NPFL099 Statistical Dialogue Systems

2. Data & Evaluation

http://ufal.cz/npfl099

Ondřej Dušek, **Vojtěch Hudeček** & Tomáš Nekvinda

11. 10. 2021

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

CC BY SA
unless otherwise stated

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

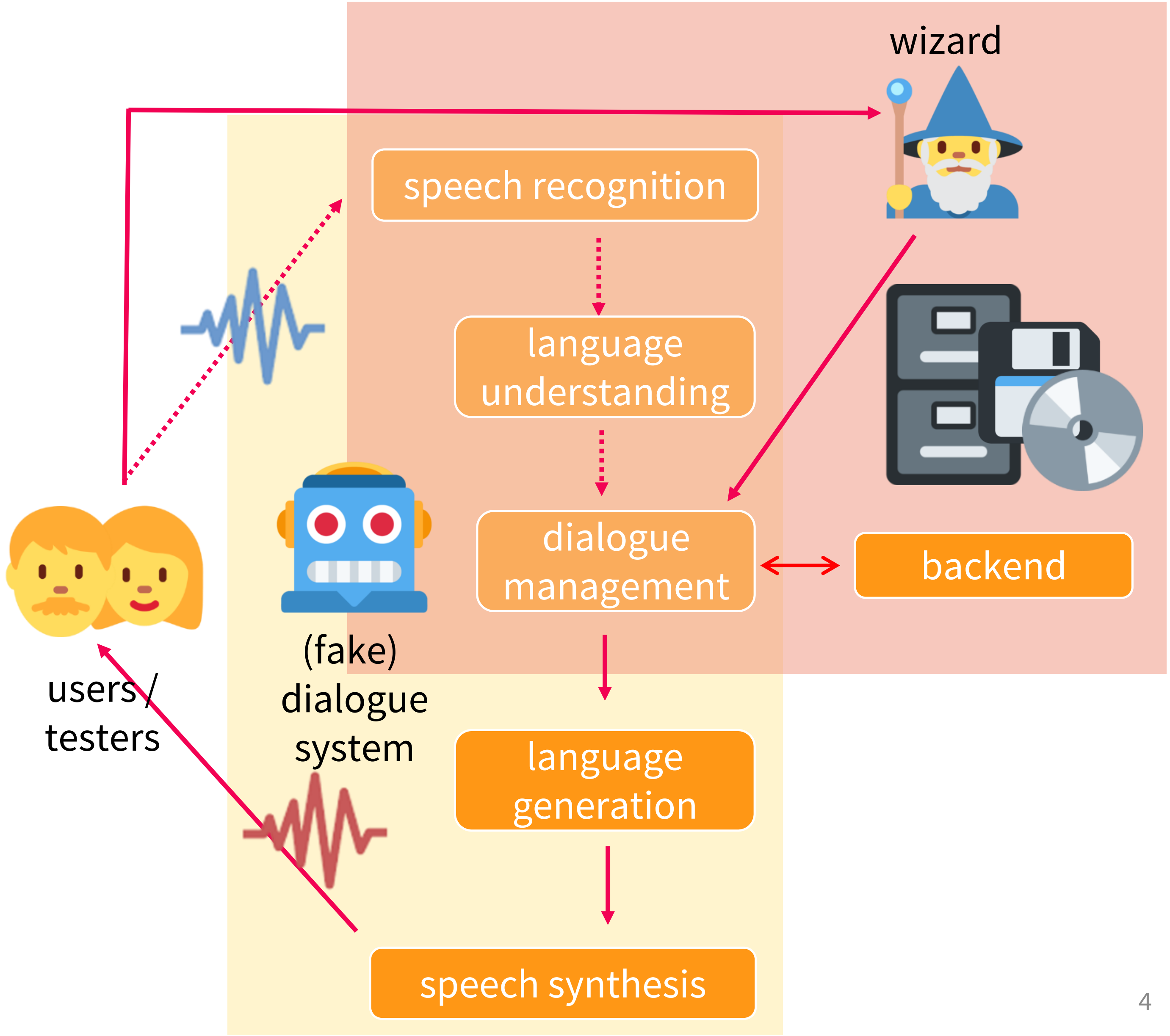
Dialogue Data Collection

- 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)

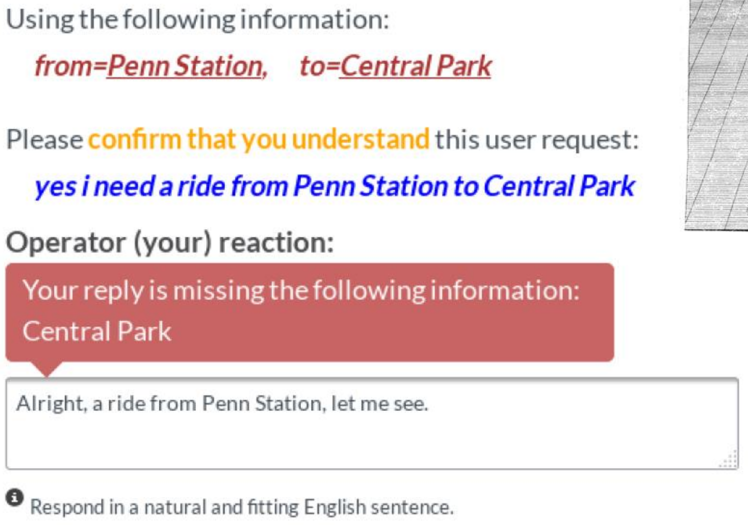
- for in-house data collection
  - also: to prototype/evaluate a system before implementing it!
- users believe they're talking to a system
  - different behaviour than when talking to a human
  - typically simpler
- system in fact **controlled by** a **human "wizard" (=you)**
  - typically selecting options (free typing too slow)

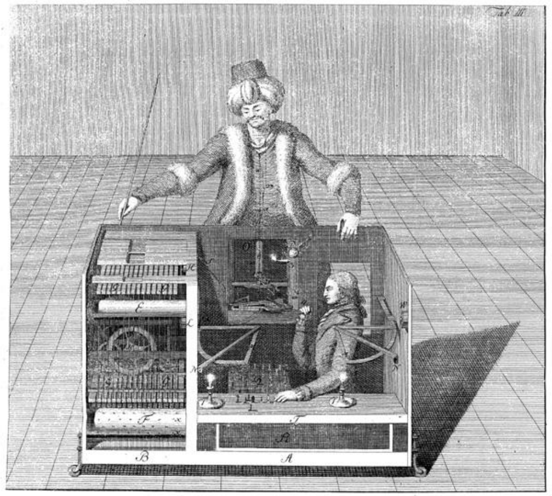


Crowdsourcing

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

- **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 (previously CrowdFlower/FigureEight)
  - Prolific.co





- 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

(Dušek & Jurčíček, 2016)
http://workshop.colips.org/wochat/@lrec2016/documents/02_Paper_6.pdf

Available Dialogue Datasets

- There's a number of research datasets available (see labs)
  - 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)

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
    - randomly re-draw your test set (same size, some items 2x/more, some omitted)
    - recompute scores on re-draw, repeat 1000x → obtain range of scores
    - check if range overlap is less than 5% (1%...)

Subjective Evaluation: Getting Subjects

- 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 (or access your website)
  - 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

Subjective Evaluation (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, just a slider)
- Question guidelines:
  - easy to understand
  - not too many
  - neutral: not favouring/suggesting any of the replies

Question Examples

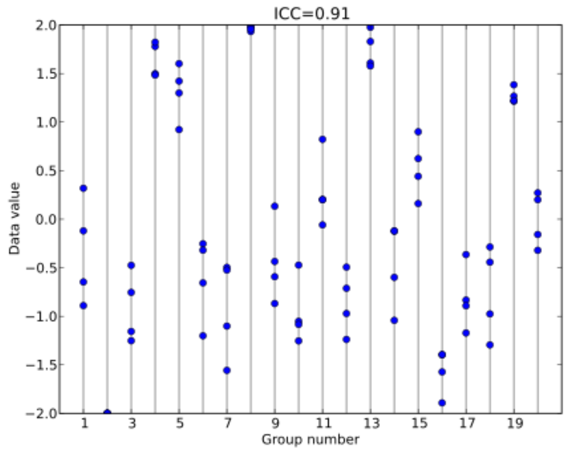
- **Success rate (task-oriented)**: Did you get all the information you wanted?
  - typically different from objective measures!
- **Future use**: Would you use the system again?
- **Likeability/engagement**: Did you enjoy the conversation?
- **ASR/NLU**: Do you think the system understood you well?
- **NLG**: Were the system replies fluent/well-phrased?
- **TTS**: Was the system's speech natural?

System	# calls	Subjective Success Rate	Objective Success Rate
HDC	627	82.30% (±2.99)	62.36% (±3.81)
NBC	573	84.47% (±2.97)	63.53% (±3.95)
NAC	588	89.63% (±2.46)	66.84% (±3.79)
NABC	566	90.28% (±2.44)	65.55% (±3.91)

(Jurčíček et al., 2012)
https://doi.org/10.1016/j.csl.2011.09.004

Question Types

- Aiming at rater consistency (multiple people rating the same)
  - high intraclass correlation coefficient (or other measure of agreement)
- **Likert** vs. **continuous**
  - Continuous scales seem to increase consistency
- alternatives: mainly for individual system outputs
  - too hard to do for whole dialogue
  - also work better than Likert
  - **Relative ranking** / Best-worst scaling
    - sort outputs from best to worst
    - variants: ties allowed / not
  - **Magnitude estimation**: continuous + reference value
    - rank-based: ask to assign values to multiple outputs at once
      - indirectly ranking



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

(Santhanam & Shaikh, 2019)
http://arxiv.org/abs/1909.10122

Intrinsic Objective Evaluation: NLU

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

(F1 is evenly balanced & default, other F variants favor *P* or *R*)

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)

$P = 1 / 3$
$R = 1 / 2$
$F = 0.2$

Intrinsic Objective Evaluation: 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

Extrinsic / Intrinsic Objective Evaluation: 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
- Problems:
  - cost: the simulator is basically another dialogue system
  - might not be fair (depending on the simulation accuracy)
    - typically your system would work better with a simulator than with humans

Extrinsic / Intrinsic Objective Evaluation: NLG

- No single correct answer here
  - many ways to say the same thing
- **Word-overlap** with reference text(s): **BLEU score**

range [0,1]
(percentage)

brevity penalty (1 if output longer than reference, goes to 0 if too short)

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{4} {}^{1}/_{4} \log\left(p_n\right)\right)$$

geometric mean

***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 = 11 / 15

matching bigrams: The Richmond, address is, is 615, 615 Balboa, Balboa Street, number is, is 4153798988, 4153798988 .
p_2 = 8 / 14
p_3 = 5 / 13, p_4 = 2 / 12, BP = 1, 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

Extrinsic / Intrinsic Objective Evaluation: NLG

- 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)

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

Extrinsic / Intrinsic Objective Evaluation: NLG

- **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

$$-\frac{1}{N}\sum_{i=1}^{N} \log q(x_i)$$

- 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

Extrinsic Objective Evaluation

- **Analyzing the logs** of people/testers/simulator interacting with the system
  - **multi-turn evaluation can work out differently from single-turn**

(Takanobu et al., 2020)
https://www.aclweb.org/anthology/2020.sigdial-1.37/

- 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

Retrieval metrics

- For retrieval/ranking systems
- **Recall**: $\mathbf{R}_{\boldsymbol{N}}@\boldsymbol{k}$
  - 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

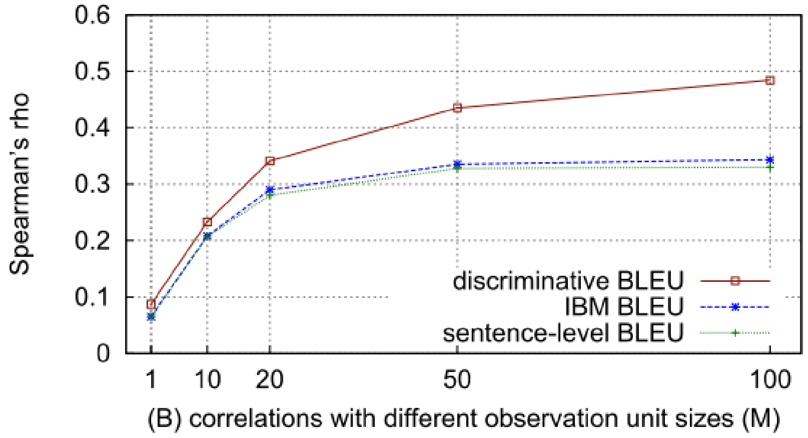
		Parameter	Description
current turn	*Exchange level*	ASRRecognitionStatus	ASR status: *success*, *no match*, *no input*
		ASRConfidence	confidence of top ASR results
		RePrompt?	is the system question the same as in the previous turn?
		ActivityType	general type of system action: *statement*, *question*
		Confirmation?	is system action confirm?
whole dialogue	*Dialogue level*	MeanASRConfidence	mean ASR confidence if ASR is success
		#Exchanges	number of exchanges (turns)
		#ASRSuccess	count of ASR status is success
		%ASRSuccess	rate of ASR status is success
		#ASRRejections	count of ASR status is reject
		%ASRRejections	rate of ASR status is reject
last 3 turns	*Window level*	{Mean}ASRConfidence	mean ASR confidence if ASR is success
		{#}ASRSuccess	count of ASR is success
		{#}ASRRejections	count of ASR status is reject
		{#}RePrompts	count of times RePromt? is true
		{#}SystemQuestions	count of ActivityType is question

“reject” = ASR output doesn’t match in-domain LM

(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/

ΔBLEU

- 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)
    - for appropriateness $\in [-1,1]$
  - **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

Trained Dialogue Metrics (works as intrinsic for NLG too)

- Train a supervised machine learning model
  - predict a score of “goodness” of each response
- Inputs may vary:
  - dialogue context + reference response (RUBER, USR)
    - works similar to BLEU
    - predict if the response fits the context
    - alternative (**adversarial evaluation**): is the response human-written or not?
  - context + training human ratings = **quality estimation**
    - can be used at system runtime – e.g. select best reply candidate
  - just context (FED)
    - using a pretrained language model
    - how likely the sentence is (~ fluency)
    - how likely it is that something positive/negative comes afterwards
- Better correlation with people than BLEU, but still not great (~0.4-0.5)

(Tao et al., 2018)
http://arxiv.org/abs/1701.03079
(Mehri & Eskenazi, 2020)
https://aclanthology.org/2020.sigdial-1.28/

(Bruni & Fernandez, 2017)
http://aclanthology.org/W17-5534

(Dušek et al., 2017; 2019)
https://arxiv.org/abs/1708.01759
https://arxiv.org/abs/1910.04731

(Mehri & Eskenazi, 2020)
https://aclanthology.org/2020.acl-main.64/

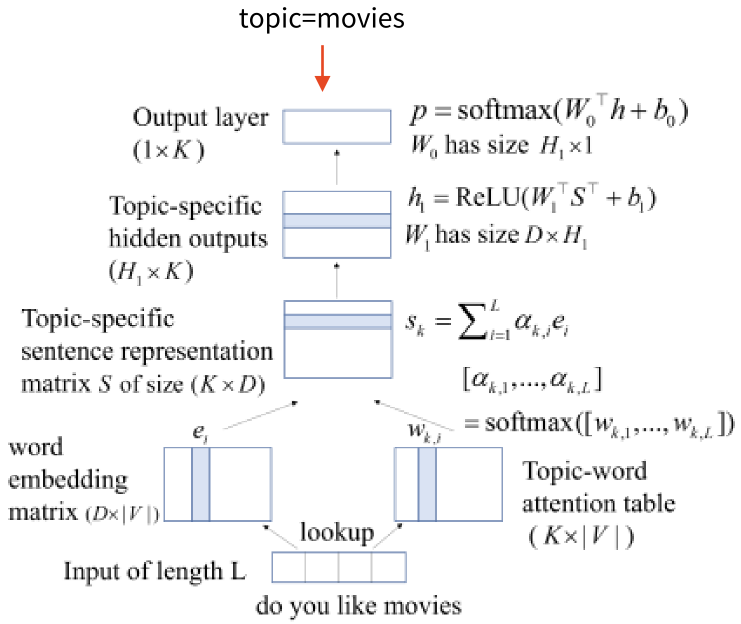
Chatbots: Self-play

- **Let the system be its own user simulator**
- Have it talk to itself + **measure some dialogue properties**
  - sentiment: sentiment classification + changes over dialogue
  - semantics/embedding: coherence ~ embedding similarity
  - engagement: # words + # ?'s in responses
- Result = linear combination of ↑, on 10-turn generated dialogues
  - seems to work pretty good (correlation ~0.7)
  - better than individual metrics, better than measuring individual turns

(Ghandeharioun et al., 2019)
http://arxiv.org/abs/1906.09308

Chatbots: 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

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 an unseen **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: Machine learning

Thanks

Contact us:

https://ufaldsg.slack.com/
{odusek,hudecek,nekvinda}@ufal.mff.cuni.cz
Zoom/Slack/Troja (by agreement)

Get the slides here:

http://ufal.cz/npfl099

Lab 17:20
1st homework assignment

Next Lecture
Monday 15:40
(no lab)

References/Further:

- Deriu et al. (2019): Survey on Evaluation Methods for Dialogue Systems: http://arxiv.org/abs/1905.04071
- Santhanam & Shaikh (2019): Towards Best Experiment Design for Evaluating Dialogue System Output https://www.aclweb.org/anthology/W19-8610/
- Takanobu et al. (2020): Is Your Goal-Oriented Dialog Model Performing Really Well? Empirical Analysis of System-wise Evaluation https://www.aclweb.org/anthology/2020.sigdial-1.37/
- Filip Jurčíček's slides (Charles University): https://ufal.mff.cuni.cz/~jurcicek/NPFL099-SDS-2014LS/
- 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