NPFL123 Dialogue Systems
12. Chatbots
(non-task-oriented dialogue)

https://ufal.cz/npfl123

Ondřej Dušek, Patrícia Schmidtová, Vojtěch Hudeček & Jan Cuřín

22. 5. 2023

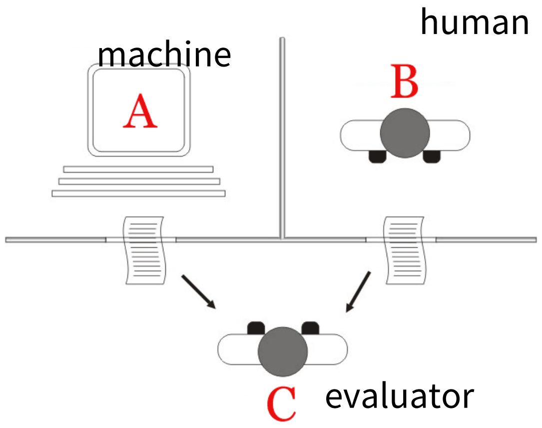
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Chatbots / Chatterbots

- dialogue systems for **open-domain** dialogue / chitchat
- **non-task-oriented**
  - main goal: keep the user entertained
  - standard evaluation: conversation length, user engagement
- (more or less) different architecture
  - may have the same structure as task oriented (NLU → DM → NLG)
  - often simpler, integrated
  - it’s hard to have explicit NLU for open domain
    - no task to guide a meaning formalism
  - some of them don’t need a DB connection (but some use it)
- beware: <u>anything</u> is called chatbots nowadays
  - this lecture: only **chatterbots** / **non-task-oriented systems**

Chatbot tests

- **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+)
  - 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+)
  - 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”

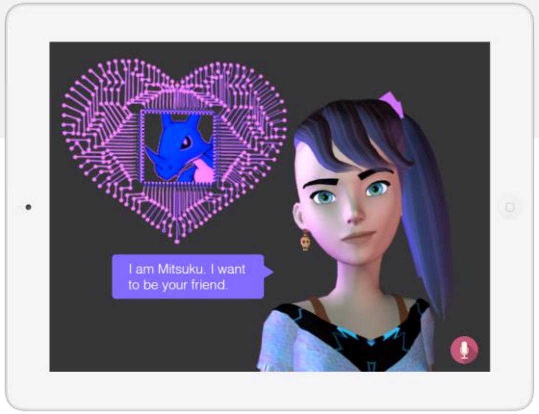
alexa prize

Chatbot 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+ – huge surge of generative models

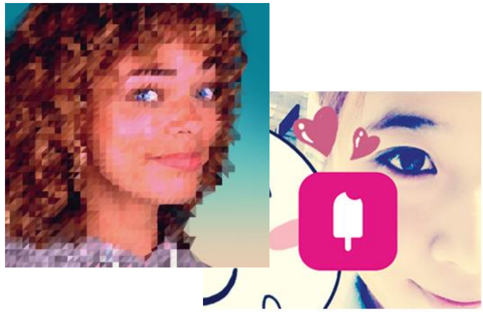
Notable/hyped chatbots

- **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
- **XiaoIce** (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/

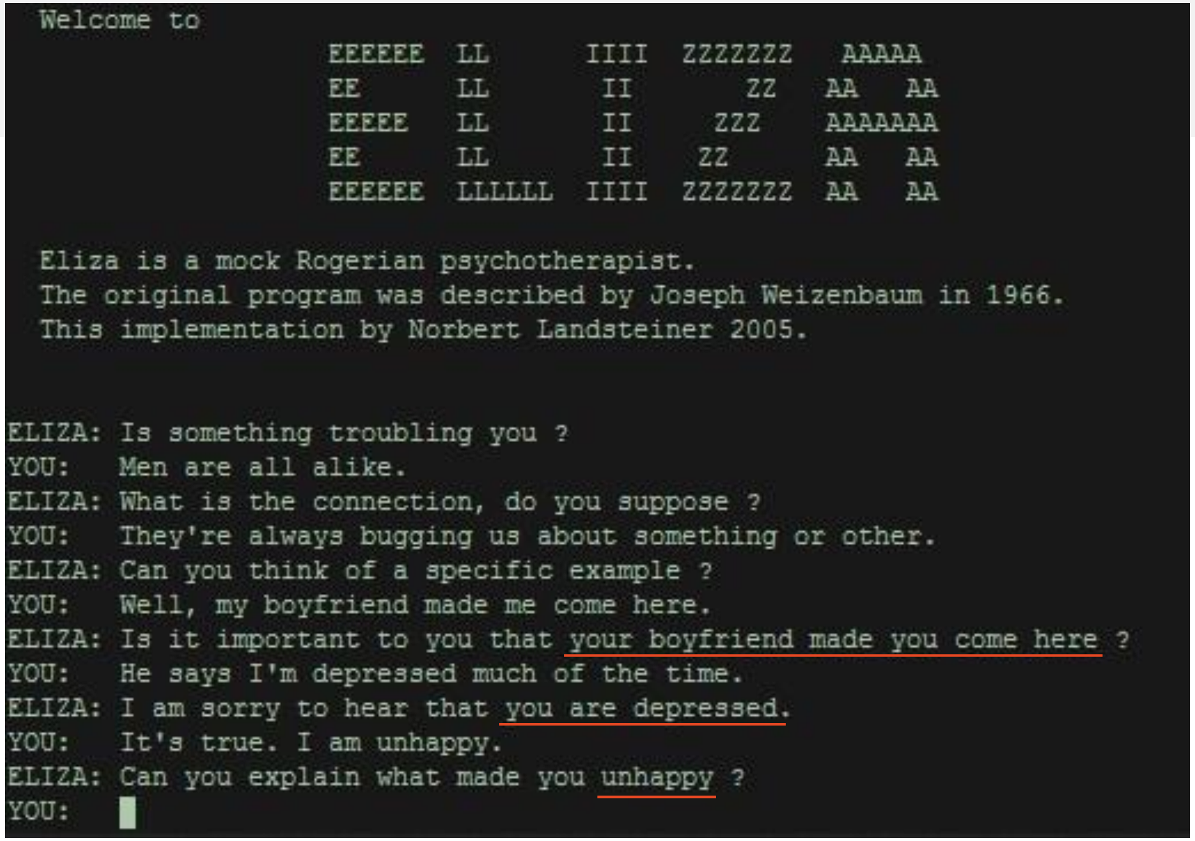
https://youtu.be/z3jqIGT-kmg
http://nautil.us/issue/33/attraction/your-next-new-best-friend-might-be-a-robot

Chatbot basic architectures

- **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/XiaoIce)
    - chatbots differ in the sophistication of reply selection
  - **generative** – (typically) seq2seq-based models
    - 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 chatbots)

- 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
                    EEEEE   LL       II     ZZZ    AAAAAAA
                    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 chatbots
  - 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)

2 categories reduced via srai to the same pattern

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

1/more words

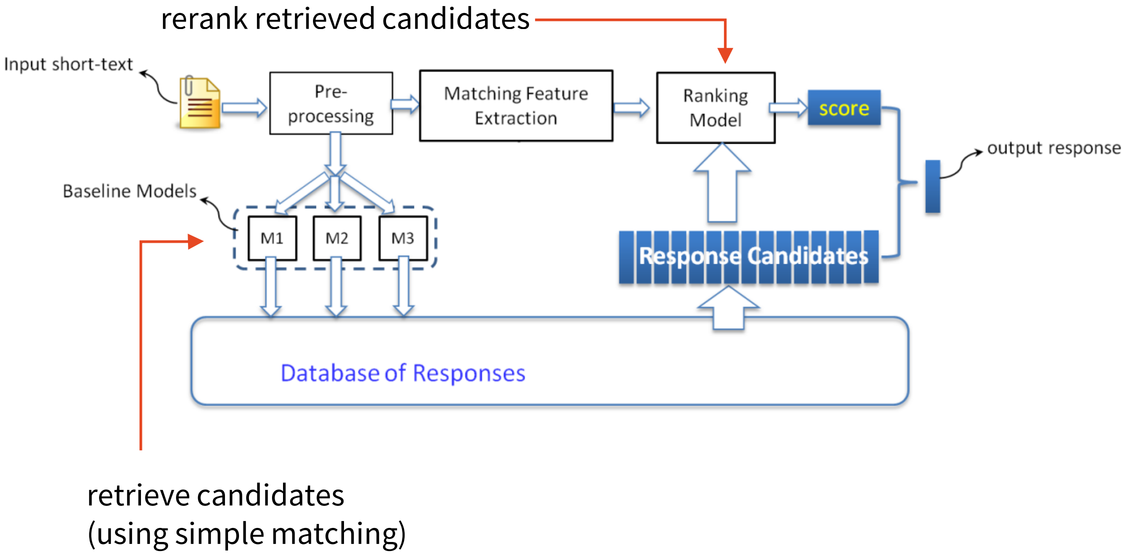
```
<category><pattern>YOU EAT *</pattern>
<template><srai>DIET</srai></template>
</category>
```

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

using a variable

Retrieval-based chatbots

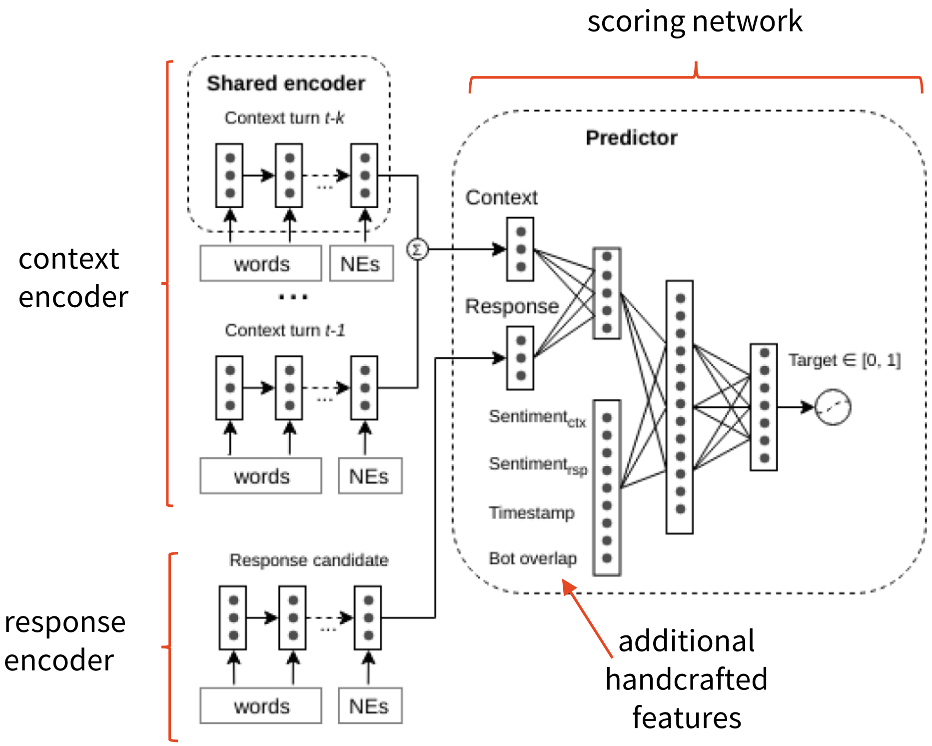
- remember a large corpus
  1) check for similar inputs in the corpus
  2) retrieve & rerank corresponding outputs
- needs 2 steps
  1) rough retrieval
    - needs to be fast to search the whole corpus
    - e.g. TF-IDF
  2) more accurate **reranking** for candidates
    - most research focuses on this 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(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)

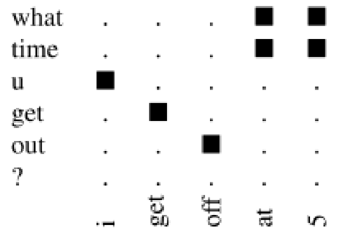


(Shalyminov et al., 2018)
https://www.aclweb.org/anthology/W18-5701

Inspiration from MT

(Ritter et al., 2011)
http://www.aclweb.org/anthology/D11-1054

- Early generative approach – convert to **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



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

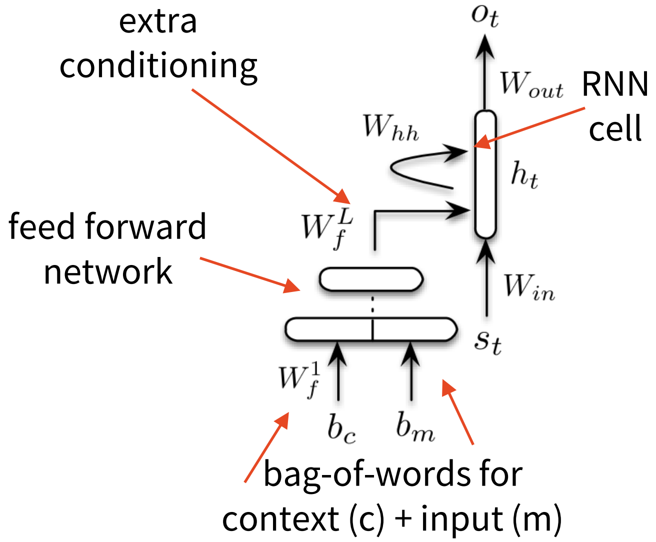
U: *i want to go to a bonfire right about now.*
S: *that does sound fun bonfire i wan na go !*

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 ?*

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, 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



- Vanilla **Seq2seq** 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 .*

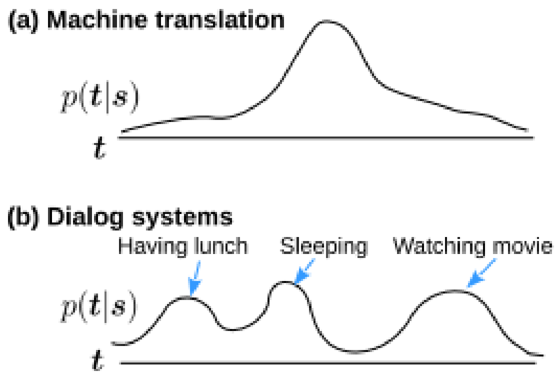
U: *what is the purpose of life ?*
S: *to serve the greater good .*

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

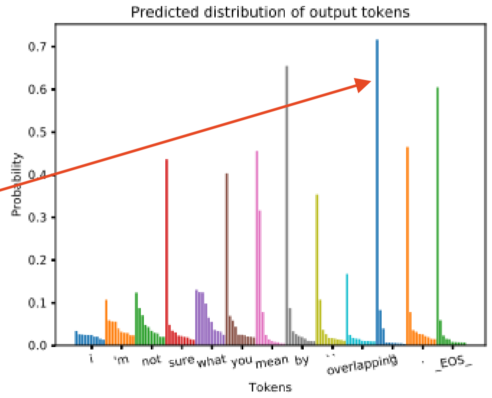
Seq2seq Chatbot Problems

- **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

(Wei et al., 2019)
https://arxiv.org/abs/1712.02250
https://ieeexplore.ieee.org/document/8682634



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?*

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

Improving upon Seq2seq

- **Reranking**
  - 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

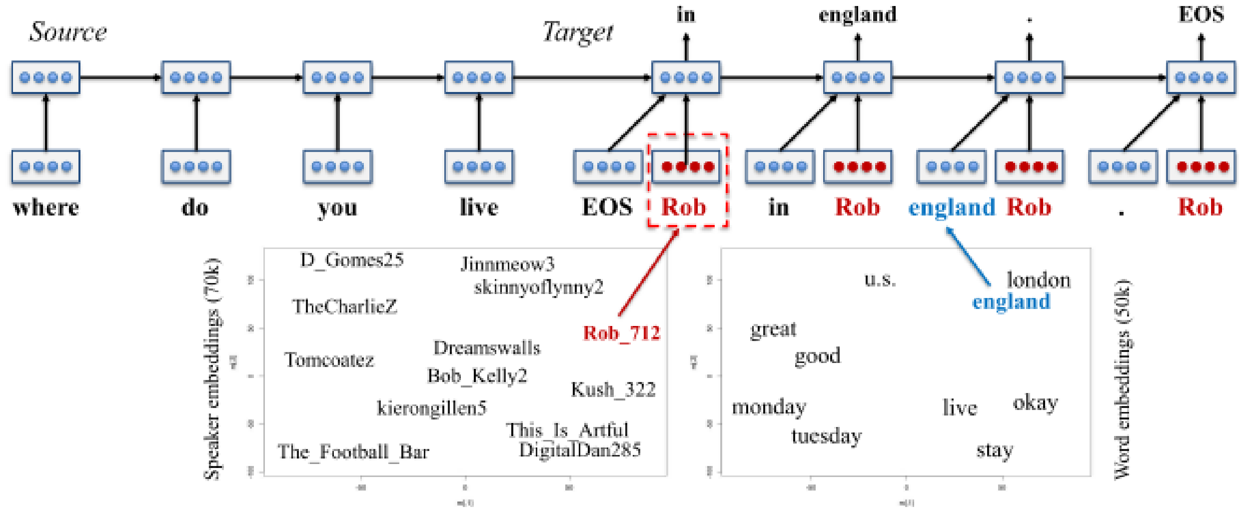
(Li et al., 2016)
https://www.aclweb.org/anthology/N16-1014

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.

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

C = context
R = reply

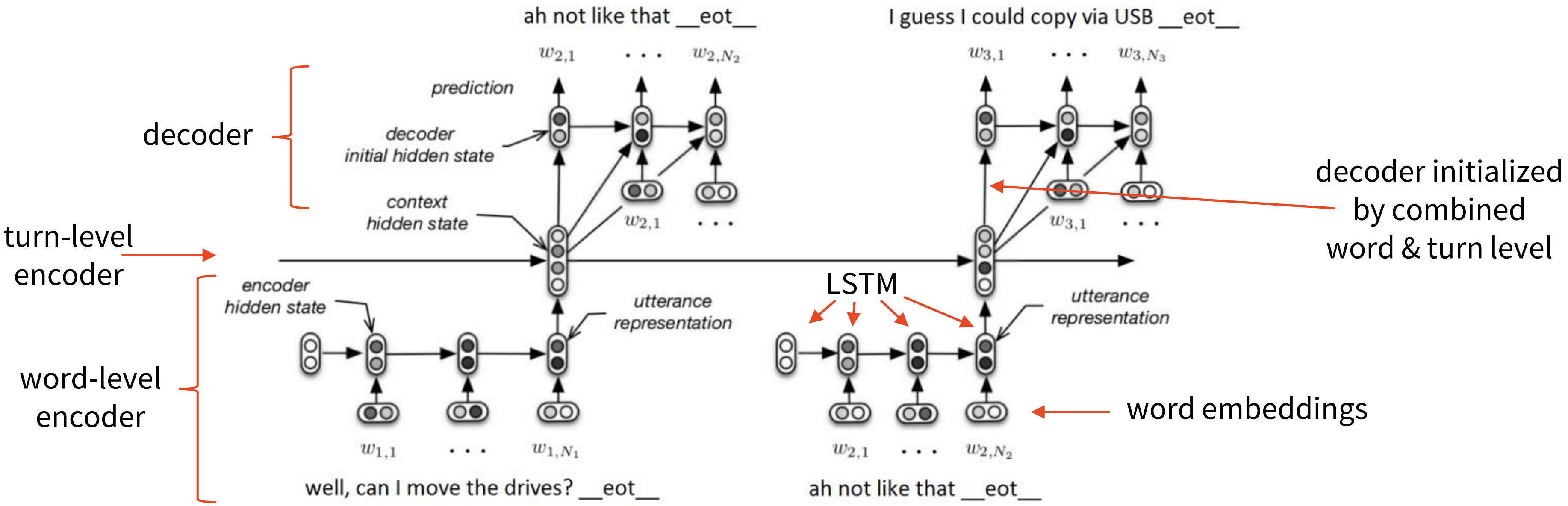
- **Persona extension**
  - improve consistency
  - train speaker embeddings
    - this is a little data-picky
  - use speaker + word embeddings in the decoder
    - can also be used in the encoder



(Li et al., 2016)
https://www.aclweb.org/anthology/P16-1094

Improving upon 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

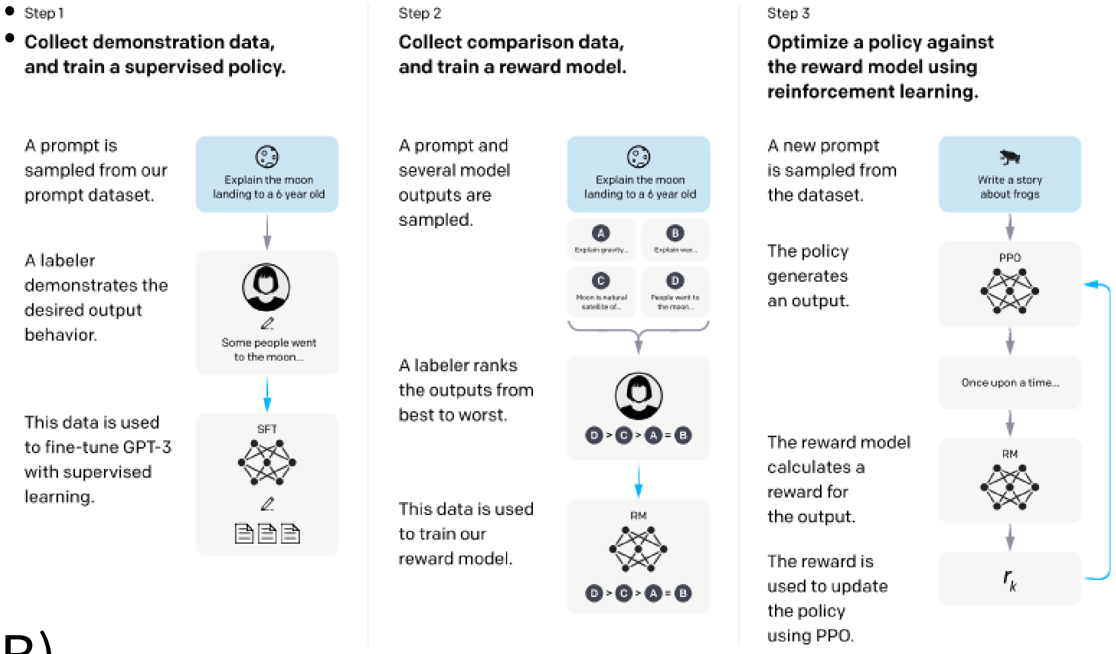


Transformer Chatbots

- **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** (chitchat SotA now) (Roller et al., 2021) https://aclanthology.org/2021.eacl-main.24/
  - huge encoder-decoder Transformer (has smaller versions)
  - 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

Instruction finetuning

- Pretrained LMs finetuned on instruction & solution pairs
  - typically starting from non-dialogue-specific model (GPT3 → **InstructGPT**, GPT3.5/4 → **ChatGPT**, LlaMA → **Alpaca**, **OpenAssistant**…)
  - Kinda task-oriented, but open-domain & unstructured
- Training from human feedback (aka RLHF):
  1) General fine-tuning, next-word prediction on instruction data
  2) Get lots of outputs evaluated by humans & train a reward model based on that
  3) Apply RL with reward model as finetuning
  - Makes models more efficient
  - The main point is global loss, not RL
  - Answers still often hallucinated (no external DB)



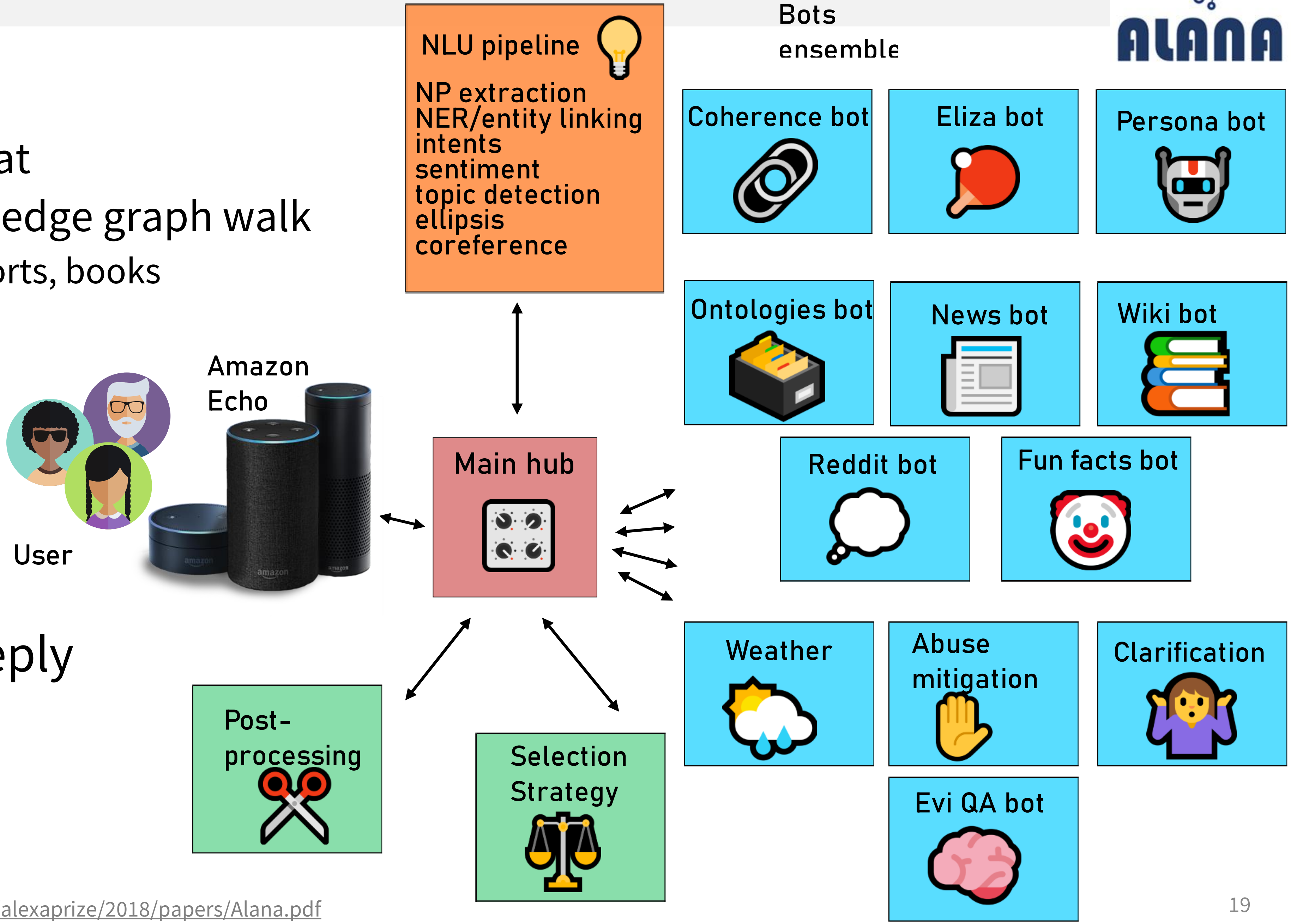
(Ouyang et al., 2022) http://arxiv.org/abs/2203.02155 (Köpf et al., 2023) http://arxiv.org/abs/2304.07327 (Zhao et al., 2023) http://arxiv.org/abs/2305.10425
https://openai.com/blog/chatgpt/ , https://open-assistant.io/chat , https://www.youtube.com/watch?v=0A8ljAkdFtg

Hybrid / Ensemble Chatbots (most Alexa Prize Entries)

- "Production" SotA (~safer than ChatGPT et al.): **combining all methods**
  - 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

Alana (Heriot-Watt University, 2017 & 2018 3rd)

- Bots:
  - Rule-based chit-chat
  - Ontologies – knowledge graph walk
    - movies, music, sports, books
  - Retrieval
    - Reddit trivia
    - news
    - Wikipedia
    - fun facts
  - Specific services
- Bots compete for reply
  - priority list
  - bots can “lock”



http://arxiv.org/abs/1712.07558
http://dex-microsites-prod.s3.amazonaws.com/alexaprize/2018/papers/Alana.pdf

Alana Example dialogue

coherence: stay on topic
& ask further questions

Coherence bot

Alana: Welcome back then John. Since I remember that you like movies, [...] What movie are you looking forward to watching?

User: Vanilla Sky –> I'm looking forward to watching Vanilla Sky **[Ellipsis resolution]**

Ontologies bot

Alana: Oh I heard something interesting about Vanilla Sky. **[… Movie trivia]** Speaking of movies. What actors do you like? I like Leonardo DiCaprio […]

User: Interesting

Eliza bot

Alana: I just think the world is full of strange and unexpected things like that. So, talking about movies, What famous actor or actress would you like to meet? […]

User: I want to talk about music

Coherence bot

Alana: So, talking about music. Who is a singer that you like? I really like Taylor Swift […]

User: Kendrick Lamar

Ontologies bot

Alana: Yes. One of my friends once told me something interesting about Kendrick Lamar. **[… Music trivia]** What's a band that you would like to see? […]

User: any news about him –> any news about Kendrick Lamar **[Coreference resolution]**

Alana Example dialogue

user model:
name + preferred topics

bot personality

News bot

Alana: Hmm... about Kendrick Lamar, so John, I read a story with this headline **[... News headline]** Shall I say some more about this?

User: could we talk about something else

Coherence bot

Alana: So John, since I know you like books, what have you read recently? I am reading [...]

User: Let's talk about Tom Hardy

Clarification

Alana: Oh I know Tom Hardy. Do you mean Tom Hardy the English actor, screenwriter and producer or Thomas Hardy the English novelist and poet?

User: I mean the novelist

Ontologies bot

Alana: Yes. Some friends told me something interesting about Thomas Hardy. **[... Trivia]**

User: What do you know about pizza?

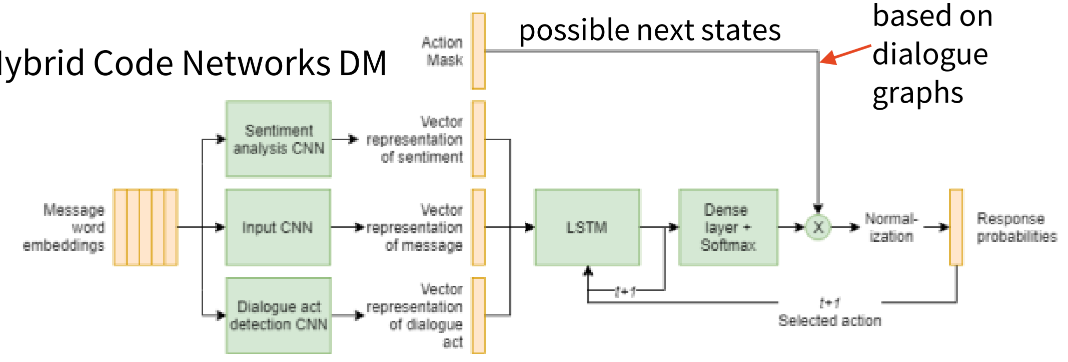
Reddit bot

Alana: I was thinking that whenever a pizza commercial has stretchy melty cheese, it looks good and appealing, but when it happens in real life it's annoying and messy.

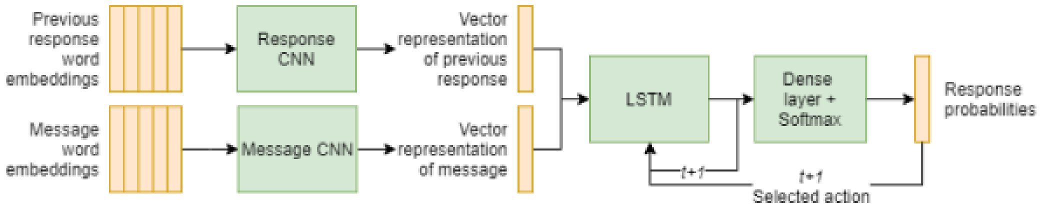
Alquist (Czech Technical University, ‘17+‘18 2nd, ‘19/20 3rd, ‘20/21 1st)

- full NLU pipeline (similar to Alana)
- 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

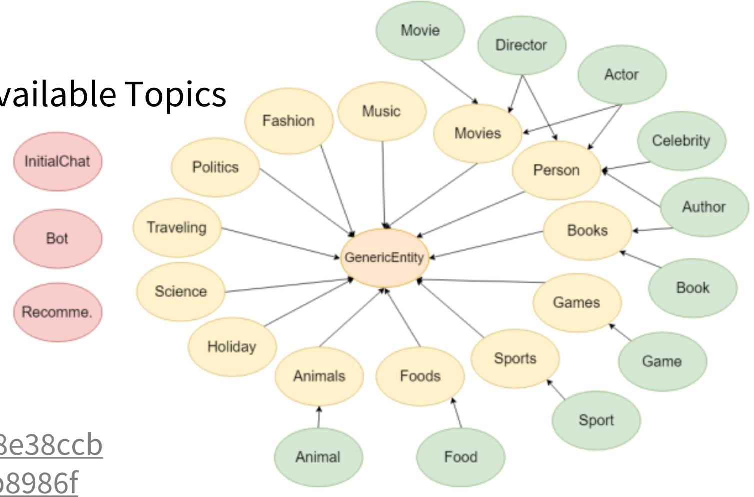
Hybrid Code Networks DM



Topic Switch Detector



Available Topics



http://alquistai.com/

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?
negotiation	2	I'm five.	Glad to hear it! We could talk about robots, Batman, or Superman. Say "next", to chat about other things.
thoughts	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?
facts	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"
movies	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?

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

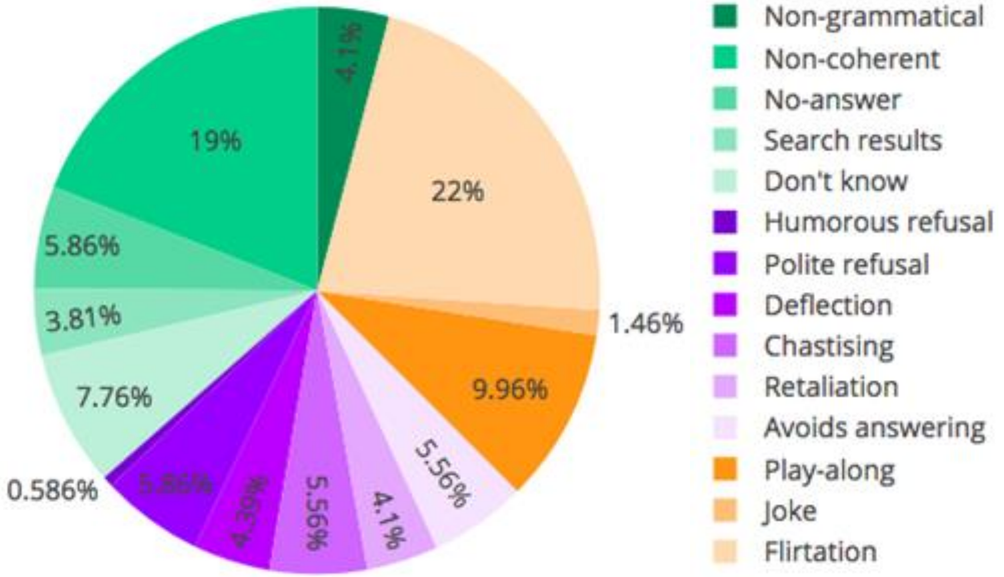
Alexa Prize bottom line

- understanding is the bottleneck
  - ASR problems – chat-specific ASR improved things, but it's by far not perfect
  - vague concept of dialogue state, despite full NLP pipelines
    - result: typically very crude intents + list of named entities
  - recognizing multiple/fine-grained intents is a problem
- it's still more about social engineering than "AI"
  - a lot of strategies for not-understanding (switching topics, questions…)
- machine learning helps, but pure ML is not enough
  - lack of annotated data → often relatively simple methods
  - ML helps mainly in NLU, end-to-end seq2seq doesn't work
- interesting content is crucial
  - the more handcrafted topics, the better
  - fluent NLG not so much (but prosody helps!)
- brutal variance in the evaluation – very subjective

Chatbot Ethics

(Cercas Curry & Rieser, 2018)
http://aclweb.org/anthology/W18-0802

- Ca. 4% of our 2017 data is sexually explicit
- Different harassment types:
  - comments on gender/sexuality
  - sexualized comments
  - sexualized insults
  - sexual requests & demands
- Chatbots/voice assistants' responses
  - various systems:
    - commercial (Alexa, Google…)
    - rule-based (Pandorabots, adult chatbots)
    - data-driven (seq2seq)
  - systems often present as women, have a woman's voice
  - responses often nonsense / play-along
    - conflict of interest for bot builders: be ethical vs. cater to abusive users



Summary

- chatbots – **non-task oriented** systems
  - 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
  - **hybrid** – combining all of the above
    - typically mainly rule-based + retrieval, machine learning in NLU only
- open-domain NLU is still an unsolved problem
  - despite that, many people enjoy conversations with chatbots
  - interesting content is crucial

Thanks

This is the Last Lecture
Lab in S4 in 10 mins
Next week: exam date

Contact us:

https://ufaldsg.slack.com/
{odusek,hudecek}@ufal.mff.cuni.cz
Skype/Meet/Zoom (by agreement)

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 XiaoIce: 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+% = B, 88+% = A