

The Pragbot corpus

Christopher Potts
Stanford Linguistics

SUBTLE meeting, Penn, Nov 4, 2010



Overview

- 1 The Pragbot scenario and data-gathering tool.
- 2 The Pragbot corpus
(439 transcripts; 12,280 utterances; 192,039 events).
- 3 Experiments: Towards a cooperative bot.

Gameboard

TYPE HERE

Yellow boxes mark cards in your line of sight.

You are on 2D

Task description: Six consecutive cards of the same suit

Received: hi
Sent: I have the JH
Received: I have the SH

Type text here:
Disable Sound

I'm on 2D, which isn't too useful. There are cards to my right and below, though. I'll check them out.

Gather six consecutive cards of a particular suit (decide which suit together). Each of you can hold only three cards at a time, so you'll have to coordinate your efforts. You can talk all you

P1 turns remaining: 546
P2 turns remaining: 599

Indicate Task Complete

up
Click a card to pick it up:
2D

left
Click a card to drop it from your hand:
JH
right

down

The cards you are holding

Move with the arrow keys or these buttons.

Scenario

Gather six consecutive cards of a particular suit (decide which suit together). Each of you can hold only three cards at a time, so you'll have to coordinate your efforts. You can talk all you want, but you can make only a limited number of moves.

Transcripts

Server, 0: TASK_COMPLETED2010-06-13 01:01:02

Server, 0: PLAYER_1A10BNPQ9TFS88E

Server, 0: PLAYER_2A253Q11TZPQPIZ

Server, 56: MAX_LINEOFSIGHT3

Server, 118: CREATE_ENVIRONMENT NEW_SECTION

```

-----;          1,2:2D;1,7:KH;1,7:9S;1,11:6C;1,13:QC;1,14:QS;
-                -;          2,18:3H;2,18:9H;
- ----- - - - -;          3,19:4H;4,8:AC;4,19:3D;
- - - - - - - - - -;          4,19:KD;
- ---- - - - - - -;          5,14:QH;5,15:5S;5,15:2S;5,16:4D;5,16:10C;5,18:4
- - - - - - - - - -;          6,11:KC;6,15:9C;
- b - - - - - -;          7,11:2H;7,13:7S;
- - - - - - - - - -;          8,2:QD;8,4:AD;8,11:JC;8,20:8S;
- - - - - - - - - -;          9,9:10S;9,9:6H;9,9:8C;9,10:7H;9,14:JS;
- - - b - - - - -;          10,1:2C;10,10:8D;11,14:6D;11,14:10H;
- - - - - - - - - -;          11,18:4C;11,18:9D;
- - - - - - - - - -;          12,10:3S;12,12:6S;12,16:5H;12,16:JD;12,20:3C;
- - - - - - - - - -;          13,4:5C;13,4:JH;13,15:KS;
- - - - - - - - - -;          14,2:5D;14,20:10D;15,2:AH;
- - - - b-----;          15,13:7D;15,15:8H;15,17:AS;15,20:7C;
-                -;
----- - - - - -;          Server, 118: MAX_CARDS3
----- - - - - -;          Server, 118: GOAL_DESCRIPTION [...]

```

Transcripts

Player 1, 566650: PLAYER_MOVE7,11

Player 2, 567771: CHAT_MESSAGE_PREFIXwhich c's do you have again?

Player 1, 576500: CHAT_MESSAGE_PREFIXi have a 5c and an 8c

Player 2, 577907: CHAT_MESSAGE_PREFIXi jsut found a 4 of clubs

Player 1, 581474: PLAYER_PICKUP_CARD7,11:8C

Player 1, 586098: PLAYER_MOVE7,10

Sample run

START ← BACK FORWARD → FINISH (keyboard navigation: ← back | → forward | ↓ start | ↑ finish)

PragBot gameboard

The gameboard is a 10x10 grid with a black maze. Red squares are located at (3,2), (5,4), (6,4), and (8,4). A cyan square is at (9,9). The maze is composed of thick black lines forming a complex path through the grid.

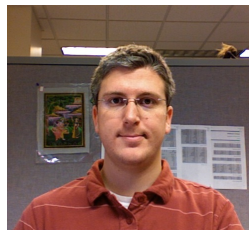
Player 1 holding: 8S, 7S, 5S
 Player 2 holding: 6S, 9S, 10S
 Current event Player 1, 902946: CLOSE_SOCKETS

1. Player 1 (36836): OK, we need to choose a suit - spades, etc. Any preference?
2. Player 2 (63662): I just got here. Doesn't matter to me.
3. Player 1 (73356): Ok, how about spades?
4. Player 2 (79177): Sounds good.
5. Player 1 (98577): Use arrow keys to move around maze, look for yellow boxes that have S
6. Player 1 (122058):
7. Player 1 (136792): Picked up 8 S
8. Player 2 (143520): Great!
9. Player 1 (170895): We need consecutive: 5,6,7, or 9,10,11
10. Player 1 (286170): picked up 10 Spades; we need the 9S
11. Player 2 (302728): I haven't found any yet. I'm still looking.
12. Player 2 (350053): I keep getting red squares. Do you know what they mean?
13. Player 1 (362998): No, I was wondering.
14. Player 2 (538526): I have a 6 and 9S
15. Player 1 (552754): picked up 7 spades. I can only hold 3 cards at a time.
16. Player 2 (568965): So what do we have now?
17. Player 1 (570774): Wow!
18. Player 1 (588214): We either need 9S or 5S
19. Player 2 (624182): OK. I just saw the 2 and 3.
20. Player 2 (630221): I'll keep going.
21. Player 1 (688140): 5S is at lower right corner - I can't pick up
22. Player 1 (711580): Should I drop 10?
23. Player 2 (711668): Outside the maze?
24. Player 1 (730054): Our mazes might be different. Maybe only MY maze has it at lower right. I'll drop 10S
25. Player 1 (787180): Ok - my cards are 5, 7, and 8 of spades
26. Player 2 (808180): I have 6 & 9.
27. Player 2 (818921): I can grab 10
28. Player 2 (853741): Will that be it?
29. Player 2 (850685): So now I have 6, 9, 10. What do you have?
30. Player 1 (855116): Terrific! I think so -- right? I'm 5 and 7 and 8
31. Player 2 (865446): Yeah! That's it! Thanks.

Done

The Pragbot platform

Extensible Java program developed by Karl Schultz. Handles high traffic well. Intuitive transcript design and helpful logging. Plays nicely with the outside world.



- Specify the task (or task family).
- Design the map (simple text format).
- Set all high-level contextual parameters (line of sight, max moves, max cards, hidden walls).
- Two humans, or one human and one bot.

Data collection

- Data collection in June 2010.
- PHP wrapper to Pragbot written by Victoria Schwanda.
- Server-side configuration by Chriz Czyzewicz.
- Players recruited via Amazon's Mechanical Turk.
- Payment: \$1.00 per player per game, with occasional \$0.50 bonuses to especially thoughtful or competent players.
- 439 good transcripts (out of 479 in all)
- Collection time: 5 batches each lasting about 5 hours, spread out over two work weeks.
- At peak times: 30 transcripts per hour.
- Total cost: about \$1,000

Email feedback from our Turkers

That was actually a pretty fun hit.

The game with chat was great and like to see more HITs from you.

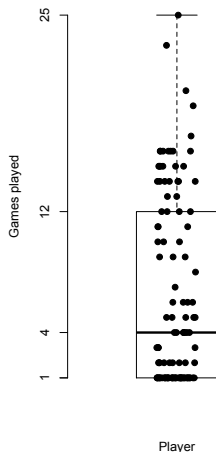
These HITs were really enjoyable. Hopefully you will put more on the site. You state that we can keep doing them, but right now if I click on your HIT, it tells me there are no more available for me. Is there something I can do to try again? Thanks.

I waited 1.22 before someone showed up. They never talked to me and didn't finish the job before leaving. Am I still out because they didn't cooperate?

Basic corpus stats

- 439 transcripts
- 111 unique players
- Game length mean: 465 actions (median 392, sd 263)
- Actions:
 - Card pickup: 8,330
 - Card drop: 6,105
 - Move: 175,503
 - Utterance: 12,280
 - Utterance length mean: 5.28 words (median 4, sd 4.78)
 - Total word count: 64,900
 - Total vocabulary: 3,149 (stemmed and with card-reference regularization: 2,255)

Expert effects



- The more a person played, the fewer utterances they used. This is true regardless of whether their partner was also experienced.
- If both players were experienced, the effect was even more dramatic.
- Expert transcripts were not necessarily shorter, though; some experts exhaustively searched independently, gathered subsets of the cards, and then assessed what they had found.

Novice strategy

Player 1: Hello. Are you here?

Player 2: yes

Player 2: do you see any cards

Player 1: Yes. I see a yellow spot. Those are our cards. We'll only be able to see the ones that are in our view

Player 1: until we move with our arrows.

Player 2: i see 3 of them

Player 1: We only have a certain number of moves, so we should decide how we're going to do this before we use them, do you think?

Player 2: sure

Player 1: Ok. So, we have to pick up six cards of the same suit, in a row...

Player 1: each of us can hold three, so...

Player 1: I think I should get my three, then you should get your three or vice versa

Player 2: ok

Player 2: you go ahead

Player 1: What suit should we do?

Player 1: And which six cards do you want to try for?

Player 2: whatever you want

Player 1: I'm Courtney, by the way- nice to meet you.

Player 2: i'm becky....nice to meet you too

Player 1: Hi Becky. How about we go for hearts? And take 234567
[...]

Expert strategy

Player 2: hi

Player 1: hi--which side r u on?

Player 2: right side

Player 2: u?

Player 1: left/middle

Player 1: ok i gathered everything in my area

Player 2: i think i have all of them also

Player 1: how bout 5C - 10C?

Player 2: ok

Player 1: i have 5C, 8C, 9C, and you should have 6C, 7C, 10C

Player 2: got them

A variation: Some games are impossible

Pragbot gameboard

START ← BACK FORWARD → FINISH (keyboard navigation: ← back | → forward | ↓ start | ↑ finish)

Player 1 holding:
Player 2 holding:
Current event Player 2, 1230: PLAYER_INITIAL_LOCATION3,13

Done

A variation: Some games are impossible

Praghot gameboard

file:/// - Praghot gameboard

START ← BACK FORWARD → FINISH (keyboard navigation: ← back | → forward | start | ↑ finish)

Player 1 holding: AS, QS
Player 2 holding: KH
Current event Player 1, 342025: CLOSE_SOCKETS

- Player 2 (13373): hey?
- Player 1 (16325): what suit?
- Player 1 (22079): S?
- Player 2 (32299): sure
- Player 1 (42813): i have QS and AS
- Player 2 (44381): 9s
- Player 1 (50474): ?
- Player 2 (51965): JS
- Player 2 (54381): JS here
- Player 2 (81960): like the numbers that are 9 and up =)
- Player 1 (95047): yeah, im trapped
- Player 2 (112546): are you inside?
- Player 2 (133664): there is a new rule
- Player 1 (137018): yeah, red blocks trap me?
- Player 1 (139805): what is it?
- Player 2 (147326): theres how i did last game
- Player 2 (150256): hems*
- Player 2 (160799): just try to find three similar cards in there
- Player 1 (160923): what??
- Player 2 (178833): wheter it be like 9D (D, AD
- Player 2 (194003): thum i find the other 3 out here
- Player 2 (206533): get it?
- Player 1 (224363): yeah but there isnt 3
- Player 2 (252950): gather up all the cards in there and put them in a coner, like put the S in one space, the D in another pace
- Player 2 (261980): space*
- Player 2 (273934): it helps a lot with organization
- Player 1 (308359): there is two of every suit
- Player 1 (312080): thats it
- Player 2 (315740): alright
- Player 2 (321549): that means its impossible
- Player 1 (331875): yeah, appears so
- Player 2 (333305): are we gonna determine this to be unsolvable?
- Player 1 (337743): yes

Done

Annotations

Syntax

Card-reference regularization ⇒

Stanford parses (Penn-style and dependency)

Pragmatics (2 annotators/transcript, resolving all issues at each stage before moving to the next)

- 1 Speech act boundaries.
- 2 Tagging with a subset of the Switchboard Dialog Act tags.
- 3 Tagging for which high-level discourse issue the speech-act engages.

Tags

- Questions
 - qy Y/N question (includes declarative questions)
 - qw Wh- question (includes declarative Wh- questions)
 - qo Other questions (when question but none of the above, including backchannel, rhetorical, tag, echo)
- Answers
 - ny Yes/Accept (includes affirmative non-yes, partial accept)
 - nn No/Reject (includes affirmative non-no, partial reject, displays of dispreference)
 - bk Response Acknowledgement (includes ok, uh-huh, etc.)
 - no Other answers (when answer but none of the above)
- ad Directive (includes imperatives and indirect suggestions)
- Politeness
 - fp Conventional opening
 - fa Apology
 - fc Conventional closing
 - ft Thanking
 - fo Other politeness (when polite but none of the above)
- o Iff none of the above options is appropriate

High-level discourse issues

- General issues about the world or the nature of Pragbot
 - FAMILIARITY: The players' familiarity with the game
 - GAME TYPE: How the game is played in general, the nature of the board, controllers, etc.
 - WORLD: Anything not about this particular game or how the game is played
- Cards and strategies
 - WHICH CARDS: Which cards to go for that not involving a SUIT or SEQUENCE strategy
 - WHICH SUIT: Which suit to go for that not involving a SEQUENCE strategy
 - WHICH SEQUENCE: Which sequence to go for not presupposing a SUIT
 - CARD DETAILS: Others issues about specific cards
- Game-internal player states
 - LOCATION: Any issue about the location of a player
 - ACTION: Issues concerning what the players should do
 - HOLDING: Issues concerning what the players are holding

Experiments: Towards a cooperative bot

Our goal is to implement a bot that plays like a human (though perhaps more efficiently). To do this properly, we need models of the following decisions:

- 1 If I hear an utterance, how should I interpret it?
- 2 If I am looking at a card, should I pick it up?
- 3 If I am holding a card, should I keep it or put it down?
- 4 If the state of play changes, what (if anything) should I say to my partner in response?

Language in context

Our code library turns each transcript into a data structure that is intuitively a list of temporally-ordered states

$$(\text{context, event})$$

The context includes

- local information (the state of play at that point)
- historical information (the events up to that point)
- global information (limitations of the game, the task, etc.)

When the event is an utterance, we can interpret it *in context*.

This is what pragmatics is all about, but it is very rare to have a dataset that truly lets you do it.

Relevance and the task

Example (Gather six consecutive cards of a particular suit.)

2H ⇒

8H 9H 10H JH
QH
7H KH
6H AH
5H
4H 3H 2H

Relevance and the task

Example (Gather six consecutive cards of a particular suit.)

Context: We're holding {4H, 5H}

2H ⇒

8H 9H 10H JH
QH
7H KH
6H AH
5H 4H 3H 2H

Underspecified referential expressions

Goal

To use the (evolving) task to make educated guesses about what underspecified card-oriented nominals pick out.

Player 2: Look for 2.

Player 1: and the 3?

Hypothesis

For any nominal referring expression, the intended referent will be the one that is (i) consistent with the information specified; and (ii) would bring the players closest to a solution *given the cards they are holding in the context of utterance*.

Underspecified referential expressions

Goal

To use the (evolving) task to make educated guesses about what underspecified card-oriented nominals pick out.

Context: The players are holding {4H,KH}

Player 2: Look for 2.

Player 1: and the 3?

Hypothesis

For any nominal referring expression, the intended referent will be the one that is (i) consistent with the information specified; and (ii) would bring the players closest to a solution *given the cards they are holding in the context of utterance*.

Annotations

Nominals referring to cards:

`[_FEATURES text]_{DENOTATION}`

- I have `[_3H 3H]_{3H}`
- Need `[_8 8]_{8H}`
- I'll drop `[_9|DEF|SG the 9]_{9H}`
- try `[_H|INDEF|PL h]_{2H, 3H, 4H, 5H, 6H, 7H, 8H, 9H, 10H, JH, QH, KH}`
- got `[_X|PRO|SG it]_{9H}`
- i'll look around to see if i can find `[_X|INDEF|PL any you can pick up]_{2H, 3H, 4H, 5H, 6H, 7H, 8H, 9H, 10H, JH, QH, KH}`

Results (for 10 transcripts)

Resolving the reference of the singular definite card references:

- [-(SUIT|RANK)DEF|SG the (SUIT|RANK)]_{CARD}
- [- (SUIT|RANK)]_{CARD}

Literal	103	(37%)
Requiring enrichment	172	(63%)
Total	275	
Correct inference	164	(95%)
Incorrect inference	8	(5%)
Total	172	

- The mistakes are localized: strings of card references that the system botches uniformly.
- Most inferences involve guessing the suit based on the rank, which is easier than guessing the rank based on the suit.

Pickup prediction

Goal

To build a model of how players decide whether to pick up a card.

Hypothesis

A player will decide whether to pick up a card based a number of interacting factors:

- Degree to which this card helps solve the problem, given current holdings.
- Current distance from a solution (when this high, players are more likely to pick up random cards)
- Was the card mentioned previously in the discourse?
- How many cards is the player currently holding?
- Does the card match the dominant suit of the player?
- What percentage of the way through the game are we?
- How many utterances have been made so far?

Model

8,326 pickup events and 14,276 non-pickup events (player could have picked up a card but didn't). Logistic regression classifier.

Predictor	Estimate
Intercept	-2.69
CardRelevance	0.53
PriorDistanceToSolution	0.62
CardMentioned	1.51
HoldingsLength	-0.49
MatchDominantSuit	0.94
UtterancesSoFar	-0.01
RelPosition	1.45

Table: Simple fitted model (all standard errors effectively 0).

Results

Extending the previous model with a number of interaction terms achieves 83.4% accuracy and the following effectiveness scores:

	Precision	Recall
Yes	0.75	0.82
No	0.88	0.84

We care more about recall (getting what is relevant) than precision (avoiding irrelevant stuff), because the cost of picking something up and then putting it down is lower than the cost of having to find something that you passed up earlier.

Looking ahead

Immediate plans

- Complete the annotations (by January 1).
- Leverage the annotations for increased accuracy and additional coverage of pragmatic issues.
- Similar analysis with the forthcoming Pragbot 2 data.
- Implement a bot based on the models we develop and use human partners to assess performance.

Questions for the group

- Additional data collection?
- New scenarios?
- Novel use cases for models that leverage context to interpret fragmentary utterances?