NFPL099 Statistical Dialogue Systems

10. Chatbots (non-task-oriented)

http://ufal.cz/npfl099

Ondřej Dušek & Vojtěch Hudeček

8.12.2020
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

• **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)

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

```
<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>
```

(normalization is typically applied during preprocessing)

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

- Early generative approach – convert to **phrase-based MT**
  - **align phrases** in inputs & responses
    - build **phrase table**
  - **build responses by covering input**
    - cover each phrase in input with likely response phrase
    - order & select phrases according to a language model (n-gram LM)

- it’s harder than MT
  - possible responses are much more variable than possible translations
    - using more sensitive phrase building (occurrence correlation)
    - discouraging input repetition (← would be vanilla MT behaviour)

- works, but…
  - fluency is not ideal
  - limited context

---

**Source** | **Target**
--- | ---
r1 (retweet) | thanks for the
potter | harry
ice | cream
how are you | you ?
good | morning
chuck | norris
watching | movie
i miss | miss you too
are you | i ’m
my birthday | happy birthday
wish me luck | good luck
how was | it was
miss you | i miss
swine | flu
i love you | love you too
how are you ? | are you ?
did you | i did
jackson | michael
how are you | i ’m good
michael | mj

---

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

U: riding to the arc on my bike. Hopefully I won’t melt into the pavement.
S: good luck on your bike ride arc

U: I really hate watching movies in the living room.
S: me too , what movies are you doing in room ?
Using RNN LMs (still MT-inspired)

- Non-LSTM solution:
  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

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

U: who is skywalker?
S: he is a hero.

U: who is bill clinton?
S: he ’s a billionaire.

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

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

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!!

12
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

(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

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

- **Persona embeddings**
  - 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

- **Persona copy-net**
  - using a hierarchical pointer-generator net
  - context includes short personal bio

---

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

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

- **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

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

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

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.
Diversity

- joining seq2seq (next turn generation) & autoencoding
  - multi-task learning
  - shared decoder
  - additional “fusion loss” enforcing the same encoding for both tasks
- Inference: adding a little noise to produce different outputs

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

- **CVAE with a coherence measure**
  - GLoVe average embedding cosine similarity
  - training data: coherence of target response known
    - also good for data filtering
  - inference: set coherence very high: 0.95/1.0

- **GAN-style for consistent order**
  - detect if three turns are consecutive or not
    - given 1 ordered, 1 misordered triple from previous dialogue history
  - use in dialogue generation learning: good replies are easy to check for order
    - see if misorder is easy to detect with a generated reply
  - GAN: train generator to produce good replies (where misorder is easy to detect) train detector to detect misorder in real sentences, not in generated

---

*Coherence*

- (Xu et al., 2018)
- (Wu et al., 2019)
Coherence: Additional Objectives

- **Transformer-based architectures**
- **Denoising** (autoencoder): additional decoders
  - recovering shuffled word order
  - recovering masked words
  - recovering masked utterance (mid-dialogue)
  - recovering utterance order (GRU decoding order)
- **Unlikelihood** – demoting unlikely tokens
  - penalize set of tokens selected at each time step
    \[
    \mathcal{L}_{UL}^{(i)}(p_\theta, C_{1:T}, x, y) = - \sum_{t=1}^{T} \sum_{y_c \in C_t} \beta(y_c) \log(1 - p_\theta(y_c|x, y_{<t}))
    \]
  - summing over steps (=all words of output)
  - added to regular MLE loss
  - penalized: repeating n-grams, too much high-freq. vocab, contradictions

(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
Pretrained Language Models

• **TransferTransfo** – GPT-like
  • pretrained on books, finetuned on PersonaChat
  • person embeddings (+training with swapped)
  • next-word prediction, next-utterance classification

• **DialogPT** – just GPT-2 finetuned on Reddit
  • 147M dialogues
  • no hierarchy, whole chat as a long text – next-word prediction
  • (optional) MMI reranking
  • 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

(Wolf et al., 2018)
https://arxiv.org/abs/1901.08149

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

(Adiwardana et al., 2020)
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
• 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: Hmm... about Kendrick Lamar, so John, I read a story with this 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.
• 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

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, 2019/20 3rd)

- Knowledge graph: Wikidata + User + Bot model
  - RDF triples, partially delexicalized
  - allows building user profile + referencing it

- NLU – segmenting (multiple intents)
  - BERT-based segmenting
  - actions per segment = intent-properties-entities
    - produce responses to all, then select

- DM/NLG – response based on “adjacency pairs”
  - predefined input-response pairs/sub-graphs
  - transition depends on KG search
  - delexicalized – lexicalized subsequently
  - adding prompts (questions, fun facts etc.)

(Pichl et al., 2020)
Emora (Emory Uni, 2019/20 1st) & Chirpy Cardinal (Stanford, 2019/20 2nd)

- **Emora**  
  - NLU – prominent topic & sentiment classifier
    - stress on emotion, personal experience
  - hierarchical ontology of topics & sub-topics
    - use higher level if more specific is not available
  - state machine manager
    - transitions similar to Alquist

- **Chirpy Cardinal**  
  - architecture similar to Alana
  - multiple response generators
    - treelet-based handcrafted dialogues
  - GPT-2-based chatbot
  - adding prompts to replies, same as Alquist
  - specific “navigational” intents
    - meta-dialogue: discussing what topic to talk about
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… 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}@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)

No labs today
Project questions?