Aalto University

Conversational agents: chatbots and dialogue agents

Image by Gracini Studios from Pixabay

Mikko Kurimo
SNLP lecture 8

Based on Chapter 24 in Jurafsky-Martin 3rd edition (version 2020)

Lecture schedule 2022

1. 11 Jan Introduction & Project groups / Mikko Kurimo
2. 18 jan Statistical language models / Mikko Kurimo
3. 25 jan Word2vec / Tiina Lindh-Knuutila
4. 01 feb Sentence level processing / Mikko Kurimo
5. 08 feb Speech recognition / Janne Pylkkönen
6. 15 feb Morpheme-level processing / Mathias Creutz
7. 22 feb Exam week, no lecture
8. **01 mar Chatbots and dialogue agents / Mikko Kurimo**
9. 08 mar Statistical machine translation / Jaakko Väyrynen
10. 15 mar Neural language modeling and BERT / Mittul Singh
11. 22 mar Neural machine translation / Stig-Arne Grönroos
12. 29 mar Societal impacts and conclusion / Krista Lagus, Mikko

See Mycourses for updates

Aalto University

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. 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.

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
- IR: Cleverbot
- DNN encoder-decoders etc

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

- Regular expressions as ELIZA
- Control structure
- Some language understanding
- Mental model

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/

Lecture exercise 7: Try chatbots

Discuss in breakout rooms and propose answers for these 6 questions into **MyCourses > Lectures > Lecture 7 exercise return box:**

1. Which chatbots and dialogue agents have you used?
   - What can they do, what not?
2. Try ELIZA, e.g. *https://www.eclecticenergies.com/ego/eliza* or http://psych.fullerton.edu/mbirnbaum/psych101/Eliza.htm
   - When does it fail? How to improve it?
3. Try PARRY, e.g. https://www.chatbots.org/chatbot/parry/ or https://www.botlibre.com/browse?id=857177
   - When does it fail? How to improve it?
4. Try more chatbots or dialogue agents, e.g. transformer: https://convai.huggingface.co/ or anyone from: https://www.chatbots.org/
5. What do you think: How to make better chatbots?
   - How to automatically evaluate chatbots?
6. What ethical issues do chatbots have?
   - Any suggestions how to solve them?

20 min work
10 min break

Aalto University

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*
-
- Common problems with chatbots:
  - Lack of consistent personality
  - Lack of long-term memory
  - Boring answers like ”I don't know”

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

Chatbot example: FinChat

This is **A!** Aalto University **research**

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

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. Accuracy 95% for human, 10% for chatbots (transformer vs encoder-decoder) trained on Finnish conversational data (Open Subtitles vs Suomi24)
5. Human evaluation: AED chatbot good for intellligibility 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

Examples of machine learning chatbots

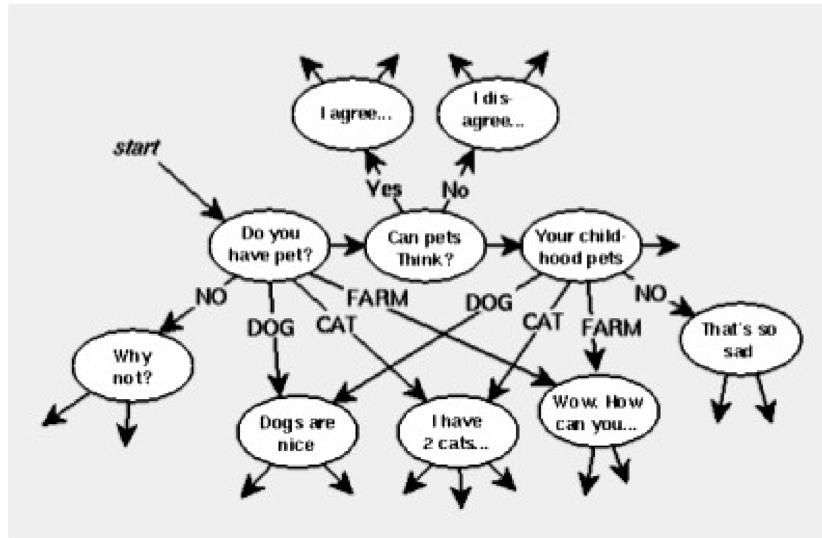
Team Names	Model Summary
Lost in Conversation	Generative Transformer based on OpenAI GPT. Trained on PERSONA-CHAT (original+revised), DailyDialog and Reddit comments.
Hugging Face	Pretrained generative Transformer (Billion Words + CoNLL 2012) with transfer to PERSONA-CHAT.
Little Baby	Profile-Encoded Multi-Turn Response Selection via Multi-Grained Deep Match Network. Modification of [9]: better model + data augmentation via translation.
Mohd Shadab Alam	Seq2Seq + Highway model. Glove + language model vector. Transfer learning strategy for Seq2Seq tasks.
ADAPT Centre	Bi-directional Attentive LSTM. Pretrained via GloVe embeddings + Switchboard, Open Subtitles.

Table of some top competitors in ConvAI 2018. For more info, see:

- *Challenge overview paper (https://arxiv.org/abs/1902.00098)*
- *http://convai.io/NeurIPSParticipantSlides.pptx*
- *https://github.com/atselousov/transformer_chatbot*
- *https://medium.com/huggingface/how-to-build-a-state-of-the-art-conversational-ai-with-transfer-learning-2d818ac26313#79c5*

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?
- 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 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?

Reminder: Project DLs

1. Project plan and Literature survey: **10 March** (uploaded to peergrade directly)
2. Peer grading for the Project plan and the Literature survey: **17 March**
3. Feedback on peer grading (rebuttal/grade): **24 March**
4. Full project report: submission of the final report. See the details below. **28 April**
5. Project Presentation video (5 min): **5 May**
6. Vote for the best Project Presentation video: **19 May**

Follow MyCourses for updates!

Next home assignment DLs

Assignment	Released	Returned
03-vsms	1 Feb	14 Feb
04-pos-tagging	8 Feb	3 March
05-mt-evaluation	16 Feb	7 March
06-subwords	1 March	14 March
07-neural-lm	8 March	21 March
08-Forum discussion	29 March	11 April

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 12 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

Feedback

Remember to fill: **MyCourses > Lectures > Feedback for Lecture 7**

Thanks for all the valuable feedback!