LING 120: Language and Computers

Semester: Fall ’17

Instructor: Sowmya Vajjala

Iowa State University, USA

01 November 2017

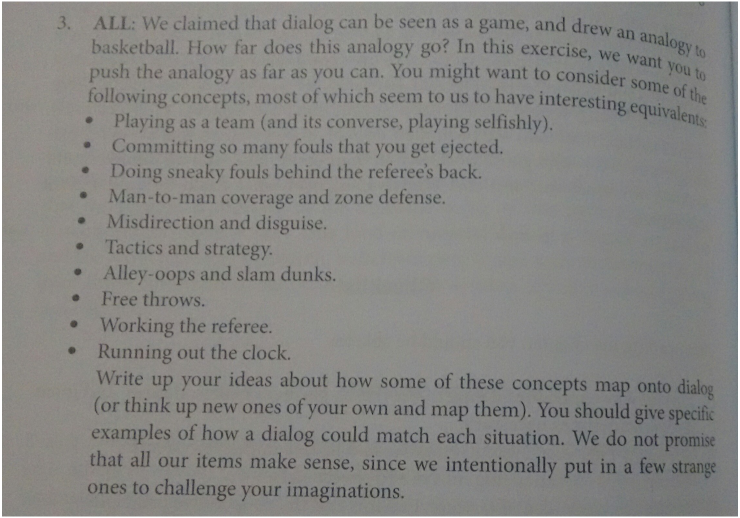
Outline for today

- Dialog Systems - Design (continued)
- Evaluating Dialog Systems
- Friday: Conclusion of dialog systems, and overview of Speech recognition.

Dialog Systems - What we discussed so far?

- What are dialog systems?
- Where are they useful?
- What are the different kinds of dialog systems?
- How are they created?
- What are the aspects of dialogues these systems should take care of while chatting with a user?

Monday’s exercise



3. **ALL:** We claimed that dialog can be seen as a game, and drew an analogy to basketball. How far does this analogy go? In this exercise, we want you to push the analogy as far as you can. You might want to consider some of the following concepts, most of which seem to us to have interesting equivalents:

   - Playing as a team (and its converse, playing selfishly).
   - Committing so many fouls that you get ejected.
   - Doing sneaky fouls behind the referee’s back.
   - Man-to-man coverage and zone defense.
   - Misdirection and disguise.
   - Tactics and strategy.
   - Alley-oops and slam dunks.
   - Free throws.
   - Working the referee.
   - Running out the clock.

   Write up your ideas about how some of these concepts map onto dialog (or think up new ones of your own and map them). You should give specific examples of how a dialog could match each situation. We do not promise that all our items make sense, since we intentionally put in a few strange ones to challenge your imaginations.

Summary of Monday's responses

- committing so many fouls you get ejected: giving rude responses until the other person quits the conversation.
- running out the clock: rambling about random things
- playing as a team: people should co-operate for smooth flow of conversation, lecture vs dialog
- doing sneaky fouls behind the referees: backhanded compliments
- mis-direction and disguise: misleading a conversation, lying with others in the conversation
- tactics: staying away from touchy subjects

Recap Questions

- How does Eliza work?

Recap Questions

- How does Eliza work?
- How can you add "Emotion" to Eliza, and make it show a variety of emotions such as happiness, sadness, anger etc?

Recap Questions

- How does Eliza work?
- How can you add "Emotion" to Eliza, and make it show a variety of emotions such as happiness, sadness, anger etc?
- What is a frame based dialog system?

Recap Questions

- How does Eliza work?
- How can you add "Emotion" to Eliza, and make it show a variety of emotions such as happiness, sadness, anger etc?
- What is a frame based dialog system?
- How does a frame based dialog system differ from a chat bot?

Recap Questions

- How does Eliza work?
- How can you add "Emotion" to Eliza, and make it show a variety of emotions such as happiness, sadness, anger etc?
- What is a frame based dialog system?
- How does a frame based dialog system differ from a chat bot?
- What is a Wizard of Oz simulation?

Another Example of Dialog System Development

Interactions - `https://www.interactions.com/` - came with the idea of human-assisted dialog systems

- The idea is to have a human in the loop to improve the accuracy of dialog systems.
- i.e., not fully automated, but automation with human support
- (does that sound like a step backward? or a case of sensible implementation?)

Adaptive Intelligence technology videos (`https://goo.gl/aTVFk3`, `https://goo.gl/LP9re7`)

Evaluation of Dialog Systems-1

- According to you, how should these systems be evaluated objectively?

Evaluation of Dialog Systems-1

- According to you, how should these systems be evaluated objectively?
- Let us say, I am using a bot to assist customers do airline bookings. How do I know if the bot is doing well?

Evaluation of Dialog Systems-1

- According to you, how should these systems be evaluated objectively?
- Let us say, I am using a bot to assist customers do airline bookings. How do I know if the bot is doing well?
- How do we evaluate a multi-purpose system like Siri?

Evaluation of Dialog Systems-1

- According to you, how should these systems be evaluated objectively?
- Let us say, I am using a bot to assist customers do airline bookings. How do I know if the bot is doing well?
- How do we evaluate a multi-purpose system like Siri?
- Even before the bot starts interacting with customers, how can I evaluate it during development?

Evaluation of Dialog Systems-1

- According to you, how should these systems be evaluated objectively?
- Let us say, I am using a bot to assist customers do airline bookings. How do I know if the bot is doing well?
- How do we evaluate a multi-purpose system like Siri?
- Even before the bot starts interacting with customers, how can I evaluate it during development?
- As it turns out, this question has no easy answer!

Evaluation of Dialog Systems-2

- Whether the dialog system manages to make a human feel he/she is chatting with another human

Evaluation of Dialog Systems-2

- Whether the dialog system manages to make a human feel he/she is chatting with another human
- Asking several human beta users to evaluate the system in terms of various features (speech recognition, response generation, language understanding, emotions, ability to converse etc)

Evaluation of Dialog Systems-2

- Whether the dialog system manages to make a human feel he/she is chatting with another human
- Asking several human beta users to evaluate the system in terms of various features (speech recognition, response generation, language understanding, emotions, ability to converse etc)
- If we do not have users yet: Simulated dialog between two computers instead of a human and a computer

Evaluation of Dialog Systems-2

- Whether the dialog system manages to make a human feel he/she is chatting with another human
- Asking several human beta users to evaluate the system in terms of various features (speech recognition, response generation, language understanding, emotions, ability to converse etc)
- If we do not have users yet: Simulated dialog between two computers instead of a human and a computer
- More straight forward in some scenarios such as slot-filling ones in frame-based dialog systems - we can look for slot error rate.

Evaluation of Dialog Systems-2

- Whether the dialog system manages to make a human feel he/she is chatting with another human
- Asking several human beta users to evaluate the system in terms of various features (speech recognition, response generation, language understanding, emotions, ability to converse etc)
- If we do not have users yet: Simulated dialog between two computers instead of a human and a computer
- More straight forward in some scenarios such as slot-filling ones in frame-based dialog systems - we can look for slot error rate.
- If possible response vocabulary is small, we can evaluate based on word-overlap between machine generated, and human generated response for the same question.

Evaluation of Dialog Systems-Recent Advances

Uses text classification, our previous topic!

- ADEM: Train a classifier on a set of responses in context, which are labeled as appropriate or inappropriate for the context by humans. You then use this classifier to evaluate responses of the dialog system.

Evaluation of Dialog Systems-Recent Advances

Uses text classification, our previous topic!

- ADEM: Train a classifier on a set of responses in context, which are labeled as appropriate or inappropriate for the context by humans. You then use this classifier to evaluate responses of the dialog system.
- Adversarial Evaluation: Train a classifier to distinguish between human and machine generated responses first. If the dialog system successfully fools this evaluator, it is a good system.

Note: This kind of work is very recent (above two examples are from July and September 2017 resp. !).

Evaluation of Dialog Systems-Recent Advances

Uses text classification, our previous topic!

- ADEM: Train a classifier on a set of responses in context, which are labeled as appropriate or inappropriate for the context by humans. You then use this classifier to evaluate responses of the dialog system.
- Adversarial Evaluation: Train a classifier to distinguish between human and machine generated responses first. If the dialog system successfully fools this evaluator, it is a good system.

Note: This kind of work is very recent (above two examples are from July and September 2017 resp. !). Something that is fresh-er (Today!): `https://www.wired.com/story/the-college-kids-doing-what-twitter-wont/`

Attendance Exercise

Go to the canvas forum, take a look at the uploaded file and work in groups of 2–3 people to answer the questions in the file.