NPFL123 Dialogue Systems

12. Chatbots
(non-task-oriented dialogue)

https://ufal.cz/npfl123

Ondřej Dušek, Vojtěch Hudeček, Tomáš Nekvinda
& Jan Cuřín, Petr Fousek

9. 5. 2021
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: anything 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”
Chatbot history

• natural communication – important part of general AI
  • concerned people even before modern computers (cf. Turing)
• 1\textsuperscript{st} chatbot: \textbf{Eliza} (1966)
  • rule-based, simulates a therapist
• \textbf{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”
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

normalization is typically applied during preprocessing

```
<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>
```

```
<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>
```

0/more words

multiple options chosen at random

0/more words

(higher priority match)

1/more words

2 categories reduced via srai to the same pattern

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 (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**
      • 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)
• 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

---

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

---

**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:  
  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… (→)

(Sordoni et al., 2015)  
https://www.aclweb.org/anthology/N15-1020

NPFL123 L12 2022
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)

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

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

• **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

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

\[ C = \text{context} \]
\[ R = \text{reply} \]
Improving upon Seq2seq

- **Hierarchical seq2seq** for longer context
  - HRED (Hierarchical Recurrent Encoder-Decoder)
  - use a 2nd, turn-level LSTM encoder, word-level LSTM hidden state as input

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

- **DialoGPT** – GPT-2 finetuned on Reddit (147M dialogues) (Zhang et al., 2020)
  - no hierarchy, just decoder, whole chat as a long text – next-word prediction
  - works better than seq2seq-based ones

- **Meena** (Adiwardana et al., 2020)
  - Slightly modified Transformer
  - encoder-decoder, huge, trained on 867M dialogues (next-word prediction)
  - rule-based postprocessing

- **BlenderBot** (sorta SotA now) (Roller et al., 2021)
  - 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
Hybrid / Ensemble Chatbots (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
  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”

NLU pipeline
- NP extraction
- NER/entity linking
- intents
- sentiment
- topic detection
- ellipsis
- coreference

Bots ensemble
- Coherence bot
- Eliza bot
- Persona bot
- Ontologies bot
- News bot
- Wiki bot
- Reddit bot
- Fun facts bot
- Weather
- Abuse mitigation
- Clarification
- Evi QA bot

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: 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

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.
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

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-alexaprize-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: 
- greeting
  - Hi! This is an Alexa Prize socialbot. How’s your day?
- negotiation
  - I’m fine. Glad to hear it! We could talk about robots, Batman, or Superman. Say “next” to chat about other things.
- thoughts
  - 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
  - I guess so. Did you know that Henry Cavill almost missed the call for the role of Superman because he was playing "World of Warcraft"
- movies
  - Really, I didn’t know that. Weird, right? Speaking of Superman, did you see Superman? It was released in 1978. It’s a comedy.
  - Yes, it was hilarious. I’m glad you feel this is hilarious. Which part do you like best about this movie?
- movies
  - I think the part when he met Lewis in Utah.
  - The part when he met Lewis Utah. Interesting. McQuilin 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://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…

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
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
  • 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

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:
• Wikipedia: AIML Chatbot Cleverbot ELIZA Jabberwacky Loebner_Prize Mitsuki PARRY Turing_test Xiaoice Zo_(bot)

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