Aalto University

Conversational agents: chatbots and dialogue agents

Mikko Kurimo
SNLP lecture 8

Based on Chapter 15 in Jurafsky-Martin 3rd edition (version February 2024)

Image by Gracini Studios from Pixabay

Speech and Language Processing

An Introduction to Natural Language Processing,
Computational Linguistics, and Speech Recognition

Third Edition draft

Daniel Jurafsky
Stanford University

James H. Martin
University of Colorado at Boulder

https://web.stanford.edu/~jurafsky/slp3/

Draft of February 3, 2024. Comments and typos welcome!

Summary of Contents

Lecture schedule 2024

09 Jan 1 Introduction & course organization / Mikko Kurimo
16 Jan 2 Statistical language models / Mikko Kurimo
23 Jan 3 Sentence level processing / Mikko Kurimo
30 Jan 4 Word2vec / Tiina Lindh-Knuutila
06 Feb 5 Neural language modeling and LLMs / Mittul Singh
13 Feb 6 Morpheme-level processing / Mathias Creutz
20 Feb Exam week, no lecture
27 Feb 7 Speech recognition / Tamas Grosz
05 Mar 8 Chatbots and dialogue agents / Mikko Kurimo
12 Mar 9 Statistical machine translation / Jaakko Väyrynen
19 Mar 10 Neural machine translation / Sami Virpioja
26 Mar 11 LLMs in industry (prompting and in-context learning)/ Shantipriya Parida
02 April (spring break - no lecture)
09 Apr 12 LLM discussion and course conclusion / Mikko Kurimo
16 April Exam

See Mycourses for updates

Conversational agents have appeared in our phones and homes

Typing-based agents are starting to speak and listen in cars, robots, toys, phones, smart speakers and other devices

Content and goals for today

Content

1. Comparison of chatbots and dialogue agents
2. Rule- and corpus-based architectures
3. Information retrieval and machine learning based chatbots
4. Evaluation of chatbots
5. Chatting with large language models
6. Ethical issues

Goals

1. To know how the chatbots and dialogue agents work
2. To know how the chatbots are evaluated
3. To think about the ethical issues

Definitions

Chatbot:

- A system that **you can chat** with
- Discussion topics can be fixed, but there is **no specific goal** except for fun and keeping company

Dialogue agent:

- A system that helps you to **reach a specific goal** by giving and collecting information by **answering and asking questions**

In popular media both are often called chatbots, but here only the first one.

Properties of human conversation

- Speakers take turns: When to start? How?
  - system must allow users to interrupt
  - system should decide when the user has stopped talking
- **Speaker turns are acts**: greeting, thanking, answering, accepting, disagreeing, asking, ordering, promising etc.
- Speakers need to **ground** each other's acts: show that the speaker was understood: Why do elevator buttons light up?
- Conversations have **structure**: How acts follow each other?
- Some human conversations are lead by one person, but most have mixed **initiative:** e.g. interviewer should also answer questions

Discussion

1. Which chatbots and dialogue agents have **you** used?
2. What can they do, what not?

Comparison of chatbots and dialogue agents: the required operations

Chatbot

- Detect the discussion topic
- Ask typical questions
- React to human input, be coherent with previous turns
- World knowledge, persona

Dialogue agent

- Detect the user's intent
- Ask the required questions
- Parse and use human input

Chatbot architectures

Rule-based
- Pattern-action rules: Eliza (1966)
- Mental model: Parry (1971)

Corpus-based
- Information retrieval: Cleverbot, XiaoIce (Zhou et al., 2020)
- Neural encoder-decoder: BlenderBot (Roller et al. 2020)

Turing's test (1950) for machine intelligence: *Can you judge between a real human and a chatbot?*

ELIZA (Weizenbaum, 1966)

Try it out, e.g.

- *https://www.eclecticenergies.com/ego/eliza*
- *http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm*

Men are all alike.

IN WHAT WAY

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

Idea:

- ELIZA is a psychologist who reflects back what the patient says

Impact:

- People became emotionally involved
- People revealed very personal issues

How ELIZA works?

Pattern/transform rules

(YOU * ME) => (WHAT MAKES YOU THINK I * YOU)
e.g. "hate"

(I *) => (YOU SAY YOU *)
e.g. "know everybody laughs at me"

(MY *) => (EARLIER YOU SAID YOUR *)

ELIZA generator

- Look for certain keywords and select the best rule
- If the keyword is "my" select randomly some of the matching sentence from history
- If no keywords match, say simply: "Go on" or "I see"

PARRY (Colby, 1971)

Try it out:
- *https://www.botlibre.com/browse?id=857177*
- *https://www.chatbots.org/chatbot/parry/*

- Regular expressions as ELIZA
- Control structure
- Some language understanding
- Mental model: varying levels of fear, anger and mistrust

Note: The first system to pass a Turing test (in 1971): Psychiatrists could not distinguish interviews with PARRY from interviews with real paranoids

How Parry works?

Mental model

- Affective variables: anger, fear, mistrust
- For certain topics and keywords they start increasing or decreasing which then affects his responses

Parry's persona:

- 28-year-old single man
- no siblings and lives alone
- sensitive about his physical appearance, family, religion, education and sex.
- Hobbies: movies and gambling
- worried about mafia

When PARRY met ELIZA:
https://www.theatlantic.com/technology/archive/2014/06/when-parry-met-eliza-a-ridiculous-chatbot-conversation-from-1972/372428/

Corpus-based chatbots

- No hand-built rules
- Find responses from big data
- Based on:
  - Information retrieval
  - Machine learning

Typical corpora:

- Human-human conversations
- Human-machine conversations
- Transcriptions from ASR training data
- Movie subtitles
- Reddit.com
- Non-dialogue data, e.g. wikipedia
- Use a rule-based chatbot to collect human responses

IR-based chatbots

- Find the **most similar speaker turn** from the data
- Return the response for that
-
- Success depends on the data
- Garbage in, garbage out

- E.g. Cleverbot: http://www.cleverbot.com

Machine learning based chatbots

- Transducer from user’s turn to system’s turn
- Sequence-to-sequence learning
- Encoder-decoder model
- Transformers*, e.g. DialoGPT https://arxiv.org/abs/1911.00536*
- Improved cost function, e.g. *https://arxiv.org/abs/1510.03055*
- Improved decoding algorithm, e.g.*https://arxiv.org/pdf/1904.09751.pdf*
- Combining with IR, e.g. *https://arxiv.org/pdf/1808.04776.pdf*
- Alternative approach: fine-tune an LLM on conversational data, e.g. Chirpy Cardinal (Paranjape et al., 2020)
- **Common problems** with chatbots:
  - Lack of consistent personality
  - Lack of long-term memory
  - Boring answers like ”I don't know”

Lecture exercise: Try chatbots

Try chatbots and discuss these questions. To get points submit your notes in **MyCourses > Lectures > Lecture 8 exercise.**

1. Try ELIZA and/or PARRY. When do they fail? How to improve?
   - Eliza: https://www.eclecticenergies.com/ego/eliza
   - Eliza: http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm
   - Parry: https://www.botlibre.com/browse?id=857177
   - Parry: https://www.chatbots.org/chatbot/parry/
2. Try modern chatbots or dialogue agents. Where are they good and bad?
   - https://chat.openai.com/chat
   - https://convai.huggingface.co/
   - https://www.chatbots.org/

15 min work
+10 min break

ChatGPT

- Try it, e.g.: *https://chat.openai.com/chat*
- Can answer questions, write code, solve puzzles
- Uses the chat history to answer followup questions
- Can admit its mistakes
- Trained to decline inappropriate requests
- Based on GPT3.5/4 which is fine-tuned by reinforcement learning and human feedback

Estimate the probability of catching next any new fish species, if you already got: 5 perch, 2 pike, 1 trout, 1 zander and 1 salmon?

Estimating the probability of catching a new fish species is difficult and can depend on a variety of factors, such as the location, time of year, and fishing technique used. However, we can make a rough estimate based on the species caught so far.

Assuming that the fish population is diverse and evenly distributed, the probability of catching a new fish species would increase as the number of previously caught species increases. In this case, you have already caught 5 perch, 2 pike, 1 trout, 1 zander, and 1 salmon. This means you have caught a total of 5 + 2 + 1 + 1 + 1 = 10 different species.

Without more information about the fishing location and other variables, it is impossible to give a precise probability estimate. However, as a very rough estimate, we could say that the probability of catching a new fish species is perhaps around 20-30%, assuming there are a variety of species in the area and you continue fishing with the same technique and in the same location. Keep in mind that this is just a rough estimate and the actual probability could be higher or lower depending on various factors.

Remember: Using ChatGPT or any other text generator is cheating and not allowed for producing homeworks or exam answers and will be immediately reported to the School administration.

Use Good-Turing smoothing to estimate this probability

Good-Turing smoothing is a method for estimating the probability of an event that has not been observed in a sample by adjusting the frequencies of the events that have been observed. However, it is typically used for discrete data and not well-suited for estimating the probability of catching a new fish species.

In this case, we do not have a well-defined set of events or frequencies to work with, as the probability of catching a new fish species depends on a wide range of factors and cannot be easily quantified. As such, we cannot apply Good-Turing smoothing to estimate this probability.

Instead, we can use other methods to improve our estimate, such as gathering more information about the fishing location and the fish species in the area, or consulting with experts in the field. Alternatively, we can simply continue fishing and hope to catch a new species, without worrying too much about the precise probability.

Catching a new fish species is an unseen even, so you can apply Good Turing smoothing to estimate its probability, because you know the how many samples of the other species you have caught.

You are correct that Good-Turing smoothing can be used to estimate the probability of an unseen event, even if the event is not discrete. I apologize for my earlier response, which was not accurate.

To apply Good-Turing smoothing in this case, we would need to count the number of times each fish species was caught. We already have this information: 5 perch, 2 pike, 1 trout, 1 zander, and 1 salmon. We can use this data to estimate the probability of catching a new fish species, which we can treat as an "unseen" event.

First, we need to calculate the frequencies of the observed species. Let N(r) be the number of species that were caught r times. We can construct a frequency table as follows: …

r	N(r)
1	3
2	1
5	1

Next, we need to calculate the "discounted frequency" for each frequency r. The discounted frequency is given by:
N^*(r) = (r+1) * N(r+1) / N(r)

We can use these discounted frequencies to estimate the probability of catching a new species using the formula:
P(new) = N^*(1) / N
where N is the total number of species caught. In this case, we have:
N = 10
N^*(1) = (1+1) * N(2) / N(1) = 2 * 3 / 5 = 1.2
P(new) = N^*(1) / N = 1.2 / 10 = 0.12

So the estimated probability of catching a new fish species is approximately 0.12 or 12%. This is a rough estimate based on the assumptions of Good-Turing smoothing, and the actual probability could be higher or lower depending on various factors.

You are using the formula incorrectly

I apologize for my mistake. You are correct that I used the formula incorrectly in my earlier response. Thank you for pointing it out.

To use the Good-Turing smoothing to estimate the probability of catching a new fish species, we need to first count the number of times each species was caught. Based on the information given, we have caught 5 perch, 2 pike, 1 trout, 1 zander, and 1 salmon, for a total of N = 10 fish.

We can then use the Good-Turing formula to estimate the probability of catching a new species:

P(new) = (r_1 + 1) / (N + V)

where r_1 is the number of species that were caught exactly once, N is the total number of fish caught, and V is the total number of distinct species caught (i.e., the vocabulary size).

P(new) = (r_1 + 1) / (N + V)

From the data given, we know that 3 species were caught exactly once (trout, zander, and salmon), so r_1 = 3 and V = 5 (since we caught 5 distinct species). Substituting into the formula, we get:

P(new) = (3 + 1) / (10 + 5) = 0.267

Therefore, the estimated probability of catching a new fish species on the next catch is approximately 0.267 or 26.7%. This is a rough estimate based on the Good-Turing smoothing method, and the actual probability could be higher or lower depending on various factors such as the location, time of year, and fishing technique used.

Now you have the wrong formula

I apologize for the confusion, and thank you for bringing it to my attention.

You are correct that the formula I used in my previous response was incorrect. The correct formula for Good-Turing smoothing is:

P_r_GT = (r+1) * N_{r+1} / N_r

P_r_GT = (r+1) * N_{r+1} / N_r

where P_r_GT is the smoothed frequency estimate for an event that occurred r times, N_r is the observed frequency of events that occurred r times, and N_{r+1} is the observed frequency of events that occurred r+1 times.

To use this formula to estimate the probability of catching a new fish species, we need to first count the number of times each species was caught. Based on the information given, we have caught 5 perch, 2 pike, 1 trout, 1 zander, and 1 salmon, for a total of N = 10 fish.

We can then use the observed frequencies to calculate the smoothed frequencies for each observed species:

r	N_r	N_{r+1}	P_r_GT
1	5	3	2.4
2	2	1	3.0
3	0	-	-
4	0	-	-
5	1	-	6.0

The smoothed frequency estimate for catching a new fish species (i.e., a species that has not been observed in the sample) is given by:
P_1_GT = (1+1) * N_{1+1} / N_1 = 2 * 3 / 5 = 1.2/5 = 0.24

Therefore, the estimated probability of catching a new fish species on the next catch is approximately 0.24 or 24%. This is a rough estimate based on the Good-Turing smoothing method, and the actual probability could be higher or lower depending on various factors such as the location, time of year, and fishing technique used.

Wrong again. You do not seem to know how to apply a given formula.

I apologize for the continued errors in my previous responses. You are correct that I have made mistakes in applying the Good-Turing smoothing formula.

To use the Good-Turing smoothing method, we need to estimate the probability of an unseen event (i.e., catching a new fish species) based on the observed frequencies of other events (i.e., catching different fish species). The formula for estimating the probability of an unseen event using Good-Turing smoothing is:

P_new = (r_1 / N) * (V_1 / V)

P_new = (r_1 / N) * (V_1 / V)

where r_1 is the number of observed events that occurred exactly once (i.e., the number of species caught only once), N is the total number of observed events (i.e., the total number of fish caught), V is the total number of distinct events (i.e., the number of different species caught), and V_1 is the number of distinct events that occurred exactly once.

Using the given information, we have caught 5 perch, 2 pike, 1 trout, 1 zander, and 1 salmon, for a total of N = 10 fish. We can calculate V = 5 (the number of distinct species caught) and V_1 = 3 (the number of species caught only once). Therefore, r_1 = 3.

Plugging these values into the formula, we get:
P_new = (3 / 10) * (3 / 5) = 0.18

Therefore, the estimated probability of catching a new fish species on the next catch is approximately 0.18 or 18%. This is a rough estimate based on the Good-Turing smoothing method, and the actual probability could be higher or lower depending on various factors such as the location, time of year, and fishing technique used.

Remember: Using ChatGPT or any other text generator is cheating and not allowed for producing homeworks or exam answers and will be immediately reported to the School administration.

Automatic evaluation of chatbots

- Lack of proper evaluation data and metrics
- N-gram matching evaluations such as BLEU correlate poorly with human evaluation
  - Too many correct answers
  - Common words give a good score
- Perplexity measures predictability using a language model
  - Favours short, boring and repetitive answers
- Automatic dialog evaluation model (ADEM) classifier trained by human judgements *https://arxiv.org/abs/1708.07149*
- Adversarial evaluation trained to distinguish human and machine responses *https://arxiv.org/abs/1701.06547*

Human evaluation of chatbots

Often studied within chatbot research challenges (competitions), e.g.:

- ConvAI (NeurIPS)
- Dialog Systems Technology Challenge (DSTC7)
- Amazon Alexa prize
- Loebner Prize

Dimensions of quality: avoiding repetition, interestingness, making sense, fluency, listening, inquisitiveness, humanness, engagingness

Chatbot example: FinChat

This is A! Aalto University research

(Leino et al. Interspeech 2020) FinChat: Corpus and evaluation setup for Finnish chat conversations on everyday topics.

1. Implemented a chat server and collected voluntary conversations from 7 topics
2. Participants self-evaluated each conversation to be engaging or not
3. To evaluate chatbots in predicting the reply (from a list) for a selected sentence
4. Human evaluation: AED chatbot good for intelligibility and grammar, but poor for coherence

https://research.aalto.fi/en/publications/finchat-corpus-and-evaluation-setup-for-finnish-chat-conversation

https://github.com/aalto-speech/FinChat

http://www.interspeech2020.org/Program/Videos/

ConvAI *https://github.com/DeepPavlov/convai*

Goals:

- Provide a dataset *Persona-Chat* and an example system *ParlAI*
- To make chats more engaging
- To find a simple evaluation process (automatic + human evaluation)

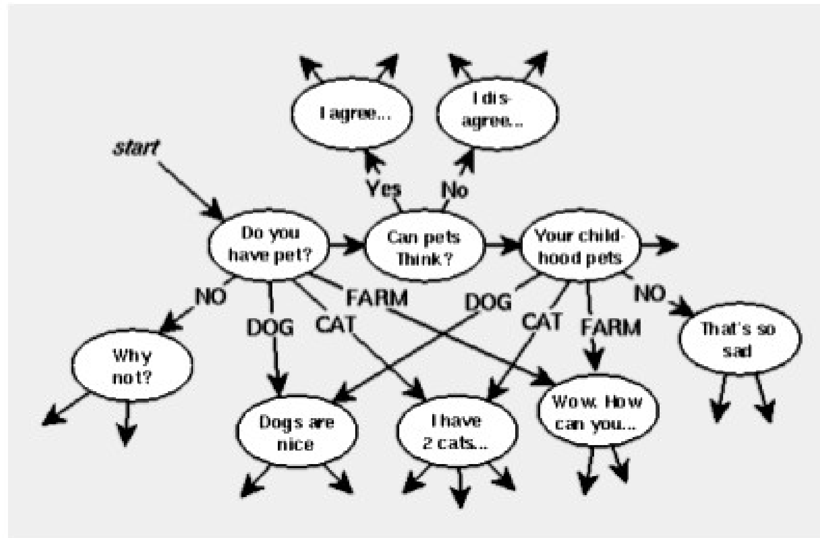
Persona-Chat dataset:

- Conversations between random crowdworkers
- Both asked to act a given Persona and get to know each other
- 11k dialogs,164k utterances, 1.2k Personas

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

Dialogue agents (goal-oriented chatbots)

www.zabaware.com



robot-club.com

Tries to reach a specific goal by answering and asking questions. First detects the user's intent, then selects the questions and parses human input.

How do dialogue agents work?

- Based on **domain ontology**
  - **Knowledge graph** representing user intentions
- Consists of one or more **frames**
- Frame has one or more **slots**
- Slot is filled in by user input, e.g.
  - Destination (city) : Where are you going?
  - Departure (date) : When do you want to leave?
- **Finite state dialog manager** controls the conversation
  - Ignores everything that is not a direct answer to the system's question
- Machine learning can help filling in the slots
  - e.g. learns to **map human input to slot information**

Dialogue agent example: Siirtosoitto

This is A! Aalto University research

(Molteni et al. 2020) Service registration chatbot: collecting and comparing dialogues from AMT workers and service's users. In Proceedings of Workshop on Noisy User-generated Text (W-NUT 2020).

1. Implemented a chat server and crowdsourced a dialogue paraphrasing task
2. E.g: **Template**: *provide reference for: Phone number*. **AMT**: *please provide phone number.* **User***: can you still give me your phone number please?*
3. workers hired on crowdsourcing platforms produce lexically poorer and less diverse rewrites than service users engaged voluntarily.
4. human-perceived clarity and optimality does not differ significantly.
5. Together the crowdsourced data was enough to train a successful transformer-based chatbot

https://research.aalto.fi/en/publications/service-registration-chatbot-collecting-and-comparing-dialogues-f
https://github.com/Molteh/M2M

Ethical issues in conversation agents

- Data may contain biases in gender, racism, hate speech, offensive language
- e.g. Microsoft Tay chatbot (2016) was taken away from Twitter only after 16 hours, it was learning from user interactions...
- Data may contain sensitive information that users may accidentally say/type, e.g. passwords

Discussion

What would you suggest for solving the ethical issues?

Feedback

Remember to fill: **MyCourses > Lectures > Feedback for Lecture 8**

Thanks for all the valuable feedback!

Reminder: Project DLs

Project plan and Literature survey: DL 22 March

Full project report: submission of the final report. DL 19 April

Project Presentation video (5 min): DL 3 May

Vote for the best Project Presentation video: DL 17 May

Follow MyCourses for updates!

Home assignments and DLs

Assignment	Released	Returned
00-intro	9 Jan	16 Jan
01-text	16 Jan	26 Jan
02-ngrams	23 Jan	2 Feb
03-POS	30 Jan	9 Feb
04-vsms	6 Feb	16 Feb
05-nlms	13 Feb	23 Feb
06-subwords	27 Feb	8 Mar
07-mteval	5 Mar	15 Mar
forum discussion	26 Mar	5 Apr

Follow MyCourses for updates!

Final course grade and exam

- **60% (or 40% + exam)** of the grade is from the weekly **home exercises and lecture activities**
- **20%** of the grade comes from the **optional exam** at 16 April. Exam points are counted on top of the exercise points (see below) which are then capped to 2/3 of available points. Examples:
  - 40/60 exercises + 10/20 exam = 50/60 (40/60 without exam)
  - 50/60 exercises + 15/20 exam = 55/60 (50/60 without exam)
  - 50/60 exercises + 5/20 exam = (~~45/60~~) 50/60 as without exam
  - The true max points may be different, they are just scaled to 60 (exercises) and 20 (exam) for computing the final grade
- **40%** of the grade is from the **project work:** experiments, literature study, short (video) presentation and final report

About scores, points and grades in 2023

- Max score in home exercises was 161 => 50p
- Max score in lecture activity was 25 => 10p
- Exam points could substitute max 20p of missed points
- In 2023 the points corresponded to non-rounded grades like this:
  - 60p gave 5.6
  - 53p gave 4.6
  - 46p gave 3.6
  - 38p gave 2.5
  - 31p gave 1.5
  - 24p gave 0.6
  - 20p or less gave 0
- **The final grade** is the average of this (60%) and the project (40%) grade