2. Data & Evaluation

https://ufal.cz/npfl123

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

21. 2. 2022
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
Data: Corpus (pl. Corpora)

- **Corpus** = collection of (linguistic) data
  - assuming access for automatic processing
  - used to train your system / inform yourself / evaluate
  - also called **dataset**

- Some of them are released openly
  - usage rights depend on a **license**
    - e.g. Creative Commons
      - BY (attribution) – SA (share alike) – NC (non-commercial) – ND (no derivatives)

- Useful for linguistic research/description, too
Dialogue Corpora/Dataset Types

- **modality**: written / spoken / multimodal
- **data source**:
  - human-human conversations
    - real dialogues
    - scripted (e.g. movies)
  - human-machine (talking to a dialogue system)
  - automatically generated (“machine-machine”)
- **domain**
  - closed/constrained/limited domain
  - multi-domain (more closed domains)
  - open domain (any topic, chitchat)

Scenario:
Determine the type of aircraft used on a flight from Cleveland to Dallas that leaves before noon.

x0201lx.x: may i see all the flights from cleveland to , dallas
x0202lx.x:ro: can you show me the flights that leave before noon , only
x0203lx.x:ro: could you sh- please show me the types of aircraft used on these flights

INDY: Let’s get out of here!
MARION: Not without that piece you want!
INDY: It’s here?
MARION nods, kicks aside a burning chair. Another burning beam falls from the roof. Indy close to him protectively.
INDY: Forget it! I want you out of here. Now! He begins dragging her out.
MARION: pointing. There! She breaks away from him, darts back and picks the hot metal loose cloth of her blouse.
INDY: Let’s go!
MARION: (looking around) You burned down my place!
INDY: I owe you plenty!

(Walker et al., 2012) [https://www.aclweb.org/anthology/L12-1657/](https://www.aclweb.org/anthology/L12-1657/)

(Dahl et al., 1994) [https://www.aclweb.org/anthology/H94-1010/](https://www.aclweb.org/anthology/H94-1010/)

[https://tla.mpi.nl/tools/tla-tools/elan/](https://tla.mpi.nl/tools/tla-tools/elan/)
Typical options:

- **in-house collection** using experts (or students)
  - safe, high-quality, but very expensive & time-consuming
  - free talk / 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
  - Appen (formerly FigureEight/CrowdFlower)
  - Prolific

• **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

https://en.wikipedia.org/wiki/The_Turk

(Dušek & Jurčiček, 2016)
https://api.semanticscholar.org/CorpusID:15546788
more often than not, you’ll need more than just recordings

**annotation** = labels, description added to the collected data:
- **transcriptions** (textual representation of audio, for ASR&TTS)
- **semantic annotation** such as dialogue acts (NLU)
- **named entity** labelling (NLU)
- other linguistic annotation: part-of-speech, syntax – typically not in DSs

getting annotation
- similar task as getting the data itself
- DIY / hiring **experts**
- **crowdsourcing**
- (semi-)**automatic** annotation
  - use rules + manual fixes, annotate small dataset & use machine learning for the rest

*I want to fly from Boston to Dallas on Monday morning.*

request(from=Boston,to=Dallas,date=Mon,daytime=morn)
Inter-annotator Agreement (IAA)

• annotation is **inherently ambiguous**
  • people sometimes don’t even hear the same thing
  • let alone interpret the same semantics
• need to test if it’s reasonably **reliable**
  – **measuring IAA**
    • 2 or more people annotate/transcribe the same thing
    • need to account for agreement by chance
      • transcriptions – too many options (words) – no big deal
      • NER – just a few categories (e.g. 7) – may play a role
• typical measure: **Cohen’s Kappa** (0<\(\kappa\)<1)
  • for categorial annotation
  • 0.4 ~ fair, >0.7 ~ great

\[
\kappa = \frac{\text{agreement} - \text{chance}}{1 - \text{chance}}
\]
Corpus Size

• Size matters here
  • need enough examples for an accurate model
  • depends on what and how you’re modelling

• Speech – 10s-100s of hours

• NLU, DM, NLG
  • handcrafting – 10s-100s of dialogues may be OK to inform you
  • simple model/limited domain – 100s-1000s dialogues might be fine
  • open domain – sky’s the limit

• TTS – single person, several hours at least
Available Dialogue Datasets

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

• Drawbacks:
  • domain choice is rather limited
  • size is very often not enough – big AI firms have much more
  • vast majority is English only
  • few free datasets with audio
    • but there are non-dialogue ones (see http://www.openslr.org/)

[11]

https://mobile.twitter.com/yoavgo/status/1467633831465394181
Dialogue Datasets: Human-Machine

Good for NLU, state tracking
• not good for learning the whole dialogue (you’d just replicate the original system)

• Dialogue state tracking challenge (DSTC)
  • DSTC1 bus schedules
  • DSTC2/3 Cambridge restaurants

• Clinc – 10 domains + “out-of-scope”
  • crowdsourcing individual turns

• ATIS – flight booking
  • WoZ collection, very old (1990’s)

DSTC1 – Let’s go (Williams et al. 2013)
https://www.aclweb.org/anthology/W13-4065/
SYS: East Pittsburgh Bus Schedules. Say a bus route, like 28X, or say I’m not sure.
USR: 61A
SYS: Okay, 61A. To change, say go back. Where are you leaving from?
USR: Downtown
SYS: Okay, downtown. You can always say go back. And where are you going to?

DSTC2 – Restaurants (Henderson et al., 2014)
https://www.aclweb.org/anthology/W14-4337/

Clinc – 10 domains + “out-of-scope”
• crowdsourcing individual turns

Clinc (Larson et al., 2019)
https://www.aclweb.org/anthology/D19-1131

can i travel to france as far as safety goes = travel_alert
i need your help finding my lost phone = find_phone
read me cat trivia = fun_fact
what is the balance in my pnc account = balance

Show flights from Boston to New York today
O O O B-dept O B-arr1-arr B-date

ATIS
https://chsasank.github.io/spoken-language-understanding.html
Datasets: Human-Human Spoken

[backchannel] B.22 utt1: Uh-huh. /
[statement, non-opinion] A.23 utt1: I work off and on just temporarily and usually find friends to babys
[statement, non-opinion] A.23 utt2: {C but } I don’t envy anybody who’s in that <laughter> situation to
[backchannel] B.24 utt1: Yeah. /

Spontaneous:

• Switchboard, Callfriend
  • phone conversations, topic given (Switchboard) or unrestricted (Callfriend)

Constrained:

• Walking around, Verbmobil, DSTC4/5
  • tasks: navigation, meeting scheduling, tourist guide
  • many more (debates, games…)

Scripted:

• Open Subtitles, Cornell Movies
  • movie subtitles/scripts from the web
  • caveats: swearing, lost visual context

Callfriend: https://ca.talkbank.org/access/CallFriend/
Verbmobil: https://www.phonetik.uni-muenchen.de/Bas/BasVM1eng.html
DSTC4: http://www.colips.org/workshop/dstc4/

(Lison & Meena, 2016)

Blade Runner script
Datasets: Human-Human Written

• easier to get than spoken
  • caveats: specific language, may be offensive

• Spontaneous:
  • Twitter (closed API), Reddit (https://pushshift.io/)
    • large, very messy
  • DailyDialog
    • language learning sites – cleaner, smaller

• Constrained:
  • Ubuntu Dialogue (>1M dialogues, Ubuntu Chat)
  • MultiWOZ
    • 10k dialogues, with detailed annotation
      • restaurants, hotels, tourist attractions, trains…
      • other: Taskmaster, Schema-guided…

I need a train leaving on a Sunday and arriving by 21:15.
Okay, I can help you with that. Where will you be traveling?
From London Kings Cross to Cambridge.
TR1681 will arrive at 20:08, would that work for you?
Yes, that sounds good. Please book a ticket on TR1681 for 6 people for me.
The booking was successful, your reference number is EAWIQ7HX. Is there anything else I can help you with?
Dialogue Datasets: Machine Generated

• Fake, but still good for testing
  • can the model learn a dataset of this complexity?
• Can be generated in any size
• Facebook bAbI
  • various tasks, mainly inference
  • auto-generated restaurant dialogues
• SimDial
  • auto-generating dialogues based on domain descriptions
### NLG Datasets

- **Needs specific annotation**
  - other datasets typically not usable for NLG
  - unless you want to generate directly without explicit NLU & DM

- **Cambridge RNNLG**
  - restaurants, hotels, laptop, TVs
  - template-like

- **E2E NLG**
  - restaurants, bigger (50k instances)
  - more complex, more messy

- **MultiWOZ (has the needed annotation)**

---

#### Inform

```
inform(type=restaurant;count='2';food=basque;kidsallowed=no;price range=moderate)
there are 2 restaurants where no children are allowed in the moderate price range and serving Basque food
```

#### Request

```
?request(near)
where would you like it to be near to
```

---

**Loch Fyne is a kid-friendly restaurant serving cheap Japanese food.**

**Serving low cost Japanese style cuisine, Loch Fyne caters for everyone, including families with small children.**

---

**Wen et al., 2016**

http://arxiv.org/abs/1603.01232

---

**Novikova et al., 2017**

https://www.aclweb.org/anthology/W17-5525/
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)
    • should be more people, so overall not so subjective 😊
    • still not repeatable (different people will have different opinions)
  • **objective** = measuring properties directly from data (~automatic)
    • might or might not correlate with users’ perception

• Evaluation discussed here is **quantitative**
  • i.e. measuring & processing numeric values
  • *(qualitative ~ e.g. in-depth interviews, more used in social science)*
Getting the Subjects (for extrinsic 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
Extrinsic – Task-Oriented (Objective)

How to measure:

1) **Record people** while interacting with your system
2) **Analyze the logs**

Metrics:

- **Task success** (boolean): 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 (fewer is better here)
- **Other**: % returning users, % turns with null semantics …
Extrinsic – Task-Oriented (Subjective)

- **Questionnaires** for users/testers
  - based on what information you need

- 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)

- Question guidelines:
  - easy to understand
  - not too many
  - neutral: not favouring/suggesting any of the replies
Extrinsic – Task-Oriented (Subjective)

Example questions:

• **Success rate:** Did you get all the information you wanted?
  • typically different from objective measures!

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

• **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% (±3.79)</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., 2012)
https://doi.org/10.1016/j.csl.2011.09.004
Extrinsic – Non-Task-Oriented

Objective metrics:
• **Duration** – most common, easiest to get
  • longer = better here
• other (non-standard):
  • % returning users
  • checks for users swearing vs. thanking the system

Subjective:
• Future use + other same as task-oriented (except task success)
• **Likeability/Engagement**: Did you enjoy the conversation?
Intrinsic – ASR

• **Word error rate**
  • ASR output (hypothesis) compared to human-authored reference
  
  \[
  \text{WER} = \frac{\#\text{substitutions} + \#\text{insertions} + \#\text{deletions}}{\text{reference length}}
  \]

  • ~ length-normalized edit distance (**Levenshtein distance**)  
  • sometimes insertions & deletions are weighted 0.5x  
  • can be >1  
  • assumes one correct answer

true: I want a restaurant  
ASR: want a rest or rant

  \[
  \text{WER} = 1 + 2 + 1 / 4 = 1
  \]
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)

  - **Example**

    **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
  - alternatively also **exact matches** on the whole semantic structure
    - easier, but ignores partial matches

- Again, one true answer assumed

- NLU on ASR outputs vs. human transcriptions
  - both options make sense, but measure different things!
  - intrinsic NLU errors vs. robustness to ASR noise
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)
  - **agenda-based** (goal: constraints, agenda: stack of pending DAs)
  - **n-gram** models over DA/dialogue turns + sampling from distribution

Problem: simulator quality & implementation cost

- the simulator is basically another dialogue system
• No single correct answer here
  • many ways to say the same thing
• **Word-overlap** with reference text(s): **BLEU score**

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

- **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

---

Intrinsic – NLG

(Papineni et al., 2002)
https://www.aclweb.org/anthology/P02-1040
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.

output: What price range would you like?

ref1: What is your price range?
ref2: What price are you looking for?

matching unigrams: the (2x), Richmond, address, is (2x), 615, Balboa, Street, . (only 1x!), number, 4153798988, What, price, range, you, ?
p_1 = 17 / 22

matching bigrams: The Richmond, address is, is 615, 615 Balboa, Balboa Street, number is,
p_2 = 10 / 20 is 4153798988, 4153798988 ., What price, price range

p_3 = 5 / 18, p_4 = 2 / 16, BP = 1, BLEU = 0.3403

**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
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 – WER-style

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

\[
D = \frac{\#\text{distinct } x}{\#\text{total } x}, \text{ where } x = \text{unigrams, bigrams, sentences}
\]

(Wen et al., 2015)
http://aclweb.org/anthology/D15-1199

output: The <hotel> ’s address is <addr> . The phone number is <phone> .
ref: The number for <hotel> is <phone> , the address is <addr> .
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
• (also, never compare scores across datasets)
  • seems obvious, but people do it
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 → pretty sure it’s not just an accident
  • more test data = more independent results → 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 → obtain range of scores
    3) check if range overlap is less than 5% (1%...)
• You **need data (corpus)** to build your systems
  • various sources: human-human, human-machine, generated
  • various domains
  • size matters
• Some models need **annotation** (e.g. dialogue acts)
  • annotation is hard, ambiguous – need to check **agreement**
• **Evaluation** needs to be done on a **test set**
  • **objective** (measurements) / **subjective** (asking humans)
  • **intrinsic** (component per se)
    • ASR: WER, NLU: slot F1 + intent accuracy, NLG: BLEU
  • **extrinsic** (in application)
    • objective: success rate, # turns; subjective: likeability, future use (…)
  • don’t forget to check **significance**
• Next week: linguistics of dialogue
Thanks

Contact us:
https://ufaldsg.slack.com/
{odusek,hudecek}@ufal.mff.cuni.cz
Skype/Meet/Zoom (by agreement)

Get the slides here:
http://ufal.cz/npfl123

References/Inspiration/Further:
Apart from materials referred directly, these slides are based on:
• Iulian V. Serban et al.’s Survey of corpora for dialogue systems (Dialogue & Discourse 9/1, 2018):
  https://breakend.github.io/DialogDatasets/
• 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
• Wikipedia: Cohen’s kappa Levenshtein distance Word error rate