NFPL099 Statistical Dialogue Systems

10. Chatbots/Open-Domain Systems

http://ufal.cz/npfl099

Ondřej Dušek, Simone Balloccu, Zdeněk Kasner, Mateusz Lango, Ondřej Plátek, Patrícia Schmidtova

5.12.2023
Chatbots / Chatterbots

- dialogue systems for **open-domain** dialogue – **chitchat**
  - i.e. “talk about anything”, though this definition is problematic
    - we don’t talk about anything with anyone, there’s a lack of shared context (common ground)
    - definitions aren’t unified across literature (may be more “social”)
- mostly **non-task-oriented** (though this changes)
  - main goal: keep the user entertained
  - standard evaluation: conversation length, user engagement
- (somewhat) different architecture
  - mostly simpler, integrated – like end-to-end DS (i.e. no separate NLU/DM/NLG)
  - it’s hard to have explicit NLU – no task to guide the meaning formalism
  - some of them don’t need a DB connection (but some use it)
- beware: *any dialogue system* is called a “chatbot” nowadays
  - this lecture: only non-task-oriented / open-domain 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** (1990-2019)
  - Turing test style, first topic-restricted 1995+ unrestricted
  - time-limited (currently 25 minutes for both conversations)
  - criticized as publicity stunt – hype but no real progress

• **Amazon Alexa Prize** (2017+, “Socialbot Grand Challenge”)
  - 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
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
  • AIML – standard for keyword spotting rules (e.g. Pandorabots platform)

• **Data-driven**
  • retrieval – remember a corpus & get replies from there
    • “nearest neighbour” approaches
    • corpus can contain past conversations with users
    • 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)
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:
  • use postprocessing, combine with rules (e.g. Cleverbot/Xiaoice bots)
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)

(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 = 👍儿 - 👎 votes

(Gao et al., 2020)
https://aclanthology.org/2020.emnlp-main.28/
Generative chatbots: RNNs

- Generic LSTM-based **Seq2seq** model
  - encode context / input, decode response
  - examples in the paper look great, but… (→)
    - the paper also reports some problems, but they’re much more prevalent than it looked

<table>
<thead>
<tr>
<th>U: who is skywalker ?</th>
<th>S: he is a hero .</th>
</tr>
</thead>
<tbody>
<tr>
<td>U: who is bill clinton ?</td>
<td>S: he ’s a billionaire .</td>
</tr>
<tr>
<td>U: what is the purpose of life ?</td>
<td>S: to serve the greater good .</td>
</tr>
<tr>
<td>U: what is moral ?</td>
<td>S: what empowered humanity , what intellectual the essence is .</td>
</tr>
</tbody>
</table>

Human: who is skywalker ?
Machine: he is a hero .
Human: who is bill clinton ?
Machine: he 's a billionaire .
Human: is sky blue or black ?
Machine: blue .
Human: does a cat have a tail ?
Machine: yes .
Human: does a cat have a wing ?
Machine: no .
Human: can a cat fly ?
Machine: no .
Human: how many legs does a cat have ?
Machine: four , i think .
Human: how many legs does a spider have ?
Machine: three , i think .
Human: how many legs does a centipede have ?
Machine: eight .
Human: what is the color of the sky ?
Machine: blue .
Human: what is the color of water ?
Machine: water .
Human: what is the color of blood ?
Machine: it is the same as a black eye .
Human: what is the usual color of a leaf ?
Machine: it is a green one .
Human: what is the color of a yellow car ?
Machine: yellow .
Human: how much is two plus two ?
Machine: four .
Human: how much is ten minus two ?
Machine: seventy - two .

(Vinyals & Le, 2015)
https://arxiv.org/abs/1506.05869
Seq2seq Chatbot Problems

• **generic/dull responses**
  - short, low-entropy, repetitive
  - problem: too many possible outputs given input
    - 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.

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 diversity & coherence: MMI, HRED

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

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

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

Input: what is your name?

<table>
<thead>
<tr>
<th>I don’t know.</th>
<th>-0.91</th>
</tr>
</thead>
<tbody>
<tr>
<td>I don’t know!</td>
<td>-0.92</td>
</tr>
<tr>
<td>I don’t know, sir.</td>
<td>-1.58</td>
</tr>
<tr>
<td>Oh, my god!</td>
<td>-0.97</td>
</tr>
<tr>
<td>My name is John.</td>
<td>-1.59</td>
</tr>
</tbody>
</table>

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

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

- joining **next turn generation & autoencoding**
  - LSTM VAE-like model, shared latent space
  - multi-task learning
  - shared decoder
  - additional “fusion loss” enforcing the same encoding for both tasks

- inference: adding a little noise to encodings
  - to produce different outputs

(Gao et al., 2019)
http://arxiv.org/abs/1902.11205
Improving coherence: Additional objectives

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

- **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

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

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

- **Meena**
  - “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**
  - 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

(Adiwardana et al., 2020)

(Zhang et al., 2020)
https://www.aclweb.org/anthology/2020.acl-demos.30

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

- 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

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

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

• LLM prompts often include “persona”
  • in their **system prompt / metaprompt / system message**
  • special prompt added before the actual conversation starts
  • ChatGPT: *You are a helpful assistant*

• Can include more details
  • personality, limitations, capabilities
  • behavior “guardrails” (*Avoid harmful or unethical content.*)

• Different personalities influence LM behavior & performance
  • adding a role help, esp. interpersonal & not too intimate (*friend, colleague*)
  • choosing the best role is tricky

(Zheng et al., 2023)
http://arxiv.org/abs/2311.10054
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

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

---

(Pandey et al., 2018) [https://aclanthology.org/P18-1123/](https://aclanthology.org/P18-1123/)
(Weston et al., 2018) [https://aclanthology.org/W18-5713/](https://aclanthology.org/W18-5713/)
(Xu et al., 2021) [http://arxiv.org/abs/2107.07567](http://arxiv.org/abs/2107.07567)
(Roller et al., 2021) [https://aclanthology.org/2021.eacl-main.24](https://aclanthology.org/2021.eacl-main.24)
Retrieval Transformer / Toolformer

- Retrieval as you generate
  - conditioned on the already generated tokens
  - allows to feed in relevant factual info

- RETRO
  - 2 nearest neighbor prefixes from DB
  - retrieved for each chunk = 4 tokens
  - retrieve, use in attention (via special layers)

- Toolformer
  - LM decodes special prefix + params for “tools” i.e. different API calls
  - finetuned on data with interleaved API calls
  - API calls sampled & filtered by loss reduction
  - QA, Wiki search, calc, calendar, MT

(Borgeaud et al., 2022) http://arxiv.org/abs/2112.04426
https://jalammar.github.io/illustrated-retrieval-transformer/

(Schick et al., 2023) http://arxiv.org/abs/2302.04761
http://arxiv.org/abs/2112.04426
Reasoning via external model

- COMET: pretrained commonsense model
  - inferring user/system emotion (xReact) + user desire (xWant) & system intent (xIntent)
  - user at runtime, system at training only (from ground truth data)
- system intent & emotion inferred at runtime
  - from user intent & emotion
  - using ChatGPT with specially crafted prompts
- ChatGPT / T5 conversation model
  - inferred intent & emotion inserted before reply
Training from feedback

• **LaMDA**: LM + retrieval + “calculator”
  - pretrained on dialogue
  - finetuned on annotated corrections of its own outputs
  - usage of retrieval & calculator annotated
  - generate multiple, filter (safety) & rerank
  - 2B/137B params versions

• **RLHF**: “standard” set by ChatGPT
  1) supervised finetuning
  2) evaluation/ranker model training
     • human annotations for model generation
  3) RL with proximal policy optimization
     • ranker model as reward
     • direct preference optimization: RL-free alternative


[https://openai.com/blog/chatgpt](https://openai.com/blog/chatgpt), [https://www.youtube.com/watch?v=0A8ljAkdfTg](https://www.youtube.com/watch?v=0A8ljAkdfTg)

[https://huggingface.co/blog/rlhf](https://huggingface.co/blog/rlhf)
Training from feedback

**JUICER**
- Get sparse user feedback & corrections
- Label rest of the data
- Train corrector LM
- Convert bad → good replies
- Retrain model on converted replies

**SYNDICOM**
- GPT-3 generated dialogues by rephrasing templates
- GPT-3 error injection
- Human feedback → corrector LM


(Richardson et al., 2023) [https://aclanthology.org/2023.sigdial-1.27](https://aclanthology.org/2023.sigdial-1.27)
• Production SotA*: combing 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 for everything else
• 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

*if you want to retain some control
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

<table>
<thead>
<tr>
<th>Turn</th>
<th>User</th>
<th>Sounding Board</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Let’s chat.</td>
<td>Hi! This is an Alexa Prize socialbot. How’s your day?</td>
</tr>
<tr>
<td>2</td>
<td>I’m fine.</td>
<td>Glad to hear it! We could talk about robots, Batman, or Superman. Say “next”, to chat about other things.</td>
</tr>
<tr>
<td>3</td>
<td>Superman.</td>
<td>It looks like you wanna chat about Superman. I was high 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?</td>
</tr>
<tr>
<td>4</td>
<td>I guess so.</td>
<td>Did you know that Henry Cavill almost missed the call for the role of Superman cause he was playing &quot;World of Warcraft&quot;</td>
</tr>
<tr>
<td>5</td>
<td>Really, I didn’t know that.</td>
<td>Weird, right? Speaking of Superman, did you see Superman? It was released in 1997. It’s a comedy.</td>
</tr>
<tr>
<td>6</td>
<td>Yes, it was hilarious.</td>
<td>I’m glad you feel this is hilarious. Which part do you like best about this movie?</td>
</tr>
<tr>
<td>7</td>
<td>The part when he met Lewis</td>
<td>Interesting. Wescott and Reilly 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?</td>
</tr>
</tbody>
</table>

http://arxiv.org/abs/1804.10202
https://sounding-board.github.io/
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
Alquist (Czech Technical University, 2017&2018 2nd)

- 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

http://alquistai.com/
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
• 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
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 well
• 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 \(\rightarrow\) safe, short, dull
    • many extensions: personality, coherence, diversity, retrieval-augmented, RLHF
  • **hybrid** – combining all of the above

• 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@ufal.mff.cuni.cz
Skype/Meet/Zoom (by agreement)

Get these slides here:
http://ufal.cz/npfl099

References/Inspiration/Further:
• Mainly individual papers referenced directly on slides
• Wikipedia: AIML Chatbot Cleverbot ELIZA Jabberwacky Loebner_Prize Mitsuku PARRY Turing_test XiaoIce Zo_(bot)