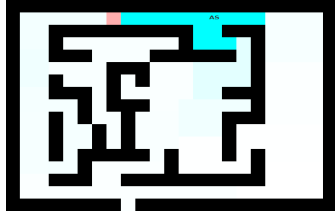
ListenerBot beliefs



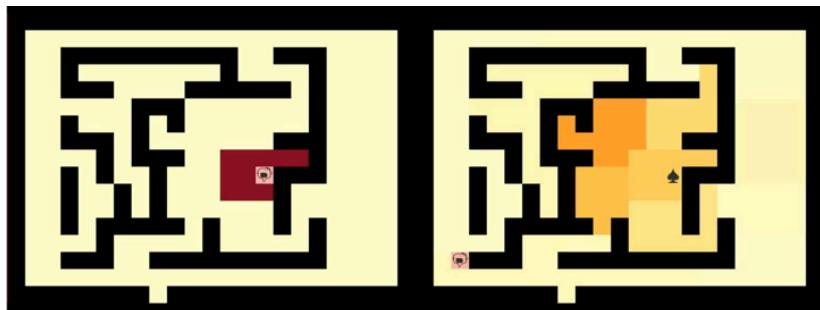
DialogBot beliefs: ListenerBot's position



DialogBot beliefs: ListenerBot's beliefs

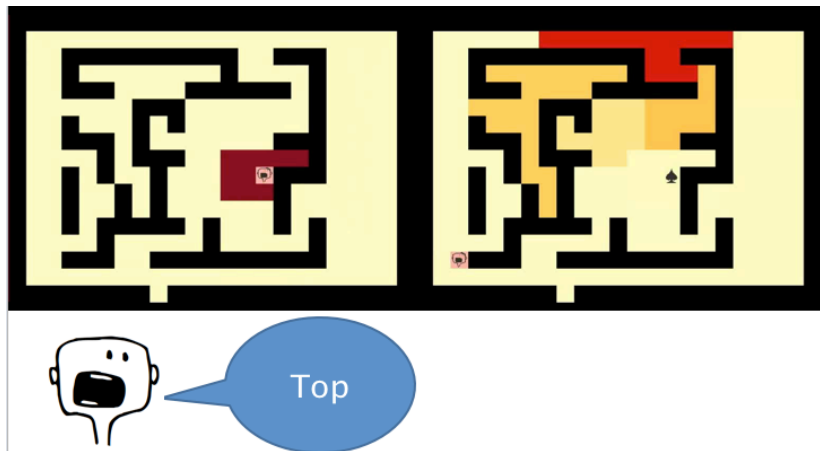


Grown-up DialogBots (a week of policy exploration)



Middle
of the
board

Baby DialogBots (a few hours of policy exploration)



Experimental results

Agents	Success	Average Moves
ListenerBot & ListenerBot	84.4%	19.8
ListenerBot & DialogBot	87.2%	17.5
DialogBot & DialogBot	90.6%	16.6

Table: The evaluation for each combination of agents. 500 random initial states per agent combination. It pays to model other minds!

Emergent pragmatics

Quality

- The Gricean maxim of quality says roughly “Be truthful”.
- For DialogBot, this emerges from the decision problem: false information is (typically) more costly.
- DialogBot would lie if he thought it would move them toward the objective.

Quantity and Relevance

- The Gricean maxims of quantity and relevance for informative, timely contributions.
- When DialogBot finds the card, he communicates the information, not because he is hard-coded to do so, but rather because it will help the other player find it.

Conclusion

- 1 Natural language is situated
- 2 Reasoning about other minds
- 3 Natural language as social
- 4 Examples of grounded NLU systems
- 5 Decision theoretic NLU agents
- 6 Conclusion**

Grounding your own NLU systems

There are many kinds of grounding, and even a little bit of grounding can help. Here are a few ideas for systems that aren't designed specifically for grounded language understanding:

- Retrofit word vectors with information from a social graph or from the environment.
- Connect to a knowledge graph for symbolic reasoning.
- Write feature functions that mix linguistic information with features from the environment.
- Think about how your system, and your data, can be seen from speaker and listener perspectives.
- Learn embeddings for non-linguistic entities and combine them with linguistic embeddings.

Corpus resources

- **SwDA**: <http://www.stanford.edu/~jurafsky/ws97/>
- **SwDA with Treebank3 alignment**:
<http://compprag.christopherpotts.net/swda.html>
- **Edinburgh Map Corpus**:
<http://groups.inf.ed.ac.uk/maptask/>
- **TRIPS**:
<http://www.cs.rochester.edu/research/cisd/projects/trips/>
- **TRAINS**:
<http://www.cs.rochester.edu/research/cisd/projects/trains/>
- **Cards**: <http://CardsCorpus.christopherpotts.net/>
- **SCARE**:
<http://slate.cse.ohio-state.edu/quake-corpora/scare/>
- **The Carnegie Mellon Communicator Corpus**:
<http://www.speech.cs.cmu.edu/Communicator/>
- **Facebook negotiation corpus**
<https://github.com/facebookresearch/end-to-end-negotiator>

Frontiers

- Deeper integration with devices and the environment.
- More sophisticated reasoning about other agents and their goals.
- Better tracking of full dialogue history; improved discourse coherence.
- Approximate state representations to address very pressing scalability issues.

References I

- Allen, James F.; Nathanael Chambers; George Ferguson; Lucian Galescu; Hyuckchul Jung; Mary Swift; and William Taysom. 2007. PLOW: A collaborative task learning agent. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, 1514–1519. Vancouver, British Columbia, Canada: AAAI Press.
- Andreas, Jacob and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1173–1182. Association for Computational Linguistics. URL <http://aclweb.org/anthology/D16-1125>.
- Campbell-Kibler, Kathryn. 2007. Accent, (ing), and the social logic of listener perceptions. *American Speech* 82(1):32–64.
- Clark, Herbert H. and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22(1):1–39.
- Cohn-Gordon, Reuben; Noah D. Goodman; and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Human Language Technologies: The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics.
- Danescu-Niculescu-Mizil, Cristian; Robert West; Dan Jurafsky; Jure Leskovec; and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International World Wide Web Conference*, 307–317. New York: ACM.
- DeVault, David and Matthew Stone. 2007. Managing ambiguities across utterances in dialogue. In Ron Artstein and Laure Vieu, eds., *Proceedings of DECALOG 2007: Workshop on the Semantics and Pragmatics of Dialogue*.
- DeVault, David and Matthew Stone. 2009. Learning to interpret utterances using dialogue history. In *Proceedings of the 12th Conference of the European Chapter of the ACL*, 184–192. Athens, Greece: Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/E09-1022>.
- Doyle, Gabriel; Dan Yurovsky; and Michael C. Frank. 2016. A robust framework for estimating linguistic alignment in twitter conversations. In *Proceedings of the 25th International World Wide Web Conference, WWW '16*, 637–648. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. doi:\bibinfo{doi}{10.1145/2872427.2883091}. URL <https://doi.org/10.1145/2872427.2883091>.
- Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12(4):453–476.
- Frank, Michael C. and Noah D. Goodman. 2014. Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology* 75(1):80–96. doi:\bibinfo{doi}{doi:10.1016/j.cogpsych.2014.08.002}.
- Frank, Michael C.; Joshua B. Tenenbaum; and Anne Fernald. 2012. Social and discourse contributions to the determination of reference in cross-situational word learning. *Language, Learning, and Development* .

References II

- Golland, Dave; Percy Liang; and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 410–419. Stroudsburg, PA: ACL. URL <http://www.aclweb.org/anthology/D10-1040>.
- Goodman, Noah D. and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences* 20(11):818–829. doi:\bibinfo{doi}{10.1016/j.tics.2016.08.005}.
- Levesque, Hector J. 2013. On our best behaviour. In *Proceedings of the Twenty-third International Conference on Artificial Intelligence*. Beijing.
- Levinson, Stephen C. 2000. *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. Cambridge, MA: MIT Press.
- Mao, Junhua; Jonathan Huang; Alexander Toshev; Oana Camburu; Alan L. Yuille; and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20. IEEE.
- Niederhoffer, Kate G. and James W. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology* 21(4):337–360.
- Potts, Christopher. 2012. Goal-driven answers in the Cards dialogue corpus. In Nathan Arnett and Ryan Bennett, eds., *Proceedings of the 30th West Coast Conference on Formal Linguistics*, 1–20. Somerville, MA: Cascadilla Press.
- Spaan, Matthijs T. J. and Nikos Vlassis. 2005. Perseus: Randomized point-based value iteration for POMDPs. *Journal of Artificial Intelligence Research* 24(1):195–220.
- Srivastava, Sameer B.; Amir Goldberg; V. Govind Manian; and Christopher Potts. 2016. Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. *Management Science* .
- Tellex, Stefanie; Ross A. Knepper; Adrian Li; Thomas M. Howard; Daniela Rus; and Nicholas Roy. 2014. Asking for help using inverse semantics. In *Proceedings of Robotics: Science and Systems*. doi:\bibinfo{doi}{10.15607/RSS.2014.X.024}.
- Vedantam, Ramakrishna; Samy Bengio; Kevin Murphy; Devi Parikh; and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. *arXiv:170102870* .
- West, Robert; Hristo S. Paskov; Jure Leskovec; and Christopher Potts. 2014. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics* 2(2):297–310.
- Winograd, Terry. 1972. Understanding natural language. *Cognitive Psychology* 3(1):1–191.

References III

Winograd, Terry. 1986. A procedural model of language understanding. In Barbara J. Grosz; Karen Sparck-Jones; and Bonnie Lynn Webber, eds., *Readings in Natural Language Processing*, 249–266. San Francisco: Morgan Kaufmann Publishers Inc.