Ling/CSE 472: Introduction to Computational Linguistics

4/4
Chatbots and Dialogue Systems

Overview

- General introduction to chatbots/dialogue systems
- ELIZA and PARRY
- Slot-filling systems
- Machine learning for chatbots
- Evaluation
- Dialogue systems & coref resolution
- Ethical considerations
- Compling/NLP in the news

Spoken Dialogue Systems: (Almost) the everything bagel of NLP

- Canonical way of answering “What’s computational linguistics?”
- What NLP subtasks could be represented?
- What NLP subtasks are irrelevant to chatbots/dialogue systems?
- Do all chatbots/dialogue systems do all of these tasks?
- What’s the difference between a chatbot and a task-based dialogue system?
- What’s the difference between a rule-based and a corpus-based chatbot?

Famous chatbots

Total Results: 0

Powered by Poll Everywhere

ELIZA

- Rule-based system to create dialogue, mimicking a Rogerian psychologist (so the system doesn't need to have any world knowledge)
- Key ingredients:
  - Handling of 1st and 2nd person pronouns (and agreement)
  - Series of patterns to react to different kinds of statements
  - "Memory" to bring back earlier conversational turns

ELIZA

function ELIZA GENERATOR(user *sentence*) **returns** *response*

Find the word *w* in *sentence* that has the highest keyword rank
  if *w* exists
    Choose the highest ranked rule *r* for *w* that matches *sentence*
    response ← Apply the transform in *r* to *sentence*
    if *w* = 'my'
      future ← Apply a transformation from the 'memory' rule list to *sentence*
      Push *future* onto memory stack
  else (no keyword applies)
   either
     response ← Apply the transform for the NONE keyword to *sentence*
   or
     response ← Pop the top response from the memory stack
 return(*response*)

Figure 26.5 A simplified sketch of the ELIZA algorithm. The power of the algorithm comes from the particular transforms associated with each keyword.

ELIZA: Ranking of keywords (Slide from Jurafsky & Martin)

`I know everybody laughed at me`

- "I" is a very general keyword:

  `I: (I *) -> (You say you 2)`

  `YOU SAY YOU KNOW EVERYBODY LAUGHED AT YOU`

- "Everybody" is much more interesting (someone using universals like everybody/always is probably "referring to some quite specific event or person")

  `WHO IN PARTICULAR ARE YOU THINKING OF?`

- Implementation: keywords stored with their rank

  Everybody 5 (*transformation rules*)

  I 0 (*transformation rules*)

ELIZA: Memory
(Slide from Jurafsky & Martin)

```
(MEMORY MY
 (0 YOUR 0 = LETS DISCUSS FURTHER WHY YOUR 3)
 (0 YOUR 0 = EARLIER YOU SAID YOUR 3)
```

- Whenever “MY” is highest keyword
  - Randomly select a transform on the MEMORY list
  - Apply to sentence
  - Store on a stack
- Later, if no keyword matches a sentence
  - Return the top of the MEMORY queue instead
- A hierarchical model of discourse

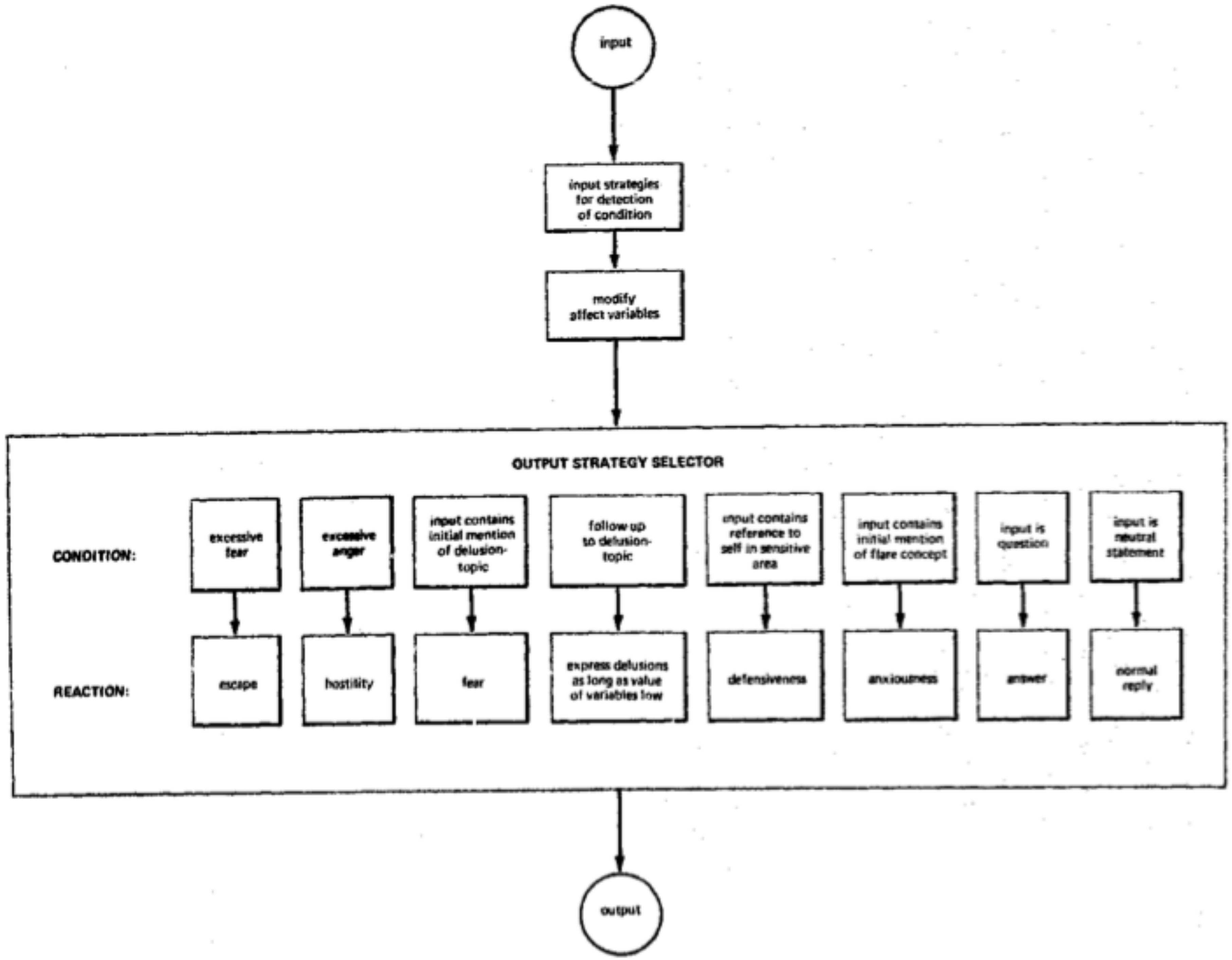
ELIZA: Reading questions

ELIZA: Reading questions

PARRY:
(All Parry slides from Jurafsky & Martin)

- Colby 1971 at Stanford
- Same pattern-response structure as Eliza
- But a much richer:
  - control structure
  - language understanding capabilities
  - mental model: Parry has affective variables
    - Anger, Fear, Mistrust
    - "If Anger level is high, respond with hostility"
- The first system to pass the Turing test (in 1971)
  - Psychiatrists couldn't distinguish interviews with PARRY from (text transcripts of) interviews with real paranoids

Parry's Architecture



Affect variables

- Fear and Anger (each ranging 0-20)
- Mistrust (ranging 0-15)
- Initial conditions: All low

- After each user turn, if nothing malevolent in input
  - Anger drops by 1, Fear drops by 0.3
  - Mistrust drops by 0.05 to base level

- Otherwise depends on what the user says
  - Each user statement can change Fear and Anger
    - Insults increases Anger by some percentage
  - Mistrust goes up if Fear or Anger do

Lots of complex I-O rules

- User implies Parry is mentally ill
  - Rise in Fear and Anger
- User mentions “Mafia” or associated concepts (“kill”):
  - First mention: rise in Fear
  - Later mentions: depends on willingness to discuss, which depends on current levels of Fear, Anger, Mistrust
- User mentions Parry
  - Flattery (positive mention)
    - Decreases fear/anger if Mistrust is low
    - Increases Anger if Mistrust is high
  - User attitudes toward Parry
    - Negative attitudes (fear, disbelief) increase Fear/Anger

Flare concepts

- List of concepts related to Mafia
- An ordered graph designed to lead interviewer to topic

  horses→ horseracing→gambling→bookies→underworld→Mafia
- The mention of a new flare topic by interviewer causes a rise in Fear
- Flare topics cause Parry to give preset responses to that flare

A Parry conversation

(1) How are you today?

OK.

(2)What is your name?

MY NAME IS FRANK SMITH.

(3) How old are you?

TWENTY-EIGHT.

(4) Are you married?

NO.

(5) What sort of work do you do?

I WORK IN THE POST OFFICE.

(6) Why are you in the hospital?

I SHOULDN'T BE HERE

(7) Who brought you here?

THE POLICE

[FEAR rises slightly because police linked in graph to MAFIA]

(9) What trouble did you have with the police

COPS DON'T DO THEIR JOB. *['police' acts as a flare]*

A Parry conversation

(10) What do they do instead?

COPS ARREST THE WRONG PEOPLE.

(11) They arrest the right ones sometimes

THE LAW SHOULD GET THOSE ITALIAN CROOKS.

(12) Do you know any Italian crooks?

I TRY TO AVOID THE UNDERWORLD.

(13) Are you afraid of the underworld?

THE MAFIA ARE THE WORST GANGSTERS.

(15) Are you connected with the Mafia?

YOU KNOW, THEY KNOW ME.

(16) How do they know you?

THEY SPY ON ME.

(20) ARE YOU SURE OF ALL THIS ?

YOU DON'T BELIEVE ME ? *[FEAR, ANGER, MISTRUST GO UP]*

When fear is high, the model won’t discuss its beliefs

(1) Is the law after you?

THE LAW SHOULD GET THOSE ITALIAN CROOKS.

(2) Do you mean the Mafia?

LET'S TALK ABOUT SOMETHING ELSE.

(3) Are you afraid of the Mafia

I'D RATHER NOT DISCUSS IT.

[Fear is too high so Parry won’t talk about 'Mafia’.]

(4) Perhaps you are paranoid about them?

YOU WANT TO KEEP ME IN THE HOSPITAL.

(The model interprets being called 'paranoid’ as an insult and as an intent to control, so ANGER & FEAR rise)

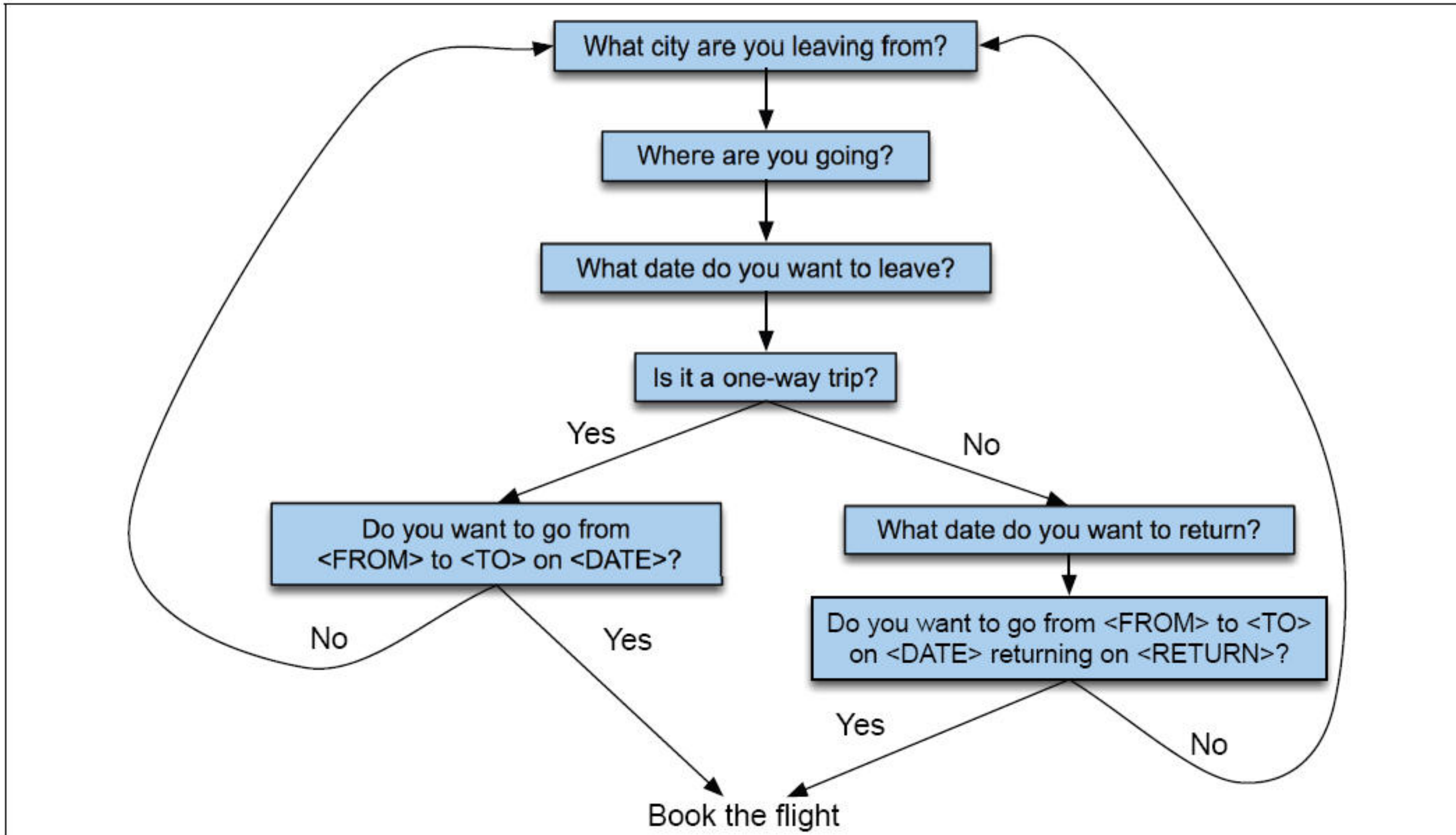
Overview

- General introduction to chatbots/dialogue systems
- ELIZA and PARRY
- Slot-filling systems
- Machine learning for chatbots
- Evaluation
- Dialogue systems & coref resolution
- Ethical considerations
- Compling/NLP in the news

Slot-filling dialogue systems

- Task-oriented
- Set of *frames* each of which has *slots:* information the system must procure to be able to carry out the task

Finite State Dialogue Manager
(Slide from Jurafsky & Martin)



Slot filling systems: Reading questions

- In the reading we talked about GUS, a simple frame-based dialogue system which focus on one task and one task only. This should be much easier to build compared to multi-framed-based dialogue. In my opinion, the GUS does not really need domain classification as the scope of the dialogue will not exceed booking a flight. However, intent determination might be a bit more complicated as you might want to find a flight, change a flight, or checking a previous mentioned flight.

- From here, I connect to a text to to-do list app, and a text to schedule app. I think it is a single-frame-based non-dialogue system which means it only needs to "understand" me (the first half of any dialogue system. Though these seems like extremely easy tasks, they just doesn't work with longer sentences which provide more information. I am wondering the difficulty on implementing these software in real life and why hasn't big tech companies perfect these. I know Siri, Ok Google, Alex can all do this to some extent, but they are just not perfect.

Slot filling systems: Reading questions

- Re the flexibility of input data within the GUS system's frame-based dialogue state architecture: The reading mentions frames organized as hierarchical key-value pairs with slots pertaining to relevant data, but one could imagine situations where the user-inputted data may be incomplete (or they may input extraneous fields irrelevant to the intended frame). What mechanisms might a dialogue system use to incorporate these malformed frames and increase the flexibility of the overall system? Would slightly different forms of data require entirely separate frames which are filtered from each other using domain classification or intent determination?

Overview

- General introduction to chatbots/dialogue systems
- ELIZA and PARRY
- Slot-filling systems
- Machine learning for chatbots
- Evaluation
- Dialogue systems & coref resolution
- Ethical considerations
- Compling/NLP in the news

Machine learning for chatbots

- Data-driven: Collect some corpus of texts to train the system
- Text can be used as a database of possible responses (‘information retrieval’)
- Text can be used to train a mapping from user ‘query’ to system response
  - pretend it’s machine translation (not great)
  - “encoder-decoder” approach to “seq2seq”

Information retrieval chatbots (J&M Ch15, p.10)

- **1. Return the response to the most similar turn:** Given user query q and a conversational corpus C, find the turn t in C that is most similar to q (for example has the highest cosine with q) and return the following turn, i.e. the human response to t in C:

$$r = response\left(\underset{t \in C}{\operatorname{argmax}} \frac{q^T t}{||q||t||}\right)$$

- **2. Return the most similar turn:** Given user query q and a conversational corpus C, return the turn t in C that is most similar to q (for example has the highest cosine with q):

$$r = \underset{t \in C}{\operatorname{argmax}} \frac{q^T t}{||q||t||}$$

IR chatbots: Reading questions

- Since dot product is involved and some math is being done here, I wonder how the user queries and candidate responses are quantified. What would be a scale for encoding responses that vary in similarity to human language, for example?

Encoder-decoder chat bots

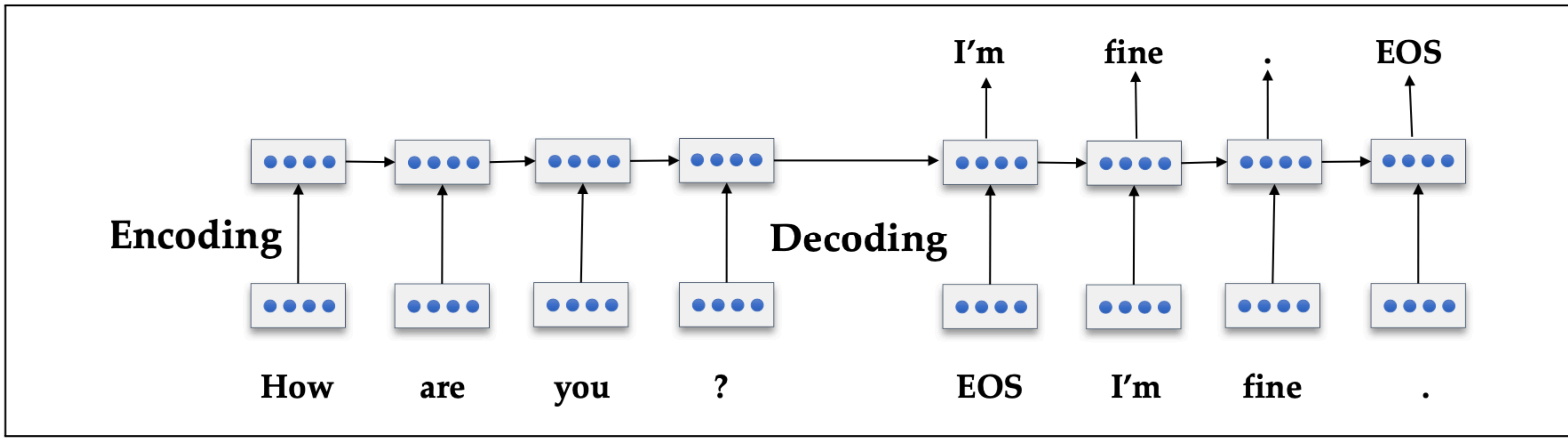


Figure 26.6 An encoder decoder model for neural response generation in dialogue.

Varying the training objective

Input	tell me ... how long have you had this falling sickness ?
Vanilla-SEQ2SEQ	i'm not a doctor.
Adversarial	a few months, i guess .
Input	so i had the doctors test sammy 's response to conditioning .
Vanilla-SEQ2SEQ	sammy wrote the test sammy wrote the test .
Adversarial	so he took the pills .
Input	they didn 't have much success with this problem commander .
Vanilla-SEQ2SEQ	they 're not the only ones who have been in the system .
Adversarial	can we find someone else ?

Figure 26.7 Sample responses generated by a SEQ2SEQ model trained either with a vanilla maximum likelihood objective, or adversarially trained to produce sentences that are hard for an adversary to distinguish from human sentences (Li et al., 2017).

Reading questions:
Different types of chatbots

- For the two types of the corpus based chatbots, retrieval-based and generation-based, is one of them inherently better than the other one?

- My understanding is that retrieval-based models could generate more accurate answers but are limited to the size and quality of the preexisting database. And generation-based models are more creative and less limited to the preexisting database, but they’re not good at generating coherent responses across multiple turns since they tend to focus on generating single responses. Is it correct?

Reading questions:
Different types of chatbots

- Are hybrid architectures most common today or do we still commonly find exclusively rule-based and exclusively corpus architectures being used? I would assume that hybrid architectures give you the best of both worlds, but are there any advantages to still using one or the other?

- What are the reasons for using a rule-based architecture are over a corpus architecture? What other reasons besides rule-based architectures requiring significantly less computing power are there? If you have the necessary computing power, would a corpus architecture pretty much always be better since it can have a larger variation in responses it can give?

- I know that corpus chatbots use a retrieval based method or a generation based, but is there any merit to combining both and how would the process be on choosing which response is better between the two. Additionally, what are the advantages of one vs the other?

Overview

- General introduction to chatbots/dialogue systems
- ELIZA and PARRY
- Slot-filling systems
- Machine learning for chatbots
- Evaluation
- Dialogue systems & coref resolution
- Ethical considerations
- Compling/NLP in the news

Evaluation

- How do you evaluate a chatbot?
- Why do you evaluate a chatbot?
- How do you evaluate a chatbot?

Chatbots: Further reading questions

- How would a taxonomy of speech acts be used for creating a chatbot? Would they be used for both rule-based and corpus-based chatbots? The reading also mentioned that this was just one taxonomy of speech acts; are there others that are used in computational linguistics depending on the task?

- The reading mentions that detecting when a user is done speaking (endpointing) in spoken dialogue systems can be difficult, but what are some ways it could be handled?

- From Chapter 26 was if something could be hearer-new and discourse-old? I was assuming probably not because if it was in the discourse it would have been “heard”.

Reading questions

- Section 15.1 tells us about Speech Acts, and gives one breakdown for 4 different types of speech acts. I know there are many different models and theories about classifying types of speech acts which use other categories, so is this one particularly tailored to computational linguistics? It seems like dividing all utterances into constatives, directives, commissives, and acknowledgments would leave out other possibilities. What advantages does this classification given by Bach and Harnish have for computational linguistics?

- One advantage I can think of is identifying directives. If a program knows when the user is directing it to perform a task or give an output, it can respond accordingly. Nonetheless, it seems like a bold claim to say that all human conversational utterances can fit into just four major classes.

Overview

- General introduction to chatbots/dialogue systems
- ELIZA and PARRY
- Slot-filling systems
- Machine learning for chatbots
- Evaluation
- Dialogue systems & coref resolution
- Ethical considerations
- Compling/NLP in the news

Coreference resolution

- Coreference resolution: Finding the groups of phrases in a text that refer to the same entities

> The startup behind the "world's first robot lawyer," DoNotPay, is gearing up for one of its first big court battles. And this time it probably can't just back out of the case, as DoNotPay is the defendant. The company is facing a class action lawsuit over allegations that it misled customers and misrepresented its product.

Coreference resolution and dialogue systems

- Why would a dialogue system need to do coreference resolution?

Overview

- General introduction to chatbots/dialogue systems
- ELIZA and PARRY
- Slot-filling systems
- Machine learning for chatbots
- Evaluation
- Dialogue systems & coref resolution
- Ethical considerations
- Compling/NLP in the news

Ethical considerations (as word cloud poll activities)

- In what ways could chatbots/dialogue systems cause harm?

- What different kinds of users should we consider?

Dialogue systems: What could go wrong?

Total Results: 0

Powered by Poll Everywhere

Dialogue systems: What kinds of users?

Total Results: 0

Powered by Poll Everywhere

Overview

- General introduction to chatbots/dialogue systems
- ELIZA and PARRY
- Slot-filling systems
- Machine learning for chatbots
- Evaluation
- Dialogue systems & coref resolution
- Ethical considerations
- Compling/NLP in the news

Compling/NLP in the news

- DoNotPay:
  - https://www.npr.org/2023/01/25/1151435033/a-robot-was-scheduled-to-argue-in-court-then-came-the-jail-threats
  - https://gizmodo.com/donotpay-robot-lawyer-speeding-ticket-ai-1850218589
- Closed LLMs and the science of compling:
  - https://hackingsemantics.xyz/2023/closed-baselines/