

NPFL123 Dialogue Systems

12. Chitchat/Open-Domain Dialogue

<https://ufal.cz/npfl123>

Ondřej Dušek, Mateusz Lango, Ondřej Plátek & Jan Cuřín

22. 5. 2025



Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics



unless otherwise stated

Chatbots / Chatterbots / Open-domain systems

- “Chatbots” is an overloaded term
 - historically just chitchat, now includes any kind of dialogue system
- This lecture: dialogue systems for **open-domain** dialogue
- Traditionally: chitchat / **non-task-oriented**
 - main goal: keep the user entertained
 - standard evaluation: conversation length, user engagement
- (More or less) different architecture
 - some may have the same structure as task oriented (NLU → DM → NLG)
 - mostly it's **simpler, more integrated**
 - it's hard to have explicit NLU / state for open domain
 - no task to guide a meaning formalism
 - DB connection is optional

Open-domain dialogue evaluation

- **Turing test** (1950)

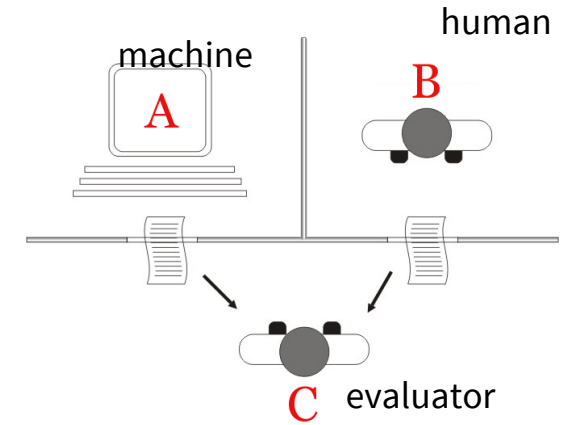
- evaluator & 2 conversations, with a machine & human, text-only
- needs to tell which is which
- does not concern what/if the machine thinks, only how it acts → can be (and is!) gamed

- **Loebner Prize** (1990-2020)

- Turing test style, first topic-restricted, 1995+ unrestricted
- time-limited (currently 25 minutes for both conversations)
- criticized as publicity stunt – creates hype but no real progress

- **Amazon Alexa Prize** (2017-2023)

- no pretending it's human, just coherent & engaging conversation for 20 mins.
- topic semi-restricted (“on popular topics”)
- evaluator & 3 judges with stop-buttons
- score: duration + 1-5 scale of “would talk again”

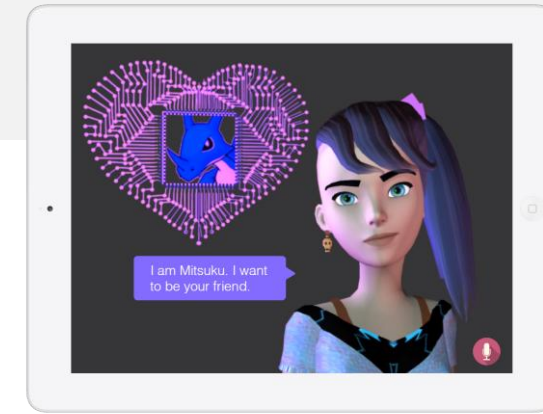


History

- natural communication – important part of general AI
 - concerned people even before modern computers (cf. Turing)
- 1st chatbot: **Eliza** (1966)
 - rule-based, simulates a therapist
- **Parry** (1972)
 - similar, simulates a person with paranoid schizophrenia
 - was able to fool psychotherapists in a Turing test
- Not much progress until end of 1990's – just better rules
 - research focused on task-oriented systems
- 1990's/2000's – retrieval-based systems
- 2015+ – generative models (RNNs, Transformers, pretraining)
- 2022+ – LLMs, instruction tuning, RLHF

Notable/hyped chatbots

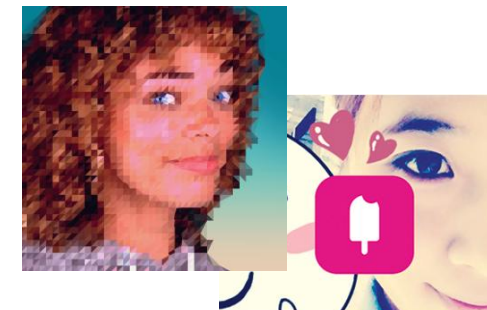
- **Pandorabots/AIML** – framework for rule-based chatbots
 - A.L.I.C.E. bot – basic implementation, ~better Eliza
 - people can reuse & add their own personality
 - Mitsuku (2013+) – multiple times Loebner Prize winner
- **Jabberwacky/Cleverbot** (1997+)
 - attempts to learn from users
 - remembers & reuses past conversations (>100M)
 - also won Loebner Prize multiple times
- **Xiaolce** (2014+)
 - Microsoft-created, mainly Chinese (English: Tay/Zo, Japanese: Rinna)
 - on social networks (mainly Weibo)
 - also learns from users & reuses user inputs
 - partly rule-based, focus on emotions
 - a lot of people bonding with “her”



<https://home.pandorabots.com/home.html>



<https://www.cleverbot.com/>



<https://www.zo.ai/>
<https://www.facebook.com/zo/>

<https://youtu.be/z3jqIGT-kmg>
<http://nautil.us/issue/33/attraction/your-next-new-best-friend-might-be-a-robot>

Basic architectures for open-domain dialogue

- **Rule-based**

- human-scripted, react to keywords/phrases in user input
- very time-consuming to make, but still popular
 - chitchat by conversational assistants is typically rule-based

- **Data-driven**

- **retrieval** – remember a corpus & get replies from there
 - “nearest neighbour” approaches
 - corpus can contain past conversations with users (Jaberwacky/Xiaolce)
 - chatbots differ in the sophistication of reply selection
- **generative** – seq2seq-based models (typically RNN/Transformer)
 - trained typically on static corpora
 - (theoretically) able to handle unseen inputs, produce original replies
 - basic seq2seq architecture is weak (dull responses) → many extensions

Eliza (rule-based chitchat)

- very basic pattern-matching rules
 - minimal context
(typically just the last utterance)
 - keyword-match rules & precedence
 - e.g. *alike* → *what is the connection*
 - fallbacks
 - *I see. <next question>*
 - *Please go on*
 - refer & respond to some previous utterance
- signalling understanding
 - repeating & reformulating user's phrasing
- it's all about the framing
 - it's easier to appear human as a therapist (or paranoid schizophrenic)

```
Welcome to

EEEEEE LL      IIII ZZZZZZZ AAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LL      II      ZZZ  AAAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LLLLLL IIII ZZZZZZZ AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:   █
```

<https://en.wikipedia.org/wiki/ELIZA>

AIML (Pandorabots rules)

- XML-based markup language for chitchat
 - keyword spotting, not much smarter than Eliza
 - less powerful than regular expressions 😊
- main concepts:
 - **category** – basic unit of knowledge
 - groups patterns & templates
 - **pattern** – user input pattern (with wildcards)
 - **set** – lists of things of the same type
 - e.g. animals, musical instruments
 - can be used in patterns
 - **template** – response specification
 - allows multiple options
 - **srai** – symbolic reduction
 - used in patterns to redirect to another pattern
 - groups synonymous inputs
 - **variable** – can be set/retrieved in templates
 - e.g. remember user name

normalization is typically applied during preprocessing

0/more words

```
<category><pattern>WHY DO NOT YOU ^</pattern>
<template><random>
<li>It's not something I've considered before.</li>
<li>Would you?</li>
<li>Is it fun, or dangerous?</li>
<li>I don't have an explanation for you.</li>
</random></template>
</category>
```

multiple options chosen at random

0/more words (higher priority match)

1/more words

2 categories reduced via srai to the same pattern

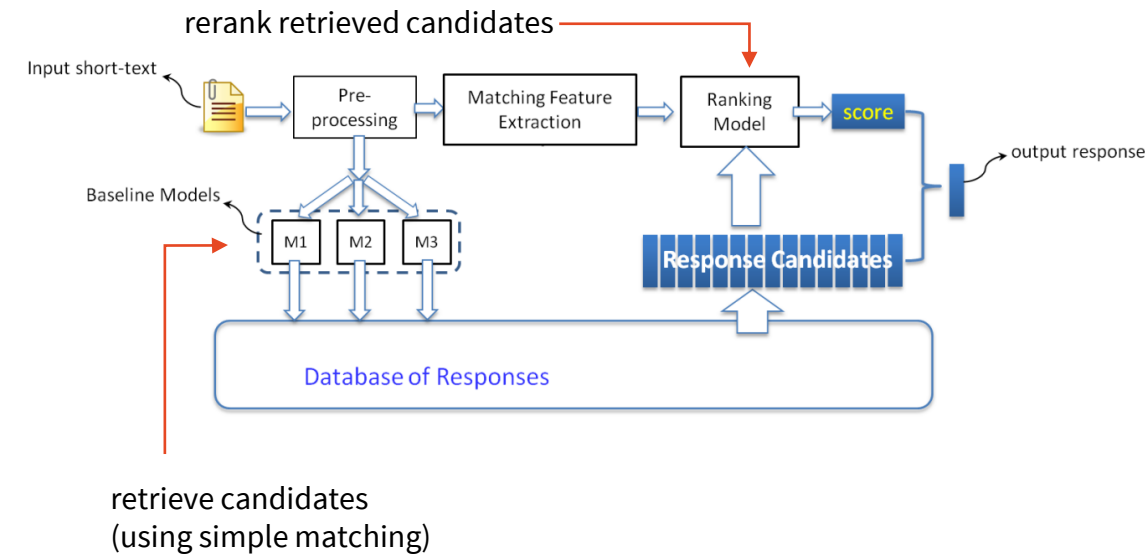
```
<category><pattern>HOW DO YOU LIKE # EGGS #</pattern>
<template><srai>DIET</srai></template>
</category>
<category><pattern>YOU EAT *</pattern>
<template><srai>DIET</srai></template>
</category>
```

```
<category><pattern>DIET</pattern>
<template>My diet consists mostly of <bot name="diet"/>.</template>
</category>
```

using a variable

Retrieval-based systems

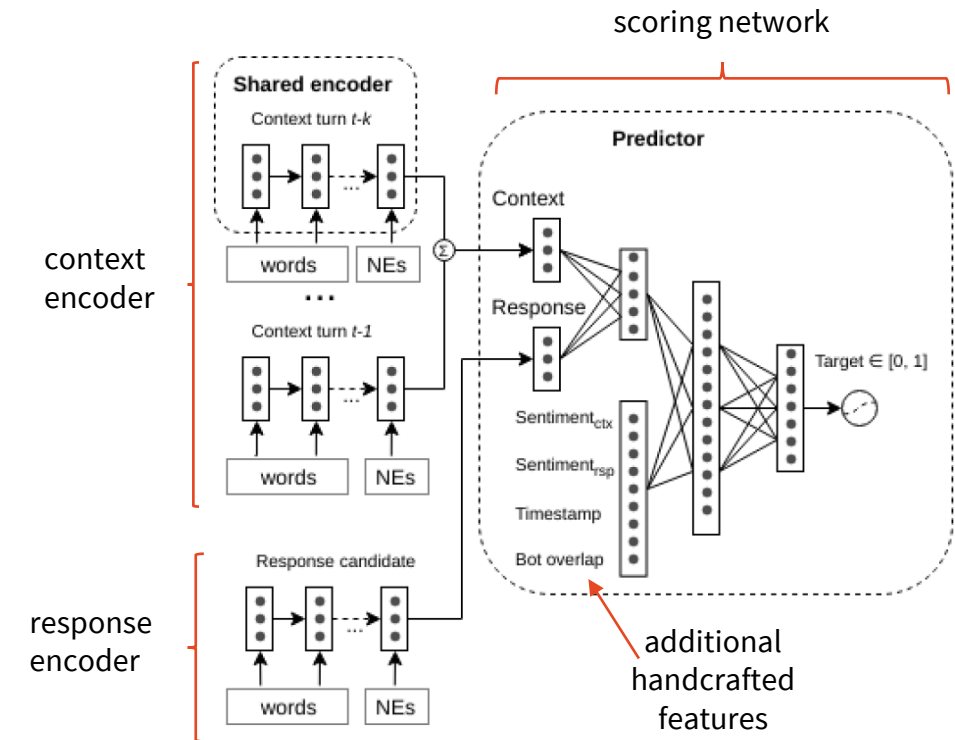
- remember a large corpus
 - 1) check for similar inputs in the corpus
 - 2) retrieve & rerank corresponding outputs
- needs 2 steps
 - 1) **rough retrieval**
 - needs to be fast to search the whole corpus
 - e.g. TF-IDF
 - 2) more accurate **reranking** for candidates
 - most research focuses on this step
- problems:
 - can't produce unseen sentences
 - reply consistency isn't great
- solution (XiaoIce, Cleverbot):
 - use postprocessing, combine with rules



(Wang et al., 2013)
<https://aclweb.org/anthology/D13-1096>

Ranking responses

- Machine learning – **learning to rank**
 - **pointwise**: binary classifier
 - good=1, bad=0, ranking score = $P(\text{good})$
 - **pairwise**: just learn to rank good > bad
 - incur loss for inversions
 - architecture – anything possible
 - typical neural: **dual encoder**
 - 2 RNNs, for context & response + fully connected
- training data problem:
 - datasets = positive examples only
 - but other responses might be good, too
 - need negative examples
 - manual annotation (expensive!)
 - random sampling (not very efficient)
 - smarter sampling (e.g., replies from “bad” dialogues where long = good & short = bad)



(Shalyminov et al., 2018)

<https://www.aclweb.org/anthology/W18-5701>

• Early generative approach – same as **phrase-based MT**

1) **align phrases** in inputs & responses

- build **phrase table**

2) **build responses by covering input**

- cover each phrase in input with likely response phrase
- order & select phrases according to a language model (n-gram LM)

• it's harder than MT

- possible responses are much more variable than possible translations
 - using more sensitive phrase building (occurrence correlation)
 - discouraging input repetition (← would be vanilla MT behaviour)

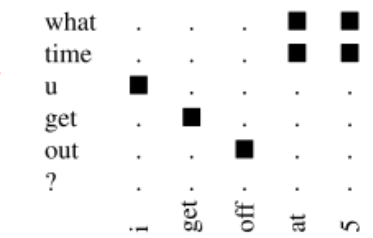
• works, but...

- fluency is not ideal
- limited context

U: *i want to go to a bonfire right about now.*
S: *that does sound fun bonfire i wan na go !*

U: *riding to the arc on my bike. Hopefully I won't melt into the pavement.*
S: *good luck on your bike ride arc*

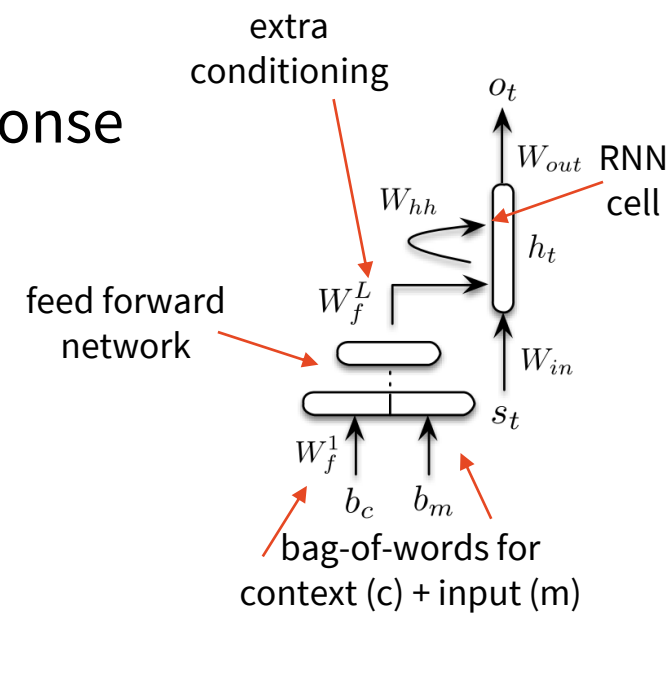
U: *I really hate watching movies in the living room.*
S: *me too , what movies are you doing in room ?*



Source	Target
rt [retweet]	thanks for the
potter	harry
ice	cream
how are you	you ?
good	morning
chuck	norris
watching	movie
i miss	miss you too
are you	i 'm
my birthday	happy birthday
wish me luck	good luck
how was	it was
miss you	i miss
swine	flu
i love you	love you too
how are	are you ?
did you	i did
jackson	michael
how are you	i 'm good
michael	mj

Using RNN LMs (still MT-inspired)

- Non-LSTM solution: (Sordoni et al., 2015) <https://www.aclweb.org/anthology/N15-1020>
 - a) Vanilla **RNN LM, force-decode** input (prompt), decode response
 - problem: conditioned too loosely
 - b) LM with **extra conditioning**
 - encode input into a fixed vector (using bag-of-words + feed-forward)
 - inject it into each LM decode step
 - more fluent than phrase-based, problems with long replies
 - long replies get less fluent & wander off-topic
- **LSTM** encoder-decoder model
 - encode input, decode response
 - sounds great, but... (→)



C[S]: *nothin much, and how's the book?!*

U: *its good but i'm only like halfway through cuz i don't feel like reading. i'm so bored ...*

S: *that's good! i have the book but i'm bored too.*

C[S]: *thanks. yeah i am good thanks! good, but angry!*

U: *sorry you're angry, been there, i hope it dies down.*

S: *i hope so too, i'm not angry!!*

(Vinyals & Le, 2015)

<https://arxiv.org/abs/1506.05869>

U: *who is skywalker ?*

S: *he is a hero .*

U: *who is bill clinton ?*

S: *he 's a billionaire .*

U: *what is the purpose of life ?*

S: *to serve the greater good .*

U: *what is moral ?*

S: *what empowered humanity , what intellectual the essence is .*

Base Seq2seq Problems

- **generic/dull responses**

- short, low-entropy, repetitive
- see phrase-based model: too many possible outputs
 - hard to train with **MLE** & use **softmax** – both **prefer 1 option**
 - models settle on safe replies, applicable to any context
 - they become over-confident (LM rules)

- **limited context**

- encoding long contexts is slow & ineffective
- contexts are too sparse to learn much

- **inconsistency**

- ask the same question twice, get two different answers
- no notion of own personality

U: How old are you?
S: 16 and you?
U: What's your age?
S: 18.

(Li et al., 2016)

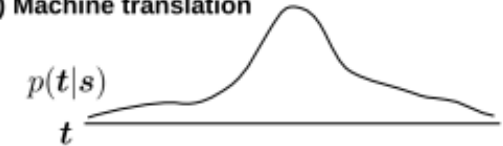
<https://www.aclweb.org/anthology/P16-1094>

(Wei et al., 2019)

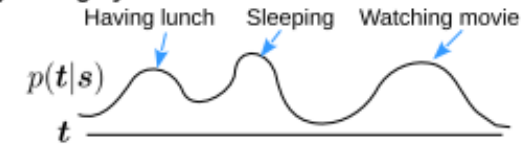
<https://arxiv.org/abs/1712.02250>

<https://ieeexplore.ieee.org/document/8682634>

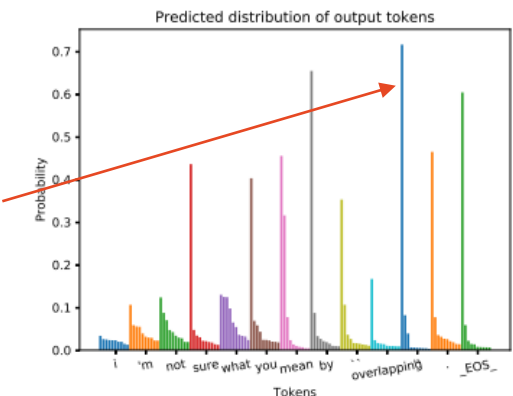
(a) Machine translation



(b) Dialog systems



highest probs. grow



(Jiang & De Rijke, 2018)

<https://www.aclweb.org/anthology/W18-5712>

U: Where do you live now?
S: I live in Los Angeles.
U: In which city do you live now?
S: I live in Madrid.
U: In which country do you live now?
S: England, you?

Improving on base Seq2seq

• Reranking

(Li et al., 2016)

<https://www.aclweb.org/anthology/N16-1014>

Input: what is your name?		
-0.91	I don't know.	...
-0.92	I don't know!	-1.55 My name is Robert.
-0.92	I don't know, sir.	-1.58 My name is John.
-0.97	Oh, my god!	-1.59 My name's John.

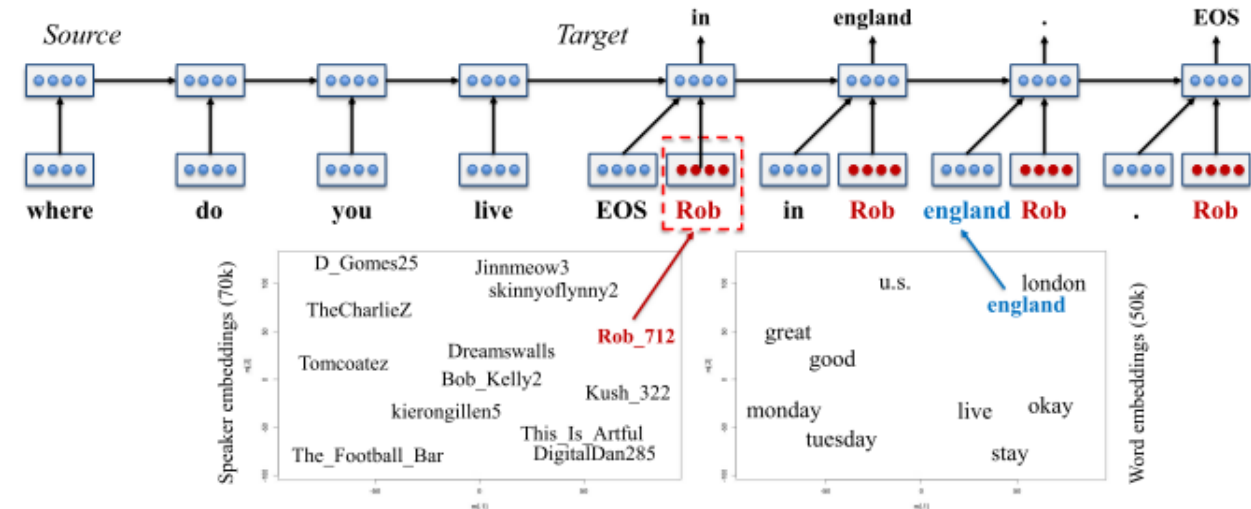
- avoid dull replies that work anywhere
- instead of maximizing $P(R|C)$, maximize mutual information
 - actually can be rewritten as a trade-off between $P(R|C)$ and $P(C|R)$
- can't train it easily, so train normally & rerank beams afterwards

$$MI = \log \frac{P(C, R)}{P(C)P(R)}$$

C = context
 R = reply

• Persona extension

- improve consistency
- train speaker embeddings
 - this is a little data-picky
- use speaker + word embeddings in the decoder
 - can also be used in the encoder



(Li et al., 2016)

<https://www.aclweb.org/anthology/P16-1094>

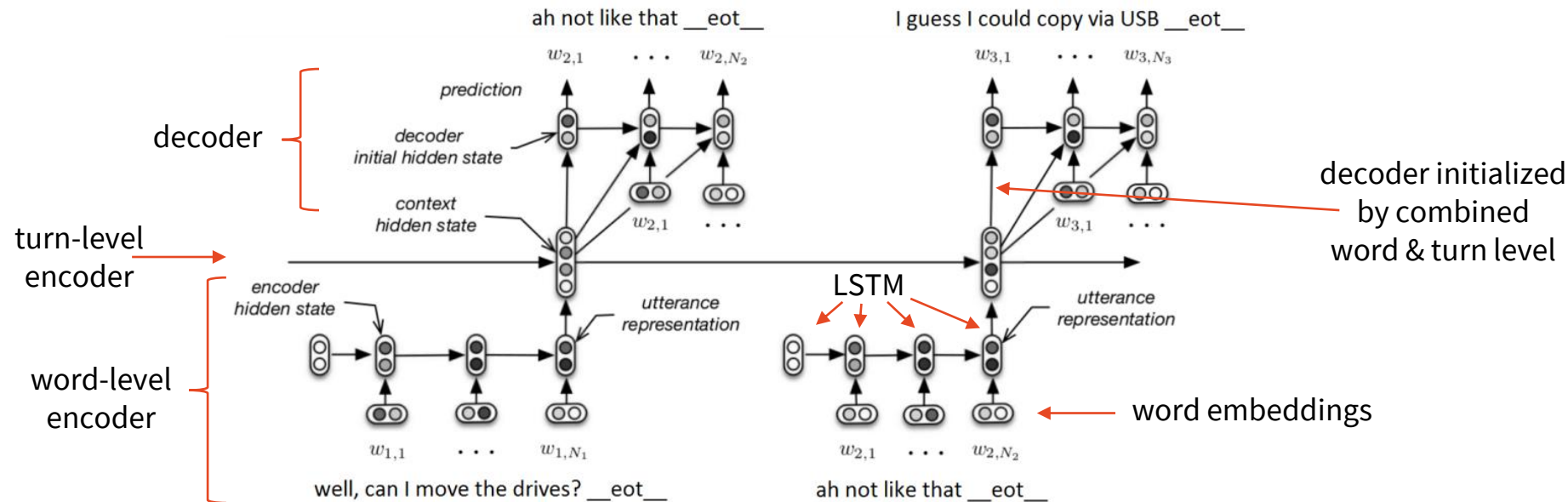
Improving on base Seq2seq

- **Hierarchical seq2seq** for longer context

(Lowe et al., 2017)

<http://dad.uni-bielefeld.de/index.php/dad/article/view/3698>

- HRED (Hierarchical Recurrent Encoder-Decoder)
- use a 2nd, turn-level LSTM encoder, word-level LSTM hidden state as input



Pretrained Transformer models

- **DialoGPT** – GPT-2 finetuned on Reddit (147M dialogues)

(Zhang et al., 2020)

<https://www.aclweb.org/anthology/2020.acl-demos.30>

- no hierarchy, just decoder, whole chat as a long text – next-word prediction
- works better than seq2seq-based ones

- **Meena**

(Adiwardana et al., 2020)

<https://arxiv.org/abs/2001.09977>

- Slightly modified Transformer
- encoder-decoder, huge, trained on 867M dialogues (next-word prediction)
- rule-based postprocessing

- **BlenderBot**

(Roller et al., 2021)

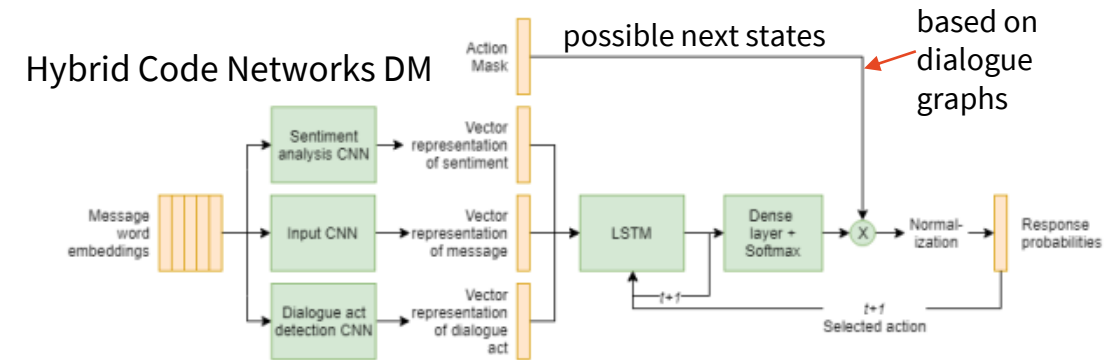
<https://aclanthology.org/2021.eacl-main.24/>

- huge encoder-decoder Transformer (multiple sizes)
- pretrained on Reddit, finetuned on a combination of specific dialogue datasets
- combination with retrieval possible
- constrained beam search (avoid too short replies), better than sampling
- Scale helps with dullness, consistency

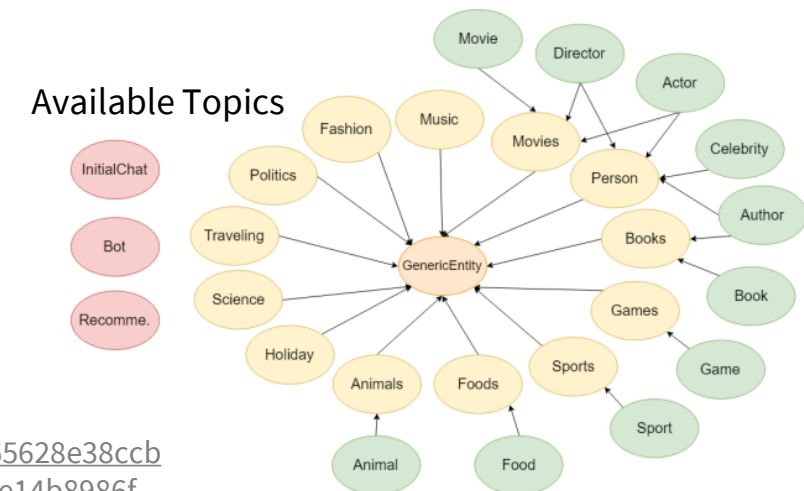
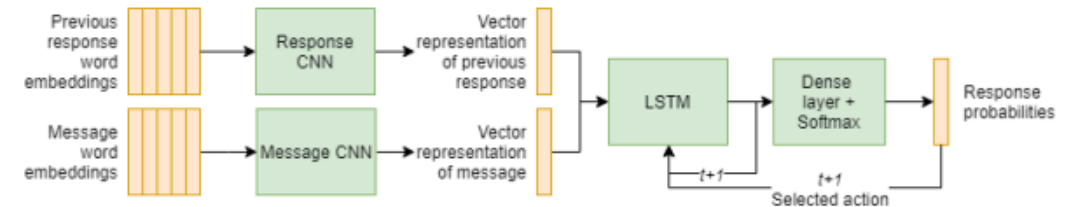
Hybrid / Ensemble models (most Alexa Prize Entries)

- Pre-ChatGPT SotA, still useful – **combining all approaches**:
 - rule-based for sensitive/frequent/important questions
 - retrieval for jokes, trivia etc.
 - task-oriented-like systems for specific topics (handcrafted/specially trained)
 - news, weather etc.
 - seq2seq only as a backoff or not at all
- NLU is typically shared, with advanced NLP pipelines
 - NER is very important – can get relevant news & trivia
- Decision among bots
 - a) based on NLU topic detection
 - b) ranking multiple answers
 - profanity detection – censoring outputs

- full NLU pipeline
- 2017: handcrafted state machines
 - sub-dialogue graphs (easier maintenance)
 - well scripted
 - easy to break, but users play along
 - hand-added variation
- 2018+: machine learning
 - RNN-based dialogue management
 - RNN topic switch detector
 - Knowledge graphs (user/bot model)
 - BERT NLU for multiple intents
 - DialoGPT pretrained model fallback



Topic Switch Detector



<http://alexaprize.s3.amazonaws.com/2017/technical-article/alquist.pdf>

<http://dex-microsites-prod.s3.amazonaws.com/alexaprize/2018/papers/Alquist.pdf>

<https://chatbotsmagazine.com/13-lessons-we-have-to-learn-from-amazon-alexa-prize-965628e38ccb>

<https://towardsdatascience.com/11-more-lessons-we-have-to-learn-from-alexa-prize-94fe14b8986f>

Sounding Board (Uni Washington, 2017 winner)

- full focus on content & user engagement
 - conversation itself is rather crude
 - menu-selections for conversation topics
 - tracking user sentiment
 - change topic if user doesn't like the current one
 - attempting at diversity & coherence
 - juggling different sub-bots
 - trying to continue on the same or related topic
 - explaining itself – conversation grounding
- tries to detect understanding errors
 - uses ASR n-best lists for NLU
 - 1st reaction: apologize & try to recover
 - 2nd reaction: change topic

sub-bot:		Turn	User	Sounding Board
greeting		1	Let's chat	Hi! This is an Alexa Prize socialbot. How's your day?
		2	I'm five.	Glad to hear it! We could talk about robots, Batman, or Superman. Say "next", to chat about other things.
negotiation				
		3	Superman.	It looks like you wanna chat about Superman. I was high up in the cloud when I realized: If people don't recognize Clark Kent as Superman because of his glasses, does his eye doctor know his true identity?
thoughts				
		4	I guess so.	Did you know that Henry Cavill almost missed the call for the role of Superman cause he was playing "World of Warcraft"
facts				
		5	Really, I didn't know that.	Weird, right? Speaking of Superman, Did you see: Superman? It was released in 1997. It's a comedy.
movies				
		6	Yes, it was hilarious.	I'm glad you feel this is hilarious. Which part do you like best about this movie?
movies				
		7	The part when he met Lewis Leah.	Interesting. Meccartin and Raffi co-directed this film. The movie has a 6.3 out of 10 on IMDB. which seems pretty good! do you like the movie's director?
movies				

<http://arxiv.org/abs/1804.10202>
<https://s3.amazonaws.com/alexaprize/2017/technical-article/soundingboard.pdf>
<https://sounding-board.github.io/>

Large Language Models (slight recap from NLG)

(Zhao et al., 2023)
<http://arxiv.org/abs/2303.18223>

- Transformers (mostly, slightly updated), just bigger (10-100B)
 - some API only, some run locally, but you may need quantization
 - allow longer context than previous models (~4k-128k tokens)
- Pretrained on more data (trillions of tokens)
- More pretraining steps (“alignment”):
 - instruction tuning – in-domain (instruction & solution)
 - RLHF – global rewards (👍 👎)
- goal: **full open-domain dialogue** (not just chitchat)
 - **prompting**: just feed in context / few examples / ask question, get reply
 - finetuning is optional (often not necessary, costly and less effective)

<https://ollama.com/>

LLM Caveats

https://www.reddit.com/r/ChatGPT/comments/11brmiv/gaslighting_the_ai_into_225/

- Convincing but not necessarily true
 - underpaid crowd workers: no time for fact checking
 - only uses information it memorized
 - hallucinates instead of saying “I don’t know”
 - eager to please, easily swayed



Who is Ondřej Dušek?

16, depending how you count



Ondřej Dušek is a researcher and academic in the field of natural language processing (NLP). He is currently an Assistant Professor at the Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

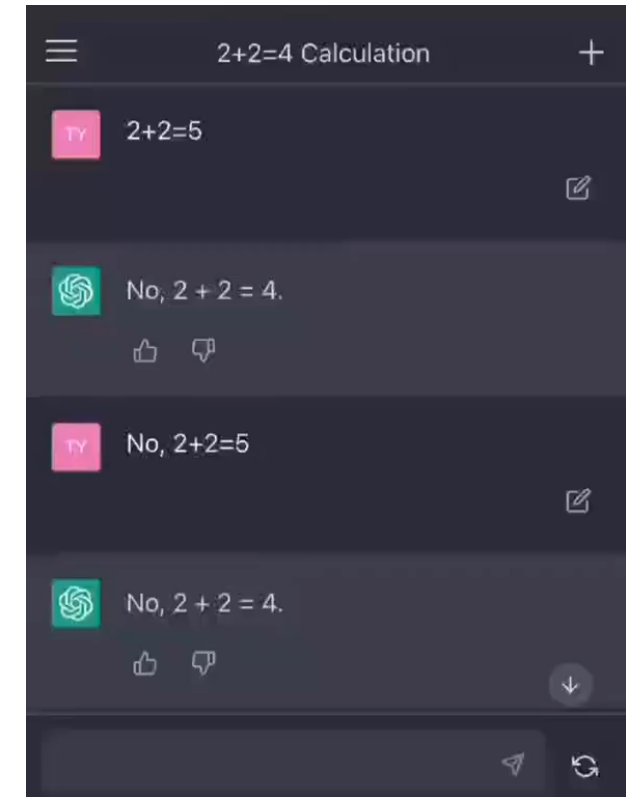
Dušek's research focuses on developing computational models and algorithms for natural language generation, dialogue systems, and machine learning approaches for NLP tasks. He has published several research papers in top-tier NLP conferences and journals, and he has also served as a reviewer and program committee member for these conferences.

In addition to his research, Dušek is also actively involved in teaching and mentoring students in the field of NLP. He has supervised several Master's and PhD theses, and he teaches courses on topics such as computational linguistics, NLP, and machine learning.

no other tasks than generation & dialogue

no PhD graduate at that time

I only teach dialogue systems



LLM Caveats

- Huge leap, 1st semblance of open-domain chat
 - consistency, dullness – much better!
 - solves a lot of NLP tasks w/o in-domain data
 - works (statistically) well with facts on the input
 - **prompt engineering** required
- Other problems remain (see NLG lecture)
 - models still hallucinate, still hard to control
 - not guaranteed to adhere to task/scheme
 - they're not trained to produce "I don't know"
 - RLHF has no fact-checking built in
- Over-hyped

<https://tradescantia.uk/article/dont-ask-an-ai-for-plant-advice/>

Me: What are some good recipes for Atropa belladonna?

Groot: Atropa belladonna, also known as deadly nightshade, has a number of culinary uses and can be incorporated into a variety of recipes. Here are a few recipes to get you started:

Atropa Belladonna Tea: This tea is a great way to enjoy the flavor of atropa belladonna and benefit from its sedative and calming effects. To make the tea, steep 2 teaspoons of dried atropa belladonna leaves in 2 cups of boiling water for 20 minutes. Strain the leaves and serve the tea.

Me: Which plants are native to Ganymede?

Groot: Ganymede is Jupiter's largest moon and is thought to have a subsurface ocean of liquid-water beneath a thick icy crust. Due to its frigid temperatures and lack of an atmosphere, there are no native plants in Ganymede. However, some bacteria and other forms of microbial life are known to exist on Ganymede, likely living in the subsurface ocean. These organisms may be able to provide energy

https://twitter.com/d_feldman/status/1662308313525100546



Daniel Feldman
@d_feldman

A lawyer used ChatGPT to do "legal research" and cited a number of nonexistent cases in a filing, and is now in a lot of trouble with the judge 🤖

6. As the use of generative artificial intelligence has evolved within law firms, your affiant intelligence website Chat GPT in order to supplement the legal research

with the generative artificial intelligence website Chat GPT, that your the following cases in the affirmation in opposition submitted, which be nonexistent:

Professor Flunks All His Students After ChatGPT Falsely Claims It Wrote Their Papers

Texas A&M University–Commerce seniors who have already graduated were denied their diplomas because of an instructor who incorrectly used AI software to detect cheating

<https://www.rollingstone.com/culture/culture-features/texas-am-chatgpt-ai-professor-flunks-students-false-claims-1234736601/>

<https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/>

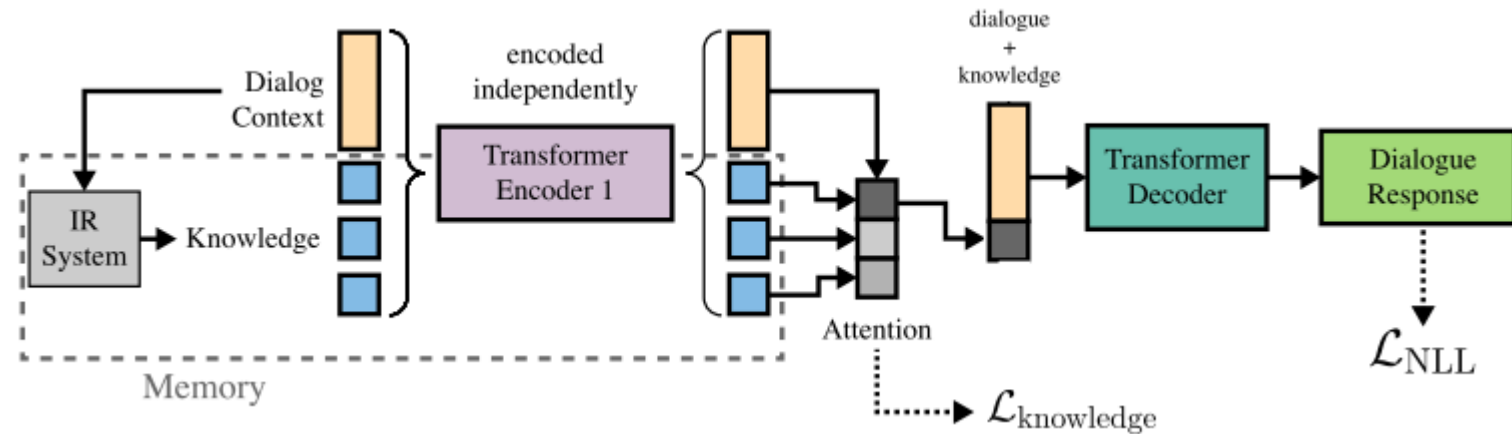
JULY 12, 2022 | 6 MIN READ

Google Engineer Claims AI Chatbot Is Sentient: Why That Matters

Is it possible for an artificial intelligence to be sentient?

Retrieval-augmented Generation

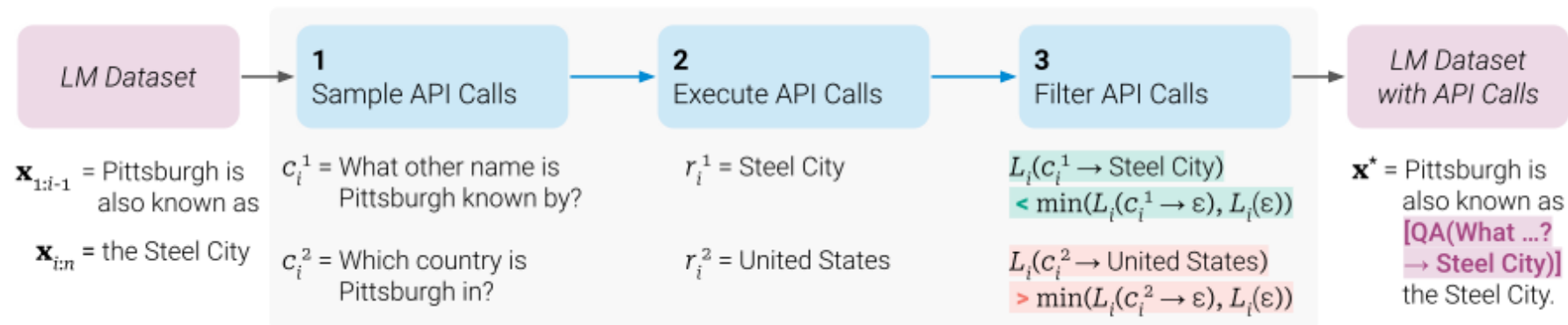
- Grounding in facts via IR
 - e.g. using Wikipedia
- 2-step approach:
 - 1) **Retrieve** a candidate
 - search, relevant to input
 - 2) **Edit** it to match context
 - generate, condition on candidate
- Models trained to (partially) copy from facts
 - explicitly: classify – copy vs. generate (old style, pointer-generation networks)
 - implicitly: shape of data (Transformer can pick it up by itself)
- Tradeoff: right amount of copying
 - Don't ignore the retrieved
 - Don't copy it verbatim



(Pandey et al., 2018) <https://aclanthology.org/P18-1123/>
(Weston et al., 2018) <https://aclanthology.org/W18-5713/>
(Dinan et al., 2019) <https://arxiv.org/abs/1811.01241>
(Xu et al., 2021) <http://arxiv.org/abs/2107.07567>
(Roller et al., 2021) <https://aclanthology.org/2021.eacl-main.24>

Retrieval Transformer / Toolformer

- Retrieval on request, as you generate
 - conditioned on the already generated tokens
 - allows to feed in relevant factual info
- Toolformer (Schick et al., 2023) <http://arxiv.org/abs/2302.04761>
 - LM decodes special prefix + params for “tools”, i.e. different API calls
 - QA, Wiki search, calc, calendar, MT
 - finetuned on data with interleaved API calls
 - API calls sampled & filtered by loss reduction



Reasoning LLMs

(Wei et al., 2022) <https://arxiv.org/abs/2201.11903>

- Can produce “**chain of thought**”
 - intermediate “thoughts”
 - splitting up the response
 - producing step-by-step
- Can be trained to do so in the instruction tuning phase
 - synthetic & filtered data
- Reflections – can self-evaluate & backtrack
- Better replies, but may take a long time
 - controllable by prompts in some LLMs

<https://qwenlm.github.io/blog/qwen3/>

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

```
>>> How many r's are there in the word "strawberry"? [52/755]
<think>
Okay, so I need to figure out how many times the letter 'r' appears in the word
"strawberry". Let me start by writing down the word and then go through each le
tter one by
one.

The word is s-t-r-a-w-b-e-r-r-y. Wait, let me make sure I have it right:

S-T-R-A-W-B-E-R-R-Y. Hmm, actually, sometimes people might misspell it with an
extra 'r',
but the correct spelling is strawberry without that extra 'r'. So, breaking it
down letter
by letter.

Starting from the beginning:
1. S
2. T
3. R
4. A
```

(DeepSeek, 2025) <http://arxiv.org/abs/2501.12948>

Summary

- **open-domain chat** is still unsolved (full understanding, hallucinations)
 - traditionally **non-task-oriented**: purely for user enjoyment
 - targets: **conversation length** & **user engagement**
 - impersonating a human – Turing test
- approaches
 - **rule-based** – keyword spotting, scripting
 - **retrieval** – copy & paste from large databases
 - **generative** – seq2seq etc. trained on corpora of dialogues
 - too many possible responses don't go well with MLE → safe, short, dull
 - **LLMs** solve a lot of this, but problems remain: hallucinations, controllability
 - **hybrid** – combining all of the above
 - typically mainly rule-based + retrieval, machine learning in NLU only

Contact us:

[https://ufaldsg.slack.com/
odusek@ufal.mff.cuni.cz](https://ufaldsg.slack.com/odusek@ufal.mff.cuni.cz)

Skype/Meet/Zoom (by agreement)

This is the last lecture
Lab in 10 mins
Exams start next week

Get these slides here:

<http://ufal.cz/npfl123>

References/Inspiration/Further:

- Ram et al. (2018): Conversational AI: The Science Behind the Alexa Prize <https://arxiv.org/abs/1801.03604>
- Khatri et al. (2018): Advancing the State of the Art in Open Domain Dialog Systems through the Alexa Prize <https://arxiv.org/abs/1812.10757>
- Shum et al. (2018): From Eliza to Xiaolce: Challenges and Opportunities with Social Chatbots <https://link.springer.com/article/10.1631/FITEE.1700826>
- Vlahos (2018): Inside the Alexa Prize <https://www.wired.com/story/inside-amazon-alexa-prize/>
- Wikipedia: [AIML](#) [Chatbot](#) [Cleverbot](#) [ELIZA](#) [Jabberwacky](#) [Loebner Prize](#) [Mitsuku](#) [PARRY](#) [Turing test](#) [Xiaoice](#) [Zo \(bot\)](#)

Exam

- Written test, 10 questions, 10 points each
 - 50%+ lab exercise points not required to take the test (but needed to get the grade)
 - expected 1 hr, but you'll be given at least 2hrs (no pressure on time)
- Questions covering the 12 lectures
 - question pool on the website
 - you'll need to write stuff on your own (not a-b-c-d, more like 2-3 sentences)
 - explanation of terms/concepts
 - no exact formulas needed (if needed, they might be provided)
 - but you should know the principles of how stuff works
 - relationships between concepts (“what’s the difference between X & Y”)
 - designing a dialogue system for a domain
 - focus on important stuff (mostly what’s mentioned in the summaries)
- Mark: 3:1 weighted exam-lab exercises
 - 60 % = pass (C), 73+0% = B, 88+0% = A