NFPL099 Statistical Dialogue Systems

10. Chatbots (non-task-oriented)

http://ufal.cz/npfl099

Ondřej Dušek, Vojtěch Hudeček & Zdeněk Kasner

5.12.2022
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 \( \rightarrow \) 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

- **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”

References:
- Pandorabots: https://home.pandorabots.com/home.html
- Cleverbot: https://www.cleverbot.com/
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/Xiaolce)
    - 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)
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
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 (Xiaolce, 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 / 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 = 👍 - 👎 votes
Generative chatbots: RNNs

- 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.*

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

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!!*
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

(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?)

(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.)

(Wei et al., 2019)  
https://arxiv.org/abs/1712.02250

(Jiang & De Rijke, 2018)  
https://www.aclweb.org/anthology/W18-5712

NPFL099 L10 2022
• **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

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

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

\[
\text{MI} = \log \frac{P(R, C)}{P(R)P(C)}
\]
• 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

---

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

- **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** (sorta SotA now)
  - 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

(Zhang et al., 2020) https://www.aclweb.org/anthology/2020.acl-demos.30
(Roller et al., 2021) https://aclanthology.org/2021.eacl-main.24/
Chat-specific PLMs (2)

- **LaMDA**: LM + retrieval + “calculator”  
  - 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


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

• 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)
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
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]


User: Interesting

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

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

User: Kendrick Lamar

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:** could we talk about something else

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

**User:** Let's talk about Tom Hardy

**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

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

**User:** What do you know about pizza?

**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

<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 &quot;next&quot;, to chat about other things.</td>
</tr>
<tr>
<td>3</td>
<td>Superman.</td>
<td>It looks like you want to 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?</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 because 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 1978. It’s a comedy.</td>
</tr>
<tr>
<td>6</td>
<td>Yes, it was hilarious.</td>
<td>I’m glad you find this hilarious. Which part do you like best about this movie?</td>
</tr>
<tr>
<td>7</td>
<td>The part when he met Lewis Utah</td>
<td>Interesting. McQuarrie and Refn 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/
**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…

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

- 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

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

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)