NFPL099 Statistical Dialogue Systems

10. Chatbots (non-task-oriented)

http://ufal.cz/npfl099

Ondřej Dušek, Vojtěch Hudeček & Zdeněk Kasner

5.12.2022

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

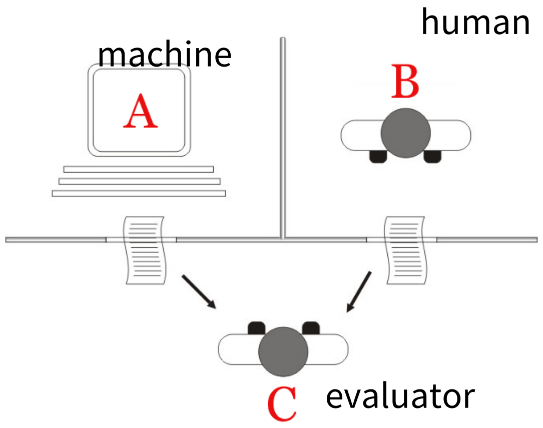
BY SA
unless otherwise stated

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 – somewhat like end-to-end DS
  - 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: *anything* can be called a “chatbot” nowadays
  - here: 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



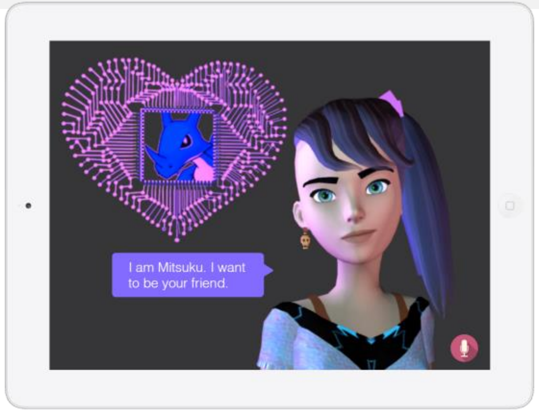
- **Amazon Alexa Prize** (2017+)
  - no pretending it's human, just coherent & engaging conversation for 20 mins.
  - topic semi-restricted ("on popular topics")
  - evaluator & 3 judges with stop-buttons
  - score: duration + 1-5 scale of "would talk again"

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

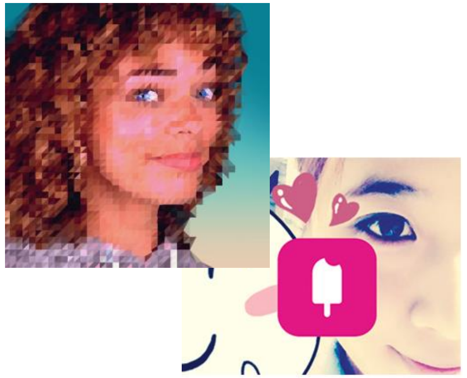


https://home.pandorabots.com/home.html

- **Jabberwacky/Cleverbot** (1997+)
  - attempts to learn from users
  - remembers & reuses past conversations (>100M)
  - also won Loebner Prize multiple times

https://www.cleverbot.com/

- **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://www.zo.ai/
https://www.facebook.com/zo/

https://youtu.be/z3jqIGT-kmg

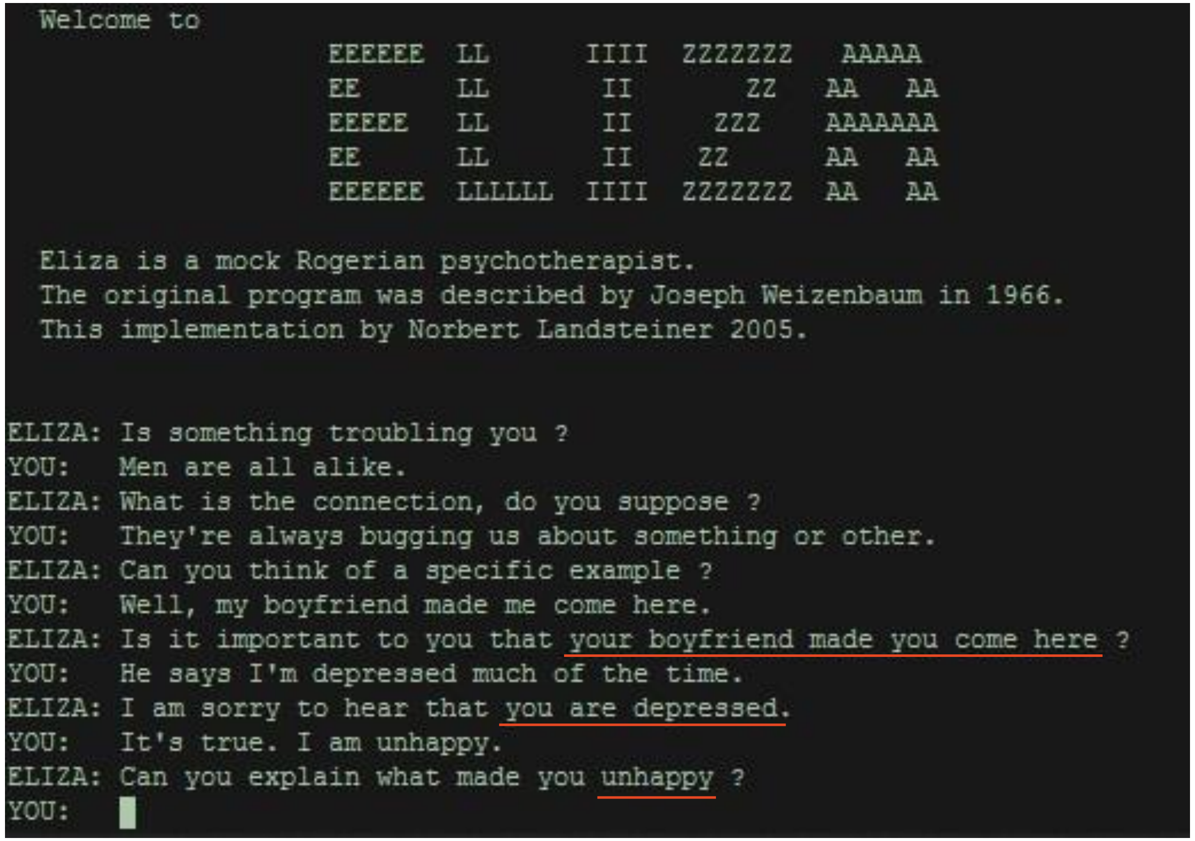
http://nautil.us/issue/33/attraction/your-next-new-best-friend-might-be-a-robot

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

1/more words

```
<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>
```

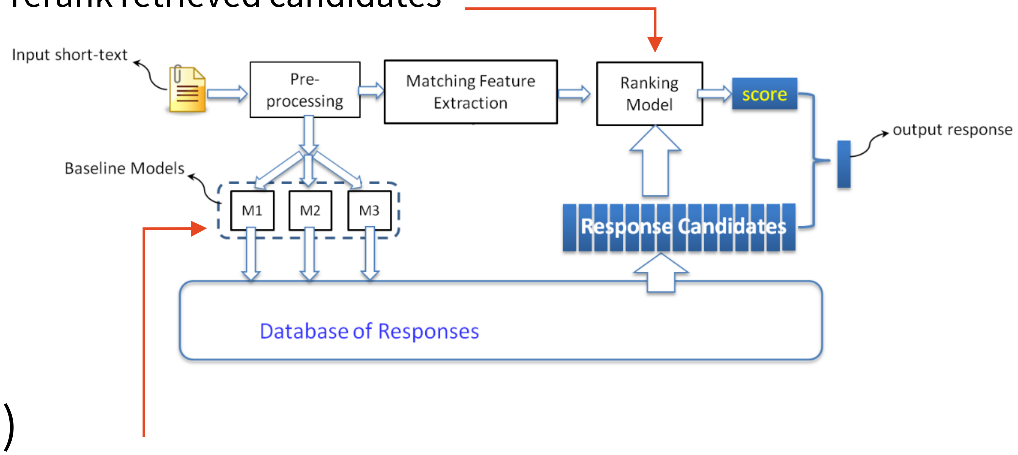
```
<category><pattern>DIET</pattern>
<template>My diet consists mostly of <bot name="diet"/>.</template>
</category>
```

using a variable

Retrieval-based 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

rerank retrieved candidates



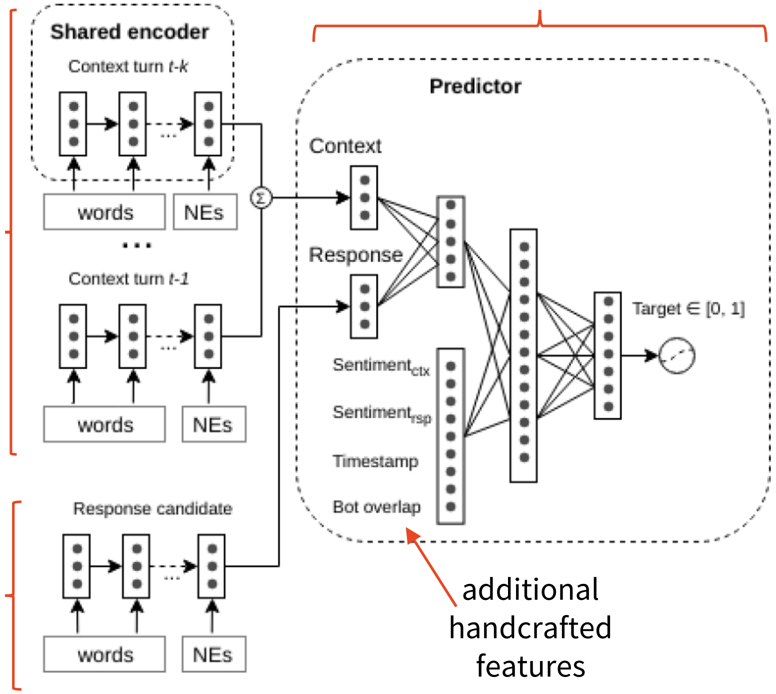
retrieve candidates
(using simple matching)

(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 / Siamese network**
      - 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)

scoring network



context encoder

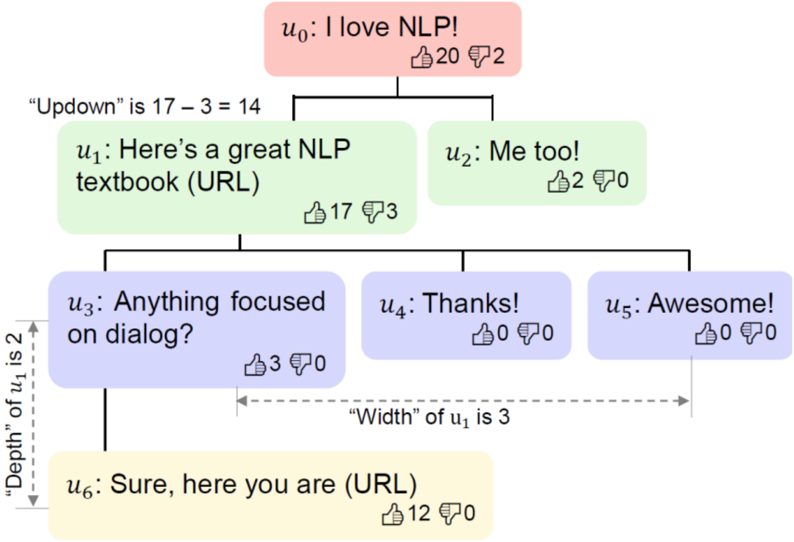
response encoder

additional handcrafted features

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

DialogRPT: Ranking with pretrained LMs

- GPT-2-based model (DialoGPT →)
- contrastive learning: feed 1 context + 2 responses, classify which is better
  - binary classification layer on top of last token
  - better than classify good/bad in isolation
- self-supervised data for ranking from Reddit
  - width = # of direct replies
  - depth = max. dialog length in subtree
  - updown = 👍-👎 votes



Generative chatbots: RNNs

(Vinyals & Le, 2015)
https://arxiv.org/abs/1506.05869

- Generic LSTM-based **Seq2seq** model
  - encode context / input, decode response
  - examples in paper look great, but… (→)

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 .*

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.*

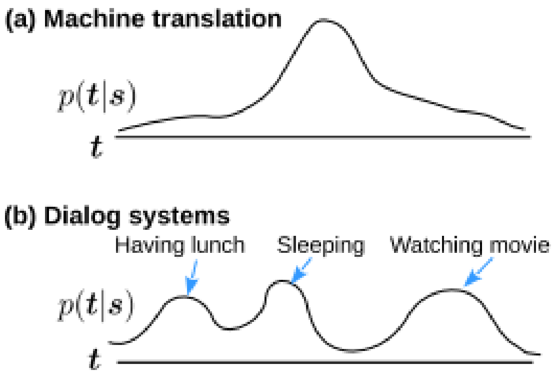
C[S]: *thanks. yeah i am good thanks! good, but angry!*
U: *sorry you're angry, been there, i hope it dies down.*
S: *i hope so too, i'm not angry!!*

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

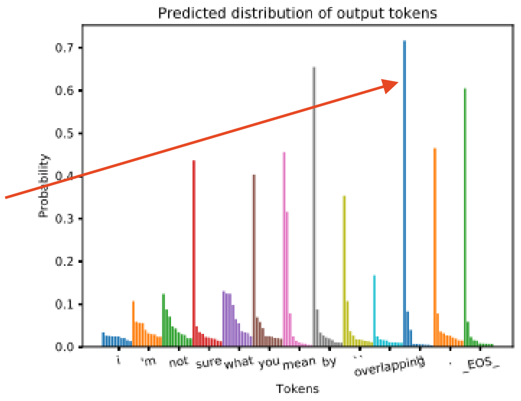
Seq2seq Chatbot Problems

- **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

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

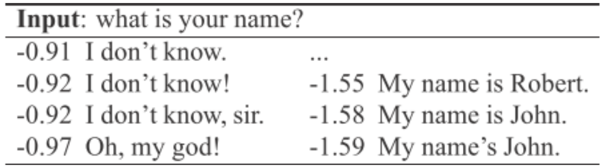
U: *How old are you?*
S: *16 and you?*
U: *What's your age?*
S: *18.*

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?*

Diversity/Coherence

- **Reranking**: MMI
  - avoid dull replies that work anywhere
  - instead of maximizing $P(Resp|Context)$, **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

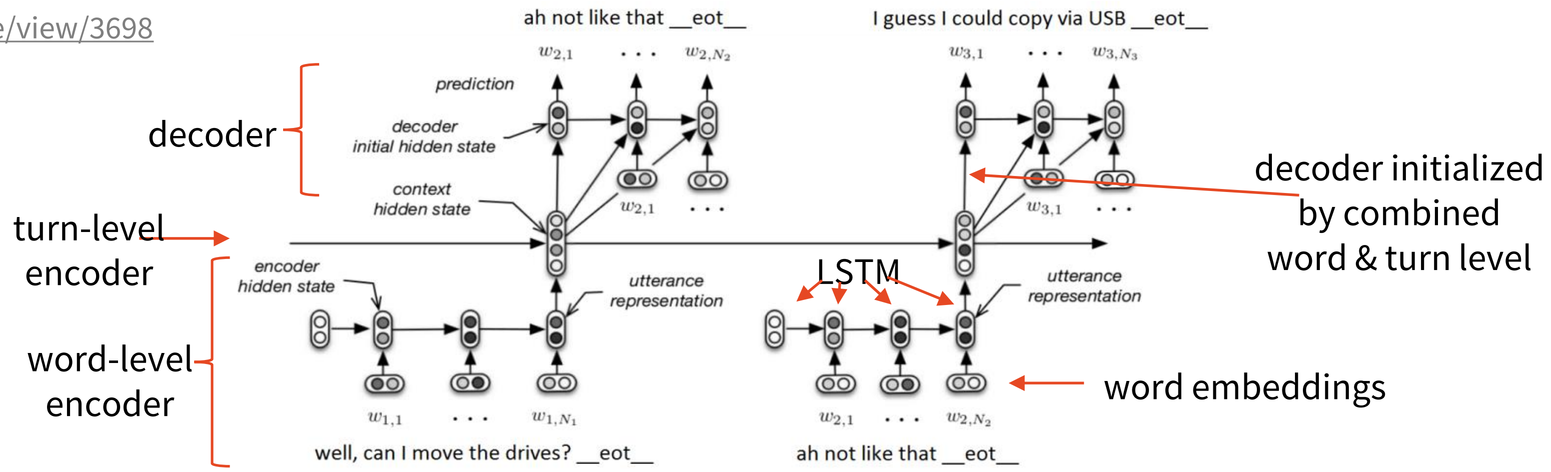


Input: what is your name?			
-0.91	I don't know.	...	
-0.92	I don't know!	-1.55	My name is Robert.
-0.92	I don't know, sir.	-1.58	My name is John.
-0.97	Oh, my god!	-1.59	My name's John.

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

- **Longer context**: HRED (Hierarchical Recurrent Encoder-Decoder)
  - 2nd, turn-level LSTM encoder, with word-level LSTM hidden state as input

(Lowe et al., 2017)
http://dad.uni-bielefeld.de/index.php/dad/article/view/3698



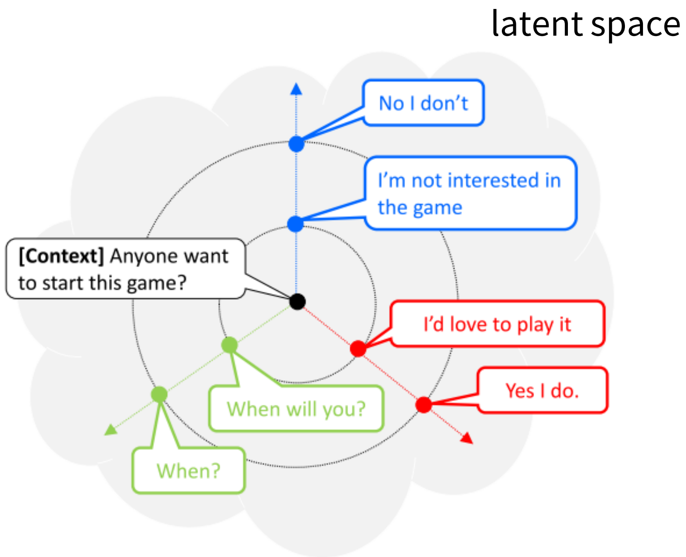
decoder initialized by combined word & turn level

word embeddings

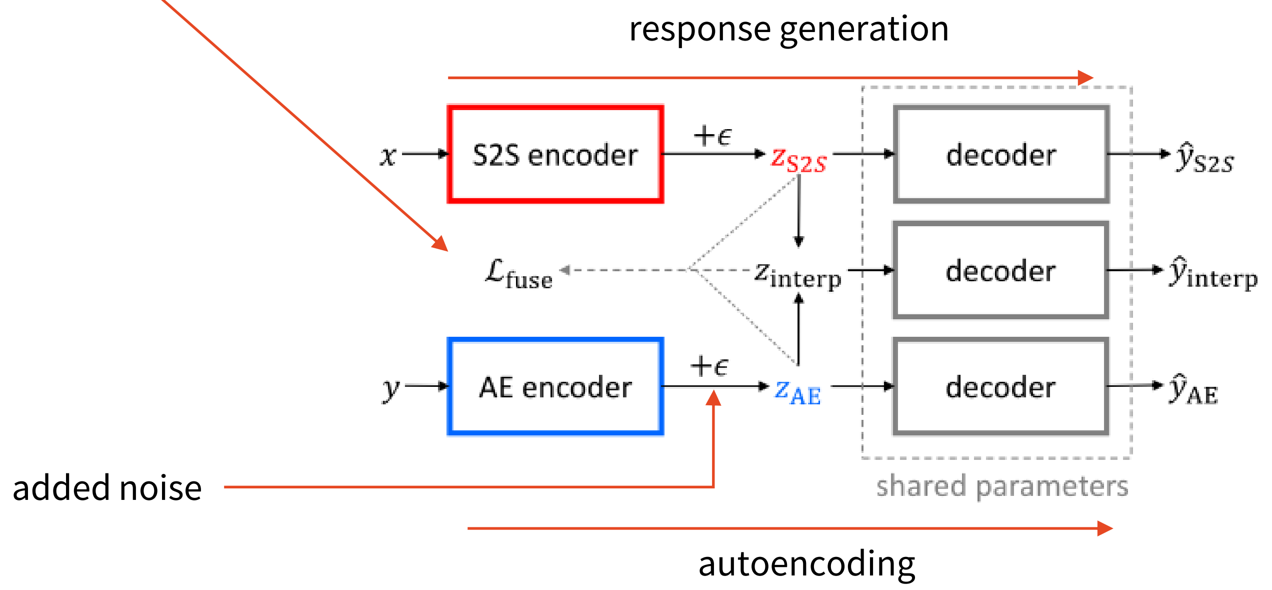
Diversity

(Gao et al., 2019)
http://arxiv.org/abs/1902.11205

- joining **next turn generation** & **autoencoding**
  - LSTM VAE-like model, shared latent space
  - multi-task learning (see last lecture)
  - shared decoder
  - additional "fusion loss" enforcing the same encoding for both tasks
- inference: adding a little noise to encodings
  - to produce different outputs



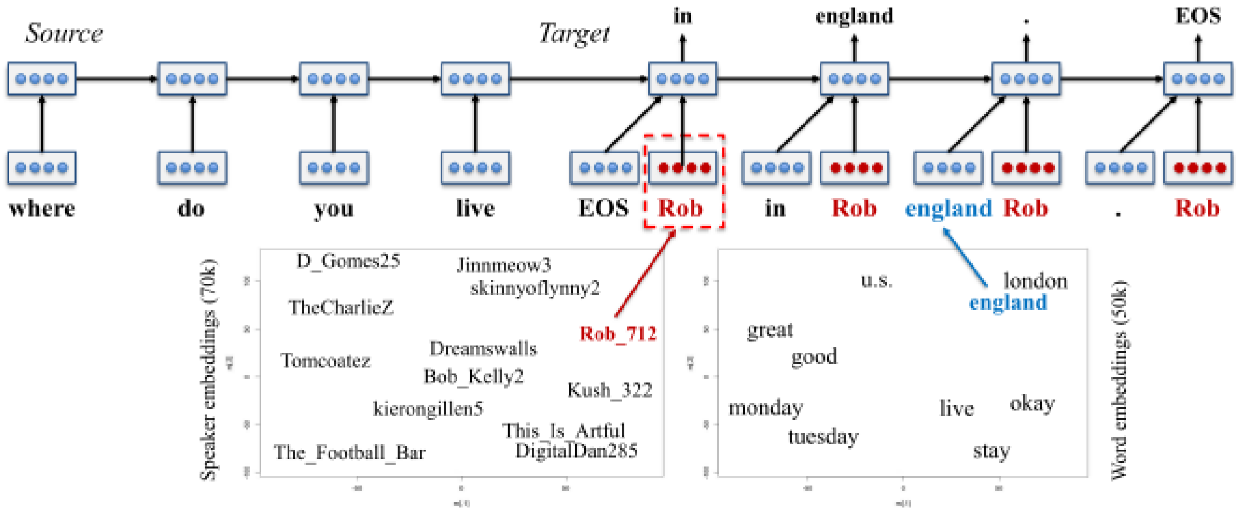
latent space
No I don't
I'm not interested in the game
[Context] Anyone want to start this game?
I'd love to play it
Yes I do.
When will you?
When?



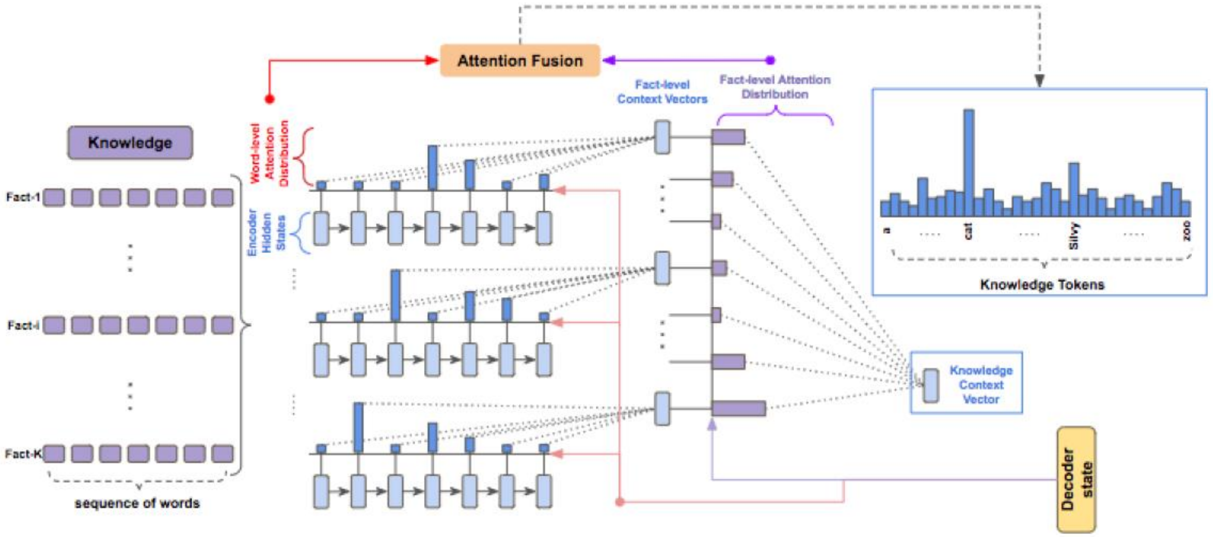
Consistency/Personality

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

- improving consistency by modelling chatbot’s personality
- **Persona embeddings**
  - train speaker embeddings
  - use speaker + word embeddings in the decoder
  - needs lots of data
- **Persona copy-net**
  - add & attend to personal bio in context
    - chunks of text
  - copy-net or pretrained LMs



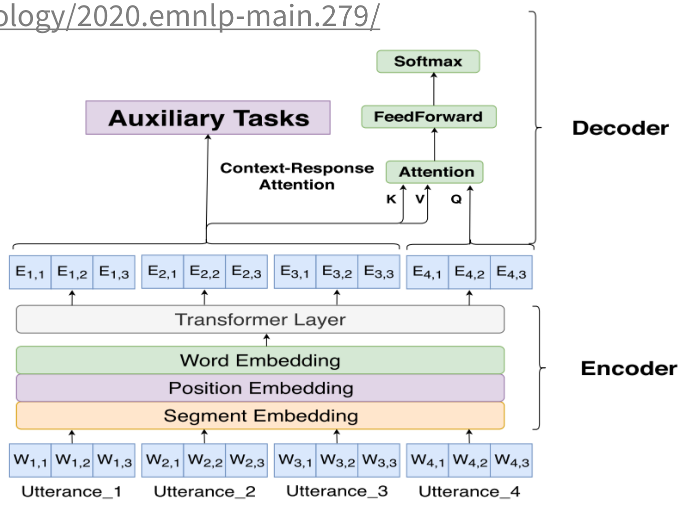
(Yavuz et al., 2019)
https://www.aclweb.org/anthology/W19-5917/



Coherence: Additional Objectives

- Transformer-based architectures
- **Denoising** (autoencoder): additional decoders
  - shuffled word order
  - masked words
  - masked utterance (mid-dialogue)
  - utterance order (GRU decoding order)

(Zhao et al., 2020) https://www.aclweb.org/anthology/2020.emnlp-main.279/


Softmax
Auxiliary Tasks
FeedForward
Decoder
Context-Response Attention
Attention
K
V
Q
$E_{1,1}$ $E_{1,2}$ $E_{1,3}$ $E_{2,1}$ $E_{2,2}$ $E_{2,3}$ $E_{3,1}$ $E_{3,2}$ $E_{3,3}$ $E_{4,1}$ $E_{4,2}$ $E_{4,3}$
Transformer Layer
Word Embedding
Position Embedding
Segment Embedding
Encoder
$W_{1,1}$ $W_{1,2}$ $W_{1,3}$ $W_{2,1}$ $W_{2,2}$ $W_{2,3}$ $W_{3,1}$ $W_{3,2}$ $W_{3,3}$ $W_{4,1}$ $W_{4,2}$ $W_{4,3}$
Utterance_1
Utterance_2
Utterance_3
Utterance_4

- **Unlikelihood** – demoting unlikely tokens
  - penalize set of tokens selected at each time step
  - repeating n-grams, too much high-freq. vocab…
  - weighted combination with regular MLE loss

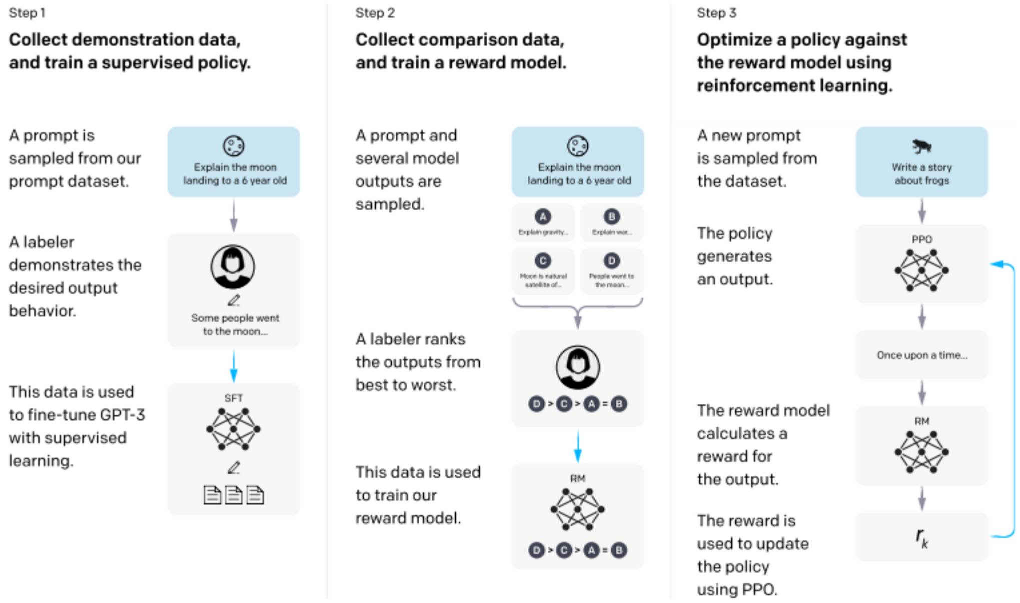
(Li et al., 2020) https://www.aclweb.org/anthology/2020.acl-main.428

Chat-Specific Pretrained Language Models

- **DialoGPT** – GPT-2 finetuned on Reddit (147M dialogues) (Zhang et al., 2020) https://www.aclweb.org/anthology/2020.acl-demos.30
  - no hierarchy, 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
  - "Evolved Transformer" architecture (Transformer + small changes automatically tuned)
  - encoder-decoder, huge, trained on 867M dialogues (next-word prediction)
  - rule-based postprocessing
  - evaluation: "making sense" & "being specific" – better on both
- **BlenderBot** (sorta SotA now) (Roller et al., 2021) https://aclanthology.org/2021.eacl-main.24/
  - again, huge Transformers (but has a smaller version)
  - retrieval & generative versions
  - pretrained on Reddit, finetuned on a combination of specific dialogue datasets
  - constrained beam search (avoid too short replies), better than sampling

Chat-specific PLMs (2)

- **LaMDA**: LM + retrieval + “calculator” (Thoppilan et al., 2022) http://arxiv.org/abs/2201.08239
  - pretrained on dialogue
  - finetuned on corrections of its own outputs
  - usage of retrieval & calculator annotated
  - generate multiple, filter (safety) & rerank
  - 2B/137B params versions
- **InstructGPT/ChatGPT:** GPT3 + feedback
  - multi-step training (from GPT3/code model)
  1) GPT3 supervised finetuning
  2) evaluation/ranker model
  3) RL with proximal policy optimization
  - more truthful, can be made safer (via feedback)
  - 1.3B params InstructGPT > 175B GPT3



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

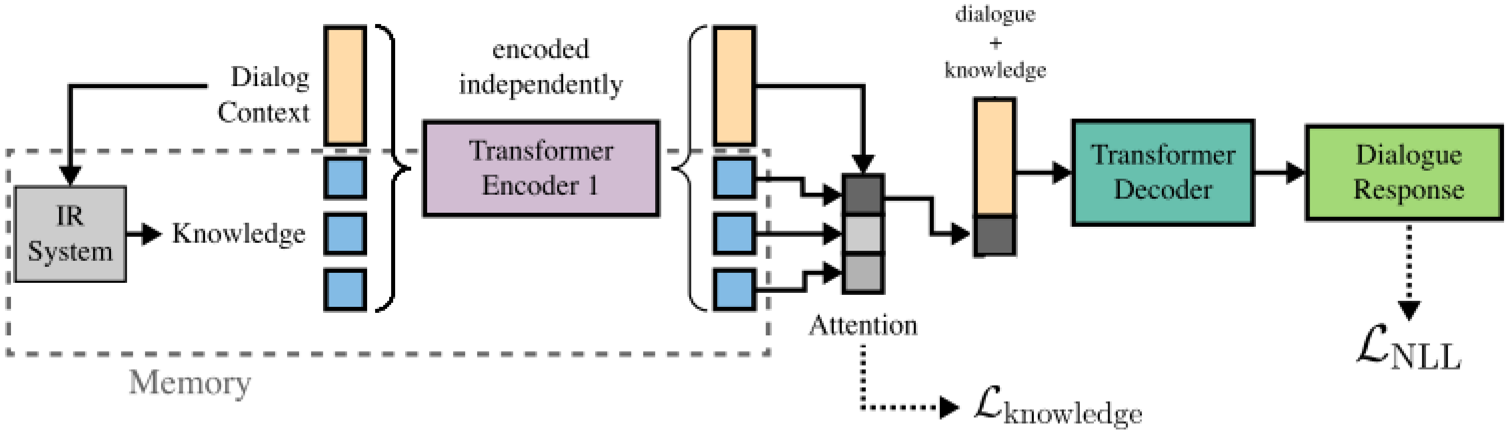
The reward is used to update the policy using PPO.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

(Ouyang et al., 2022) http://arxiv.org/abs/2203.02155
https://openai.com/blog/chatgpt/ , https://www.youtube.com/watch?v=0A8ljAkdFtg

Retrieval-augmented bots

- Combination of generation & retrieval
  1) **Retrieve** a candidate,
  2) **Edit** it using a seq2seq model to better match context
- Knowledge grounding
  - candidate = knowledge to be used in response
  - Wizard-of-Wikipedia



(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

- Problem: right amount of copying
  - Don’t ignore the retrieved
  - Don’t copy it verbatim
  - Question of parameters, tradeoff, various hacks to achieve this
  - α-blending: replace retrieved with target with some probability, to promote copying

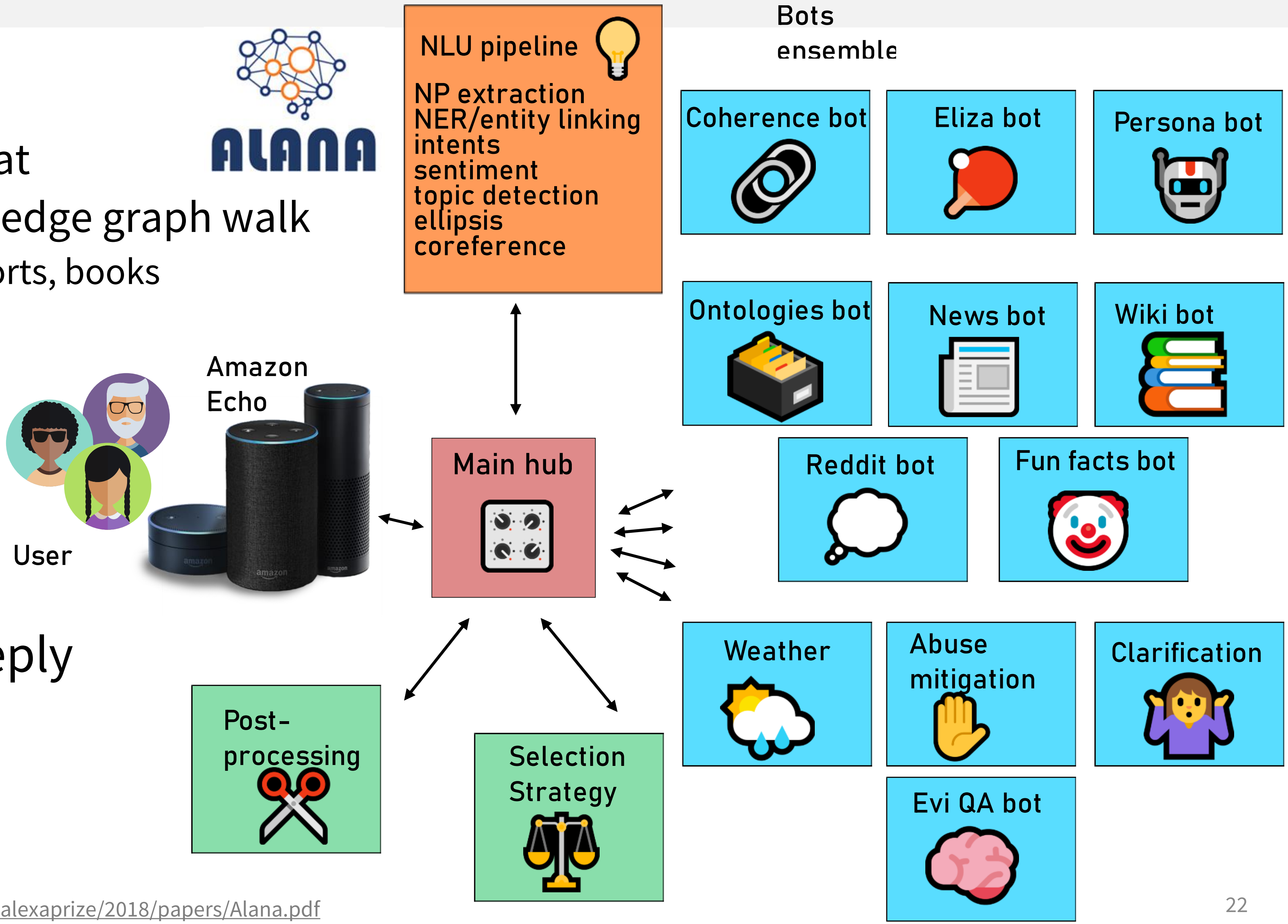
(Roller et al., 2021) https://aclanthology.org/2021.eacl-main.24

Hybrid / Ensemble Chatbots (a.k.a. most Alexa Prize Entries)

- State-of-the-art today: **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
  - based on NLU topic detection
  - 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]**

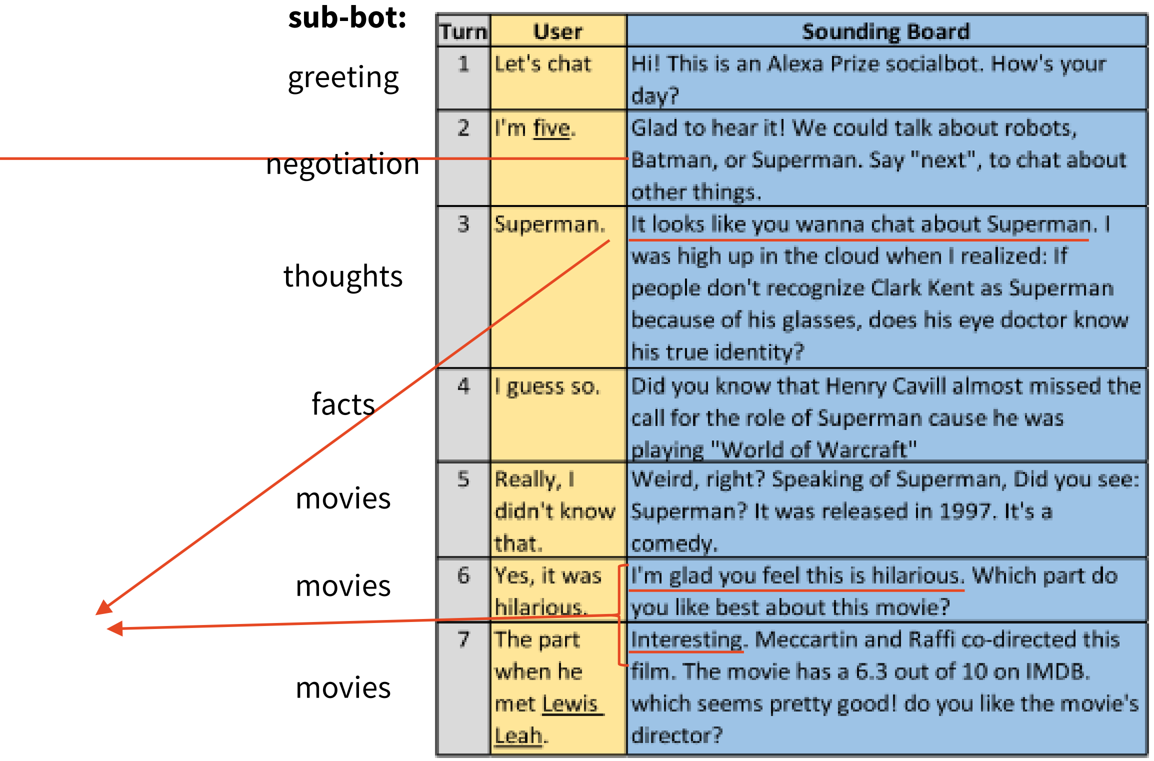
User: What do you know about pizza?

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.

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

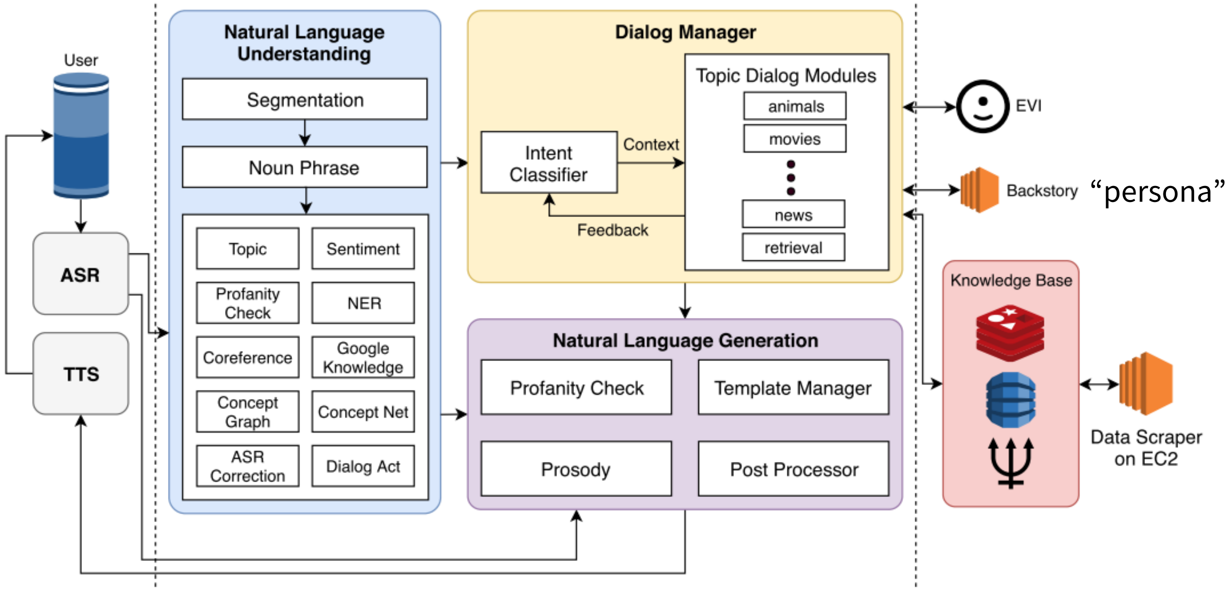


sub-bot:	Turn	User	Sounding Board
greeting	1	Let's chat	Hi! This is an Alexa Prize socialbot. How's your day?
negotiation	2	I'm five.	Glad to hear it! We could talk about robots, Batman, or Superman. Say "next", to chat about other things.
thoughts	3	Superman.	It looks like you wanna chat about Superman. 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/

Gunrock (UC Davis, 2018 winner)

- Improving ASR
  - error correction – KB fuzzy matching (allow for “typos”)
  - sentence segmentation (RNN-based)
- NLU – keyphrase extraction
  - focus on noun phrases
- Dialog manager – stack
  - return to previous topics
  - related topics
  - a lot of different topics with domain-specific KBs
    - games, psychology, travel…



“persona”

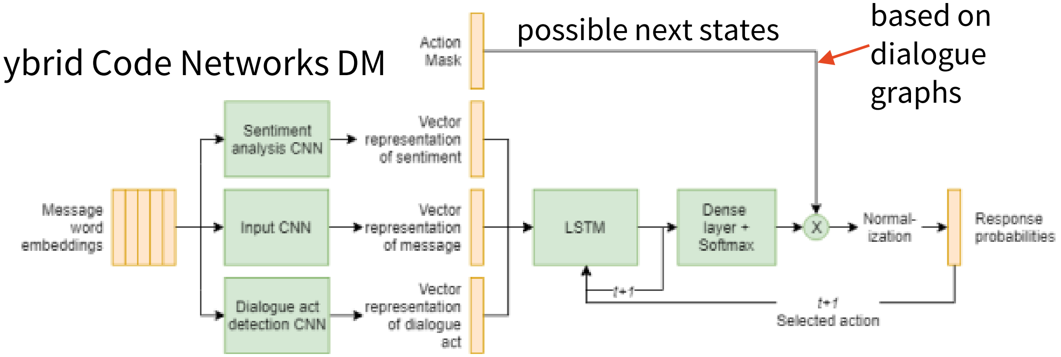
http://dex-microsites-prod.s3.amazonaws.com/alexaprize/2018/papers/Gunrock.pdf

Alquist (Czech Technical University, 2017&2018 2nd)

ALQUIST

- full NLU pipeline (similar to Alana)
- 2017 – handcrafted state machines
  - traversing sub-dialogue graphs
    - dividing for easier maintenance
  - well scripted
    - easy to break, but users play along
    - hand-added variation
- 2018 – adding machine learning
  - Hybrid Code Networks
    - RNN-based dialogue management
    - for each sub-dialogue/topic
  - topic switch detector
    - RNN-based architecture similar to HCN

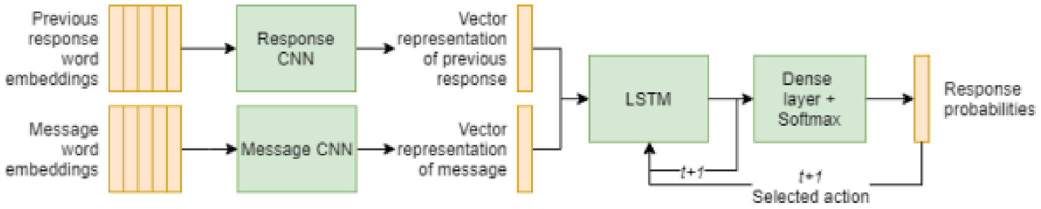
Hybrid Code Networks DM



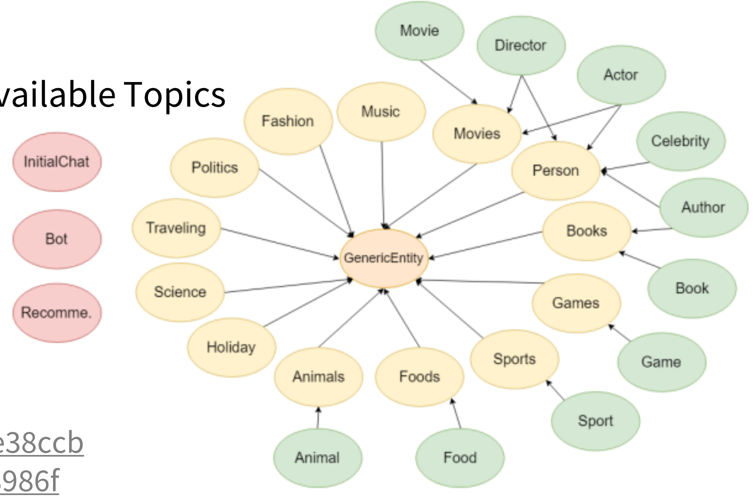
possible next states

based on dialogue graphs

Topic Switch Detector



Available Topics



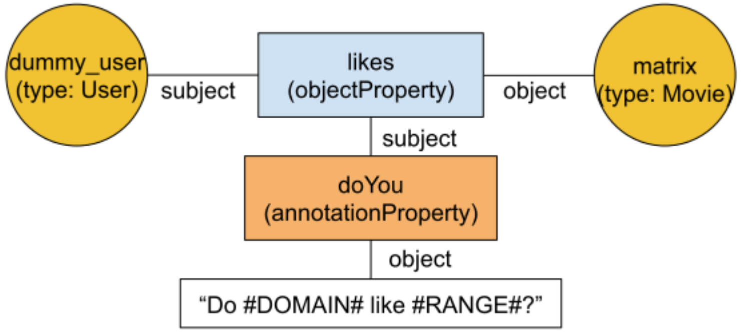
http://alquistai.com/

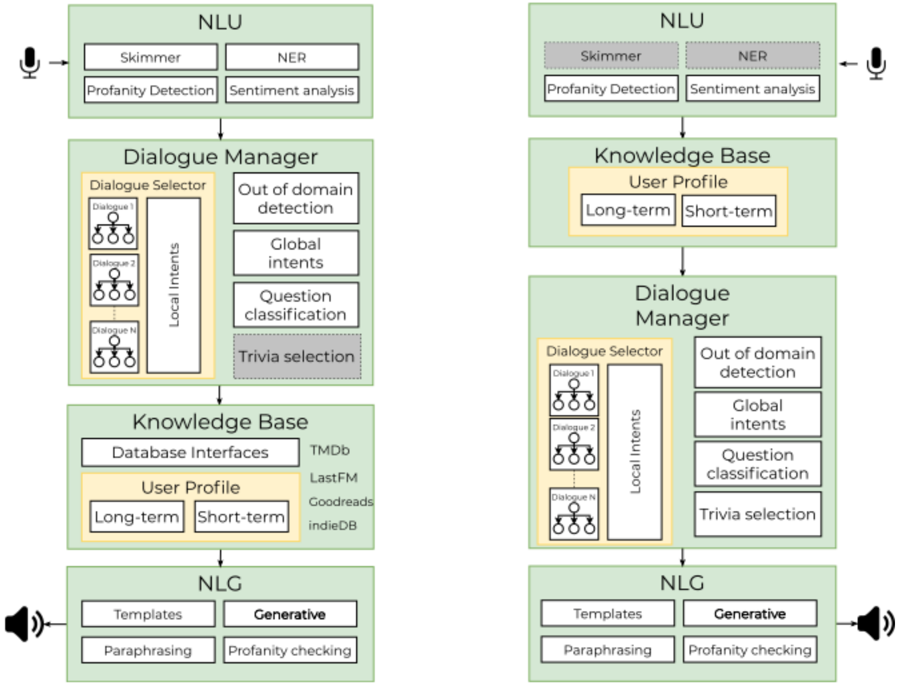
http://alexaprize.s3.amazonaws.com/2017/technical-article/alquist.pdf
http://dex-microsites-prod.s3.amazonaws.com/alexaprize/2018/papers/Alquist.pdf
https://chatbotsmagazine.com/13-lessons-we-have-to-learn-from-amazon-alexa-prize-965628e38ccb
https://towardsdatascience.com/11-more-lessons-we-have-to-learn-from-alexa-prize-94fe14b8986f

Alquist (Czech Technical University, 19/20 3rd, 20/21 1st)

(Pichl et al., 2020) https://arxiv.org/abs/2011.03261
(Konrád et al., 2021) https://arxiv.org/abs/2109.07968

- Knowledge graph: Wikidata + User + Bot model
  - RDF triples, partially delexicalized
  - allows building user profile + referencing it
- NLU – BERT-based segmenting (multiple intents)
  - produce responses to all, then select
- DM/NLG – response based on “adjacency pairs”
  - predefined input-response pairs/sub-graphs
  - transition depends on KG search
  - adding prompts (questions, fun facts etc.)
- Out-of-domain: detection & DialoGPT response
  - DialogRPT reranker
- Exploration vs. exploitation
  - first get to know user, then use this information





(a) Exploration flow (b) Exploitation flow

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

Summary

- chatbots = **non-task oriented** systems
  - targets: **conversation length** & **user engagement**
  - impersonating a human – Turing test
- approaches:
  - **rule-based** – keyword spotting, scripting
  - **retrieval** – copy & paste from large databases
  - **generative** – seq2seq/transformer trained on corpora of dialogues
    - too many possible responses don't go well with MLE → safe, short, dull
    - many extensions: personality, coherence, diversity, retrieval-augmented… still not ideal
  - **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

Contact us:

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

Labs in 10 mins
5th assignment

Get these slides here:

http://ufal.cz/npfl099

References/Inspiration/Further:

- Mainly individual papers referenced directly on slides
- 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)