Dialogue Systems
NPFL123 Dialogové systémy

3. Data & Evaluation

Ondřej Dušek & Vojtěch Hudeček
http://ufal.cz(npfl099
17. 10. 2019
Before you build a dialogue system

Two significant questions, regardless of system architecture:

1) **What data** to base it on?
   - even if you handcraft, you need data
     - people behave differently
     - you can’t enumerate all possible inputs off the top of your head
   - ASR can’t be handcrafted – always needs data

2) **How to evaluate** it?
   - is my system actually helpful?
   - did recent changes improve/worsen it?
   - actually the same problem as data
     - you can’t think of all possible ways to talk to your system
Dialogue Data Collection

Typical options:

• **in-house collection** using experts (or students)
  - safe, high-quality, but very expensive & time-consuming
  - scripting whole dialogues / Wizard-of-Oz

• **web crawling**
  - fast & cheap, but typically not real dialogues
    - may not be fit for purpose
  - potentially unsafe (offensive stuff)
  - need to be careful about the licensing

• **crowdsourcing**
  - compromise: employing (untrained) people over the web
Wizard-of-Oz (WoZ)

• for in-house data collection
  • also: to prototype/evaluate a system before implementing it!

• users believe they’re talking to a system
  • different behaviour than when talking to a human
  • typically simpler

• system in fact controlled by a human “wizard” (=you)
  • typically selecting options (free typing too slow)
Crowdsourcing

• hire people over the web
  • create a webpage with your task
  • data collection / evaluation
  • no need for people to come to your lab
  • faster, larger scale, cheaper

• platforms / marketplaces
  • Amazon Mechanical Turk
  • CrowdFlower/FigureEight

• problems
  • can’t be used in some situations (physical robots, high quality audio…)
  • crowd workers tend to game the system – noise/lower quality data
  • a lot of English speakers, but forget about e.g. Czechs
Available Dialogue Datasets

• There’s a number of research datasets available
  • (see labs assignment 1)
  • typically built as part of various research projects
  • license: some of them research-only, some completely free

• Various types:
  • human-human, human-machine, Wizard-of-Oz
  • task-oriented or non-task-oriented
  • text-based, multimodal, (audio + text – rare)

• Common drawbacks:
  • domain choice is rather limited
    • but it’s getting better
    • non-task-oriented are still not ideal (mostly discussion forums, subtitles)
  • size is very often not enough – big AI firms have much more
    • this is also improving
  • vast majority is English only
Dataset Splits

• Never evaluate on data you used for training
  • memorizing training data would give you 100% accuracy
  • you want to know how well your model works on new, unseen data

• Typical dataset split:
  • training set = to train your model
  • development/validation set = for evaluation during system development
    • this influences your design decisions, model parameter settings, etc.
  • test/evaluation set = only use for final evaluation
  • need sufficient sizes for all portions

• Cross-validation – when data is scarce:
  • split data into 5/10 equal portions, run 5/10x & test on different part each time
Dialogue System Evaluation

• Depends on dialogue system type / specific component

• Types:
  • **extrinsic** = how the system/component works in its intended purpose
    • effect of the system on something outside itself, in the real world (i.e. user)
  • **intrinsic** = checks properties of systems/components in isolation, self-contained
  • **subjective** = asking users’ opinions, e.g. questionnaires (~manual/human)
    • should be more people, so overall not so subjective 😊
  • **objective** = measuring properties directly from data (~automatic)
    • might or might not correlate with users’ perception

• Evaluation discussed here is mostly **quantitative**
  • i.e. measuring & processing numeric values
  • (**qualitative** ~ e.g. in-depth interviews, more used in social science)
Getting the Subjects (for human evaluation)

• Can’t do without people
  • **simulated user** = another (simple) dialogue system
    • can help & give guidance sometimes, but it’s not the real thing – more for intrinsic

• **In-house** = ask people to come to your lab
  • students, friends/colleagues, hired people
  • expensive, time-consuming, doesn’t scale (difficult to get subjects)

• **Crowdsourcing** = hire people over the web
  • much cheaper, faster, scales (unless you want e.g. Czech)
  • not real users – mainly want to get their reward

• **Real users** = deploy your system and wait
  • best, but needs time & advertising & motivation
  • you can’t ask too many questions
Intrinsic – NLU

• Slot Precision & Recall & F-measure (F1)

Precision

\[ P = \frac{\text{#correct slots}}{\text{#detected slots}} \]

how much of the identified stuff is identified correctly

Recall

\[ R = \frac{\text{#correct slots}}{\text{#true slots}} \]

how much of the true stuff is identified at all

F-measure

\[ F = \frac{2PR}{P + R} \]

harmonic mean – you want both \( P \) and \( R \) to be high (if one of them is low, the mean is low)

true: inform(name=Golden Dragon, food=Chinese)
NLU: inform(name=Golden Dragon, food=Czech, price=high)

\[ P = \frac{1}{3} \]
\[ R = \frac{1}{2} \]
\[ F = 0.2 \]
Intrinsic – NLU

• **Accuracy** (% correct) used for intent/act type
  • intent detection is multi-class classification (1 utterance → 1 intent)
• alternatively also **exact matches** on the whole semantic structure
  • easier, but ignores partial matches
• Assumes one true answer, which might not be accurate
  • there’s ambiguity in some user inputs
  • it’s still used since it’s too hard to account for multiple correct options

• NLU on ASR outputs vs. human transcriptions
  • both options make sense, but measure different things!
  • intrinsic NLU errors vs. robustness to ASR noise
Intrinsic – Dialogue Manager

• Objective measures (task success rate, duration) can be measured with a user simulator
  • works on dialogue act level
  • responds to system actions

• Simulator implementation
  • handcrafted (rules + a bit of randomness)
  • n-gram models over DA/dialogue turns + sampling from distribution
  • agenda-based (goal: constraints, agenda: stack of pending DAs)
  • reinforcement learning policy

• Problem: simulator implementation cost
  • the simulator is basically another dialogue system
Intrinsic – NLG / Extrinsic

• No single correct answer here
  • many ways to say the same thing

• Word-overlap with reference text(s): **BLEU score**

  \[ \text{BLEU} = BP \cdot \exp \left( \sum_{n=1}^{4} \frac{1}{4} \log (p_n) \right) \]

  - **Brevity penalty** (1 if output longer than reference, goes to 0 if too short)
  - **n-gram precision**: 
    \[ p_n = \frac{\sum_u \# \text{ matching n–grams in } u}{\sum_u \# \text{ n–grams in } u} \]

• **n-gram** = span of adjacent \( n \) tokens
  • 1-gram (one word) = unigram, 2-gram (2 words) = bigram, 3-gram = trigram
**BLEU**

Example:

output: *The Richmond’s address is 615 Balboa Street. The phone number is 4153798988.*

*ref1:* The number for Richmond is 4153798988, the address is 615 Balboa.

*ref2:* The Richmond is located at 615 Balboa Street and their number is 4153798988.

matching unigrams: the (2x), Richmond, address, is (2x), 615, Balboa, . (only 1x!), number, 4153798988

\[ p_1 = \frac{11}{15} \]

matching bigrams: The Richmond, address is, is 615, 615 Balboa, Balboa Street, number is, is 4153798988, 4153798988.

\[ p_2 = \frac{8}{14} \]

\[ p_3 = \frac{5}{13}, \quad p_4 = \frac{2}{12}, \quad \text{BP} = 1, \quad \text{BLEU} = 0.4048 \]

• **BLEU is not very reliable** (people still use it anyway)
  • correlation with humans is questionable
  • never use for a single sentence, only over whole datasets
Intrinsic – NLG / Extrinsic

Alternatives (not much):

• Other word-overlap metrics (NIST, METEOR, ROUGE …)
  • there are many, more complex, but frankly not much better

• **Slot error rate** – only for delexicalized NLG in task-oriented systems
  • delexicalized → generates placeholders for slot values
  • compare placeholders with slots in the input DA – \[ \frac{\text{#missed+added+wrong\_value slots}}{\text{#total slots}} \]

• **Diversity** – mainly for non-task-oriented
  • can our system produce different replies? (if it can’t, it’s boring)

\[ D = \frac{\#\text{distinct}_x}{\#\text{total}_x}, \text{ where } x = \text{unigrams, bigrams, sentences} \]
Intrinsic NLG / Extrinsic

Entropy / perplexity

\[ H(p) = - \sum_x p(x) \log p(x), \quad 2^H(p) \]

- intrinsic for **language modelling** / word prediction
  - fitting the test set / reference outputs: lower is better
  - actually cross-entropy

- extrinsic – model output **diversity** (Shannon entropy)
  - looking at model outputs per se, no references
  - higher is better, more diverse
  - Variant: **n-gram conditional entropy**
    - entropy with known previous context

\[ -\frac{1}{N} \sum_{i=1}^{N} \log q(x_i) \]
NLG Supervised Quality Estimation

• Training a supervised model to...
  • check if an NLG system output is good or not (give rating)
    • just given the output + corresponding NLG input (dialogue act)
    • without using reference texts
    • can be used at runtime: should we trigger a fallback?
• check which output is the best out of multiple
  • selecting from n-best list

MR: inform_only_match(name='The Cricketers', area='pacific heights')
NLG output: the only match i have for you is the hotel drisco in the pacific heights area.

Rating: 4 (on a 1-6 scale)

MR: inform(name='The Crickets', eatType='coffee shop', rating=high, familyFriendly=yes, near='Café Sicilia')
NLG 1: The Cricketers is a children friendly coffee shop near Café Sicilia with a high customer rating.
NLG 2: The Cricketers can be found near the Café Sicilia. Customers give this coffee shop a high rating. It's family friendly.

Rating: 4 (on a 1-6 scale)

Rank: better ← worse
NLG QE Model

• Encoders for input DA + NLG output(s) → fully connected → linear

• Ranking: use 2 identical networks for 2 outputs
  • can learn both things jointly

• More reliable than BLEU
  • but still quite bad absolute (noise in the ratings?)
Extrinsic – Objective

• **Analyzing the logs** of people/testers interacting with the system

Metrics:

• **Task success** (task-oriented): did the user get what they wanted?
  • testers with agenda → check if they found what they were supposed to
    • [warning] sometimes people go off script
    • basic check: did we provide any information at all? (any bus/restaurant)

• **Duration**: number of turns
  • task oriented: fewer is better, non-task-oriented: more is better

• Other (not so standard):
  • % returning users
  • % turns with null semantics (task-oriented)
  • % swearing / thanking
Extrinsic – Subjective (Questionnaires)

• **Questionnaires** for users/testers
  • based on what information you need (overall satisfaction, individual components)

• Question types
  • **Open-ended** – qualitative
  • **Yes/No** questions
  • **Likert scales** – agree … disagree (typically 3-7 points)
    • with a middle point (odd number) or forced choice (even number)
  • **“Continuous” scales** – e.g. 0-100 (or no numbers shown)

• Question guidelines:
  • easy to understand
  • not too many
  • neutral: not favouring/suggesting any of the replies
Question Examples

- **Success rate (task-oriented):** Did you get all the information you wanted?
  - typically different from objective measures!

- **Future use:** Would you use the system again?

- **Likeability/engagement:** Did you enjoy the conversation?

- **ASR/NLU:** Do you think the system understood you well?

- **NLG:** Were the system replies fluent/well-phrased?

- **TTS:** Was the system’s speech natural?

<table>
<thead>
<tr>
<th>System</th>
<th># calls</th>
<th>Subjective Success Rate</th>
<th>Objective Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDC</td>
<td>627</td>
<td>82.30% (±2.99)</td>
<td>62.36% (±3.81)</td>
</tr>
<tr>
<td>NBC</td>
<td>573</td>
<td>84.47% (±2.97)</td>
<td>63.53% (±3.95)</td>
</tr>
<tr>
<td>NAC</td>
<td>588</td>
<td>89.63% (±2.46)</td>
<td>66.84% (±3.70)</td>
</tr>
<tr>
<td>NABC</td>
<td>566</td>
<td>90.28% (±2.44)</td>
<td>65.55% (±3.91)</td>
</tr>
</tbody>
</table>

Jurčiček et al., Comp. Speech & Language 2012
Question Types

• Aiming at rater consistency (multiple people rating the same)
  • high intraclass correlation coefficient

• **Likert** vs. **continuous**
  • Continuous scales seem to increase consistency

• alternatives: mainly for individual system outputs
  • too hard to do for whole dialogue
  • also better than Likert

• **Relative ranking** / Best-worst scaling
  • sort outputs from best to worst
  • variants: ties allowed / not

• **Magnitude estimation**
  • Show reference, with a value (e.g. 100)
  • rank-based: ask to assign values to multiple outputs
    • indirect ranking

(Santhanam & Shaikh, 2019)
Retrieval metrics

• For retrieval/ranking systems

• Recall: $R_N@k$
  • assuming N candidates, 1 relevant response
  • % of time the relevant one is among top-k rated
  • e.g. $R_{100}@1$ – only the 1st out of 100 candidates

• $R_N@1$ given context = next utterance classification (NUC)

• precision possible in theory, but not used very much
  • “% of top-k rated that are relevant”
  • actually $P_N@1 = R_N@1$, assuming 1 relevant response
  • $R_N@k$ grows with higher $k$, $P_N@k \to 0$ with higher $k$
  • not many datasets have multiple outputs tagged as relevant

(Henderson, 2019)
https://www.aclweb.org/anthology/P19-1536
Turn-level Quality Estimation

Interaction Quality

• turns annotated by experts (Likert 1-5)
• trained model (SVM/RNN)
  • very low-level features
  • mostly ASR-related
  • multi-class classification
• result is domain-independent
  • trained on a very small corpus (~200 dialogues)
  • same model applicable to different datasets
• can be used in a RL reward signal
  • works better than task success

(Schmitt & Ultes, 2015; Ultes et al., 2017; Ultes, 2019)
https://doi.org/10.1016/j.specom.2015.06.003
https://doi.org/10.21437/Interspeech.2017-1032
https://aclweb.org/anthology/W19-5902/

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASRRecognitionStatus</td>
<td>ASR status: success, no match, no input</td>
</tr>
<tr>
<td>ASRCertainty</td>
<td>confidence of top ASR results</td>
</tr>
<tr>
<td>RePrompt?</td>
<td>is the system question the same as in the previous turn?</td>
</tr>
<tr>
<td>ActivityType</td>
<td>general type of system action:</td>
</tr>
<tr>
<td>Confirmation?</td>
<td>is system action confirm?</td>
</tr>
<tr>
<td>MeanASRConfidence</td>
<td>mean ASR confidence if ASR is success</td>
</tr>
<tr>
<td>#Exchanges</td>
<td>number of exchanges (turns)</td>
</tr>
<tr>
<td>#ASRSuccess</td>
<td>count of ASR status is success</td>
</tr>
<tr>
<td>%ASRSuccess</td>
<td>rate of ASR status is success</td>
</tr>
<tr>
<td>#ASRRejections</td>
<td>count of ASR status is reject</td>
</tr>
<tr>
<td>%ASRRejections</td>
<td>rate of ASR status is reject</td>
</tr>
<tr>
<td>{Mean}ASRConfidence</td>
<td>mean ASR confidence if ASR is success</td>
</tr>
<tr>
<td>{#}ASRSuccess</td>
<td>count of ASR is success</td>
</tr>
<tr>
<td>{#}ASRRejections</td>
<td>count of ASR status is reject</td>
</tr>
<tr>
<td>{#}RePrompts</td>
<td>count of times RePrompt? is true</td>
</tr>
<tr>
<td>{#}SystemQuestions</td>
<td>count of ActivityType is question</td>
</tr>
</tbody>
</table>

“reject” = ASR output doesn’t match in-domain LM
• BLEU problem for dialogue: multiple answers are OK
  • but most dialogue datasets only have 1 reference

• ΔBLEU: “discriminative” BLEU
  • get multiple references
  • have them rated (~crowdsourcing)
  • - appropriateness ∈ [−1,1]
  • weigh each n-gram match by highest-scoring reference in which it is found
    • this highest score can be negative → negative contribution to ΔBLEU
    • identical to multi-ref BLEU if all weights = 1
  • better correlation with humans

(Galley et al, 2015)
https://arxiv.org/abs/1506.06863
**ΔBLEU test set creation**

- Context-message-response triples
  - context: only 1 preceding message, ignoring the rest (sparse data)

1. Get messages with high-quality responses
2. Use IR to get alternative responses (IR on messages in training set)
3. Have other responses rated (don’t discard low-rated!)

- Rating: crowdsourced, 1-5 Likert-scale (5 raters average scaled to $[-1,1]$)

<table>
<thead>
<tr>
<th>Context $c$</th>
<th>Message $m$</th>
<th>Response $r$</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>i was about to text you and my two cousins got excited cause they thought you were &quot;rihanna&quot;</td>
<td>aww, i can imagine their disappointment</td>
<td>they were very disappointed!!!</td>
<td>0.6</td>
</tr>
<tr>
<td>yes, my ex-boyfriend, killed my cat. like i say, it was the start of a bad time...</td>
<td>i can imagine?</td>
<td>yes. luckily, the whole thing feels very much of the past now.</td>
<td>0.8</td>
</tr>
<tr>
<td>it's good.. for some reason i can't name stand out tracks but i've been playing it since it dropped</td>
<td>i can imagine, banks doesn't disappoint</td>
<td>na this is anything but a disappointment.</td>
<td>0.6</td>
</tr>
<tr>
<td>at my lil cousins dancing to “dance for you”. these kids are a mess.</td>
<td>Imnzo i can imagine.</td>
<td>they were belly rolling, hilarious.</td>
<td>0.4</td>
</tr>
<tr>
<td>what's sick about it?? do you know how long it is?? no so how is it sick?</td>
<td>i can imagine it</td>
<td>your imagination is wrong, very wrong at that.</td>
<td>-0.1</td>
</tr>
<tr>
<td>it's snow oh yeah the snow is very beautiful</td>
<td>yeah i can imagine</td>
<td>the weather in russia is very cool.</td>
<td>-0.7</td>
</tr>
</tbody>
</table>
ADEM (Automatic Dialogue Evaluation Model)

- Supervised model for dialogue response evaluation
- Inputs encoded by GRU RNNs:
  - context $c$: all prior turns (hierarchical)
  - reference gold-standard response $r$
  - system output response $\hat{r}$
- Output: dot product
  - with some (trained) transformation to keep it $\in [0,5]$
  - trained using human ratings
- Better correlated with human ratings than BLEU/ROUGE
- Robust to previously unseen models

(Lowe et al., 2017)
https://www.aclweb.org/anthology/P17-1103/
ADEM training data

- Twitter dataset
- Responses from 4 different models:
  - TF-IDF retrieval chatbot
  - neural retrieval chatbot
  - generative chatbot
  - humans (crowdsourced original alternative replies, not seeing references)
- Crowdsourced Likert scale (1-5) ratings
  - raters with low agreement removed
  - only measured overall score
    - other (topicality, informativeness…): low agreement / high correlation with overall

---

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># Examples</td>
<td>4104</td>
</tr>
<tr>
<td># Contexts</td>
<td>1026</td>
</tr>
<tr>
<td># Training examples</td>
<td>2,872</td>
</tr>
<tr>
<td># Validation examples</td>
<td>616</td>
</tr>
<tr>
<td># Test examples</td>
<td>616</td>
</tr>
<tr>
<td>( \kappa ) score (inter-annotator correlation)</td>
<td>0.63</td>
</tr>
</tbody>
</table>
Adversarial Evaluation

• bidi-LSTM encoder + attention → sigmoid classification layer
  • is the dialogue (preceding context + response) human-generated or not?
  • context limited – 1-2 utterances

• trained on 3 concatenated datasets (movies, phone transcripts)
  • negative examples: randomly sampled

• intrinsic evaluation: both model & humans aren’t great
  • accuracy around 0.7, low inter-annotator agreement (~0.3)

• detecting seq2seq outputs vs. real – discriminator better than humans
  • humans totally random, discriminator accuracy ~0.6-0.7

• might be a problem with the dataset – movies are messy
Topic-based Evaluation

- automatic evaluation for chatbots
- based on a topic classifier
  - “attentional deep averaging networks”
    - using topic-specific saliency ∀ word ~ per-topic attentions
    - few fully connected layers + final classification
  - given a turn, assign topic
    - two levels: coarse / fine (e.g. entertainment / movies )

- conversation topic breadth & depth
  - breadth: average number of distinct topics in each dialogue
  - depth: average length of sub-dialogue (consecutive turns on the same topic)

- correlates well with human overall dialogue ratings

(Guo et al, 2017)
http://arxiv.org/abs/1801.03622
Significance Testing

• Higher score is not enough to prove your model is better
  • Could it be just an accident?

• Need **significance tests** to actually prove it
  • Statistical tests, $H_0$ (**null hypothesis**) = “both models performed the same”
  • $H_0$ rejected with >95% confidence $\rightarrow$ pretty sure it’s not just an accident
  • more test data $\rightarrow$ more independent results $\rightarrow$ can get higher confidence (99+%) 

• Various tests with various sensitivity and pre-conditions
  • Student’s t-test – assumes normal distribution of values
  • Mann-Whitney $U$ test – any ordinal, same distribution
  • **Bootstrap resampling** – doesn’t assume anything
    1) randomly re-draw your test set (same size, some items 2x/more, some omitted)
    2) recompute scores on re-draw, repeat 1000x $\rightarrow$ obtain range of scores
    3) check if range overlap is less than 5% (1%...)
Summary

• You need data (corpus) to build your systems
  • various sources: human-human, human-machine, generated
  • various domains
  • size matters

• Evaluation needs to be done on a test set
  • intrinsic (component per se) / extrinsic (in application)
  • objective (measurements) / subjective (asking humans)
  • don’t forget to check significance

• Evaluation is non-trivial
  • there is no ideal metric – humans, BLEU, recall… all have their problems
  • you can try training a model for evaluation – might work better

• Next week: NLU
Thanks

Contact us:

odusek@ufal.mff.cuni.cz
hudecek@ufal.mff.cuni.cz
room 424 (but email me first)

Get the slides here:

http://ufal.cz/npfl099

References/Inspiration/Further:

• Oliver Lemon & Arash Eshghi’s slides (Heriot-Watt University): https://sites.google.com/site/olemon/conversational-agents
• Helen Hastie’s slides (Heriot-Watt University): http://letsdiscussnips2016.weebly.com/schedule.html

Labs today

14:00 SW1