Dialgue system evaluation

Milica Gašić

Dialogue Systems and Machine Learning Group,
Heinrich Heine University Düsseldorf

Evaluation

Human evaluation

Evaluation via Crowd

Dialogue Challenges

Dialogue datasets and toolkits

Dialogue system evaluation

- Achilles' heal of the dialogue system research
- This makes the dialogue system research more art than science

Component-wise evaluation

- In modular approach to spoken dialogue systems we can evaluate each module separately
- There are well-established measures how to do this.
- We need to establish what is the test set in the supervised learning case, environment in the reinforcement learning case and whether we are only considering the top output or the N-best list (or belief).

Automatic speech recognition

- Speech recognition is a supervised learning task
- Word-error-rate on test set
- Perplexity of language model

Semantic decoding

- Semantic decoding is a supervised learning task
- Accuracy of the top hypothesis
- ICE of the N-best list

Belief tracking

- Belief tracking is a supervised learning task
- Joint goal accuracy (top hypothesis)
- L2-norm (complete belief)

Dialogue management

- Reinforcement learning task
- Average reward/success/turns during training/testing on simulated or real user
- Speed of convergence as well as final performance

Natural language generation

- Natural language generation is a supervised learning task
- Slot error rate
- Naturalness
- Informativeness

Speech synthesis

- Preference test

What does this tell us about the system as a whole?

- Individual models performance are very valuable and important measures but only intermediate measure.
- When putting together the whole system many aspects come into play and it is vital to assess the performance via a human trial.

Human evaluation

- Before a dialogue system is deployed with real users it must be evaluated in a human trial.
- This means that the system (or several systems including the baseline) is evaluated with volunteers.
- When recruiting volunteers it is important to have a variety of demographics and ideally people who have not come in contact with the same or similar systems before.

Dialogue tasks

- Dialogue tasks must be carefully designed.
- They must be of varying difficulty for the system and sufficiently cover the dialogue domain or domains.

Example dialogue with a volunteer

- System: How can I help you?
- User: Chinese
- System: Can you please repeat that?
- User: Cheap
- System: Can you please repeat that?
- User: Centre

Dialogue tasks

- Dialogue tasks can be descriptive: *You would like to take out your girlfriend to a fancy restaurant to celebrate your promotion. She lives in the north of the city. If there is no suitable restaurant in the north something in the centre would be good too. She is really keen on Japanese food.*
- Descriptive tasks should use a variety of vocabulary but be written in such a way that they encourage the volunteer to use their own words while not deviating from the task.
- Pictorial tasks: it might be preferable to use pictures to describe the task so that the volunteer would use their own words when talking to the system.

During the dialogue

- When the volunteer has acquainted themselves with the tasks they can proceed to talk to the system.
- Researcher should not interfere with the interaction unless the subject gets upset or explicitly asks for help.

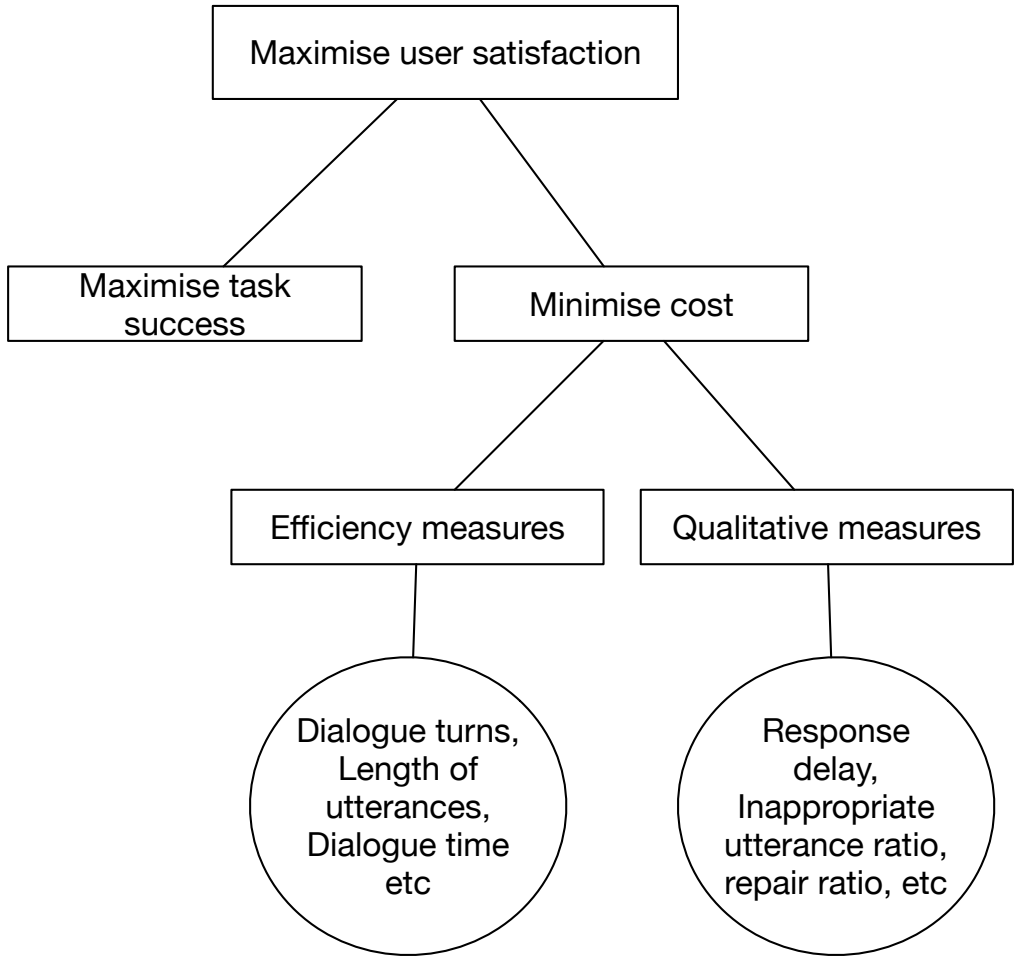
Questionnaire

- At the end of the interaction it is very important to ask the subject to fill in the previously prepared questionnaire
- The questions include:
  - Have you found what you were looking for?
  - I enjoyed the conversation (Likert scale)
- The questions can be module specific
  - The system was able to understand me (Likert scale)
  - The system output was well formulated (Likert scale)
- Note that the volunteer perception of what is wrong with the system may be different to reality.

Objective measures

- Volunteers and especially paid volunteers provide noisy feedback
- It is therefore important in addition to the subjective measures to objectively evaluate dialogues collected in the trial
- These include dialogue length, task completion, failure rate, etc.

PARADISE evaluation framework [Walker et al., 1997]



PARADISE evaluation framework

- definition of a task and a set of scenarios;
- experiments with alternate dialogue agents for the task;
- calculation of user satisfaction using surveys;
- calculation of task success
- calculation of dialogue cost using efficiency and qualitative measures;
- regression and values for user satisfaction, task success and dialogue costs;
- comparison with other agents/tasks to determine which factors generalise;
- refinement of the performance model.

How many interactions are necessary?

- It is important to acquire as many interactions as necessary to obtain statistically significant results
- This is not easy in an in-lab setting. Recruiting volunteers is costly, difficult and takes time.
- This is why nowadays it is more common to perform evaluation via crowd-sourcing.

Dialogue evaluation via crowd

- Platforms such as Amazon MTurk allow for small tasks to be posted to be performed by a human.
- Such tasks can be talking to a dialogue system and filling in a questionnaire.

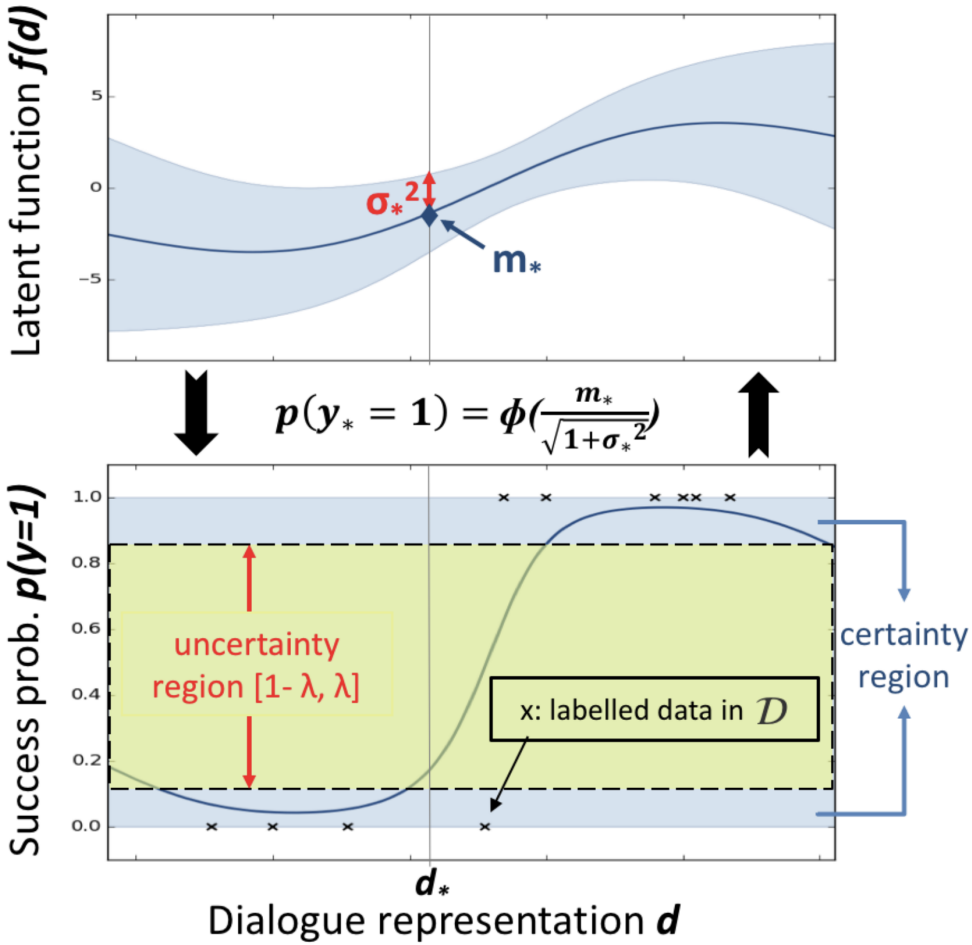
Advantages of crowd-sourcing

- Scaling up and speeding up evaluation
- Reduced cost

Disadvantages of crowd-sourcing

- Reduced quality of interactions
- Difficultly controlling who enters the experiment
- Difficulty asserting if the subjects follow the task
- Difficulty asserting the truth-fullness of the answers

Dealing with unreliable reward [Su et al., 2016]



Dialogue evaluation via real users

- Paid users have different motive from real users.
- Therefore results from a paid user trial do not always adequately reflect the system's real performance

"Let's go" Dialogue Challenge [Black et al., 2011]

- Carnegie Mellon University has for years had a system that was giving information about the Pittsburgh bus time table out of hours [Raux et al., 2005].
- This system was system-initiative, hand-coded but nevertheless popular among its users.
- In 2010 they organised a challenge: the data that their system collected over the years was released as the training data.
- Participating systems were evaluated in an in-lab human evaluation
- Best performing systems were selected for in-the-wild evaluation with real users

"Let's go" Dialogue Challenge outcome

- The results of the in-lab evaluation were not the same as in the in-the-wild evaluation
- Users were more used to system initiative questions and such systems were preferred.
- This process is called *entrainment*.

Dialogue State Tracking and Dialogue System Technology Challenges (DSTCs)

- Creating a shared dataset was a huge boost for the dialogue systems research and resulted in yearly challenges
- These were initially focused on dialogue state tracking but later examine a variety of aspects of dialogue systems

Dialogue datasets and toolkits

- A number of datasets and toolkits are available that facilitate comparing systems.
- They are typically used for development of various modules of dialogue systems though more recently they have been used for end-to-end approaches.
- Compared to speech resources and NLP resources, the resources available for dialogue are still very very modest.

ATIS

- The ATIS (Air Travel Information System) Pilot Corpus [Hemphill et al., 1990] is one of the first human-machine corpora.
- ATIS is a corpus of spoken dialogues and contains only 41 dialogues.
- The average number of turns is 25, which is quite high for a single domain dialogue.

Communicator

- The Carnegie Mellon Communicator Corpus [Bennett and Rudnicky, 2002] contains human-machine interactions with a travel booking system
- It contains 15K transcribed dialogues of average length 12 turns

DSTC datasets

- The DSTC datasets were released for the DSTC challenges.
- DSTC1 [Williams et al., 2013] features conversations with an automated bus information interface. It has 15K dialogues with 15 turns average length.
- DSTC2 introduces changing user goals in a restaurant booking system [Henderson et al., 2014]. The data was collected in a noisy car. It contains 3K dialogues with average number of turns 8.

Wizard of Oz data collection

- The way humans communicate to each other and the way they speak to machines is very different.
- This is the reason why it is often difficult to model dialogue system using human-human dialogue data.
- Human-machine dialogue data necessitates a machine and this is often a chicken-and-egg problem.
- Wizard of Oz data collection is somewhere in between: one human in conversation imitates a system and the other human does not know that they are actually speaking to a human.

MultiWOZ dialogue dataset [Budzianowski et al., 2018]

- Text dialogues collected using crowd-sourcing where several users contributed to the same dialogue.
- A dialogue can span multiple domains sometimes even within the same turn.
- Subjects produced the dialogues and labelled them in the same time.
- The corpus size is 8.5K and the average number of turns is 14.

PyDial dialogue toolkit [Ultes et al., 2017]

- Provides implementations of statistical approaches for all dialogue system modules
- It offers easy configuration, easy extensibility, and domain-independent implementations of the respective dialogue system modules.
- It has been extended to provide multi-domain conversational functionality

Pydial environments for benchmarking policy optimisation [Casanueva et al., 2018]

- Benchmarking environment where a fair comparison between different algorithms interacting can be established.
- These consist of different domains, user simulator settings and noise levels in the input.
- In total 18 environments are available.
- The implementation includes 4 state of the art dialogue policy optimisation algorithms.

Other toolkits

- Rasa - open source chatbot toolkit `https://rasa.com/`
- ConvLab - a multi-domain dialogue toolkit [Zhu et al., 2020]

Summary

- Evaluation is hard!
- Component-wise metrics are useful but not sufficient to determine the quality of the system as a whole.
- Human evaluation is essential. This includes in-lab evaluation, crowd-sourced evaluation and in-the-wild evaluation with real users.
- A number of dialogue datasets are available for training and testing various modules of a dialogue system. Still the size is modest in comparison to speech and NLP corpora.

References I


Bennett, C. L. and Rudnicky, A. I. (2002).
The carnegie mellon communicator corpus.
In Hansen, J. H. L. and Pellom, B. L., editors, *INTERSPEECH*. ISCA.


Black, A. W., Burger, S., Conkie, A., Hastie, H., Keizer, S., Lemon, O., Merigaud, N., Parent, G., Schubiner, G., Thomson, B., Williams, J. D., Yu, K., Young, S., and Eskenazi, M. (2011).
Spoken dialog challenge 2010: Comparison of live and control test results.
In *Proceedings of the SIGDIAL 2011 Conference*, SIGDIAL '11, page 27, USA. Association for Computational Linguistics.

References II


Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Stefan, U., Osman, R., and Gašić, M. (2018).
Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling.
In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.


Casanueva, I., Budzianowski, P., Su, P.-H., Mrki, N., Wen, T.-H., Ultes, S., Rojas-Barahona, L., Young, S., and Gai, M. (2018).
A benchmarking environment for reinforcement learning based task oriented dialogue management.


Hemphill, C. T., Godfrey, J. J., and Doddington, G. R. (1990).
The ATIS spoken language systems pilot corpus.
In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27,1990*.

References III


Henderson, M., Thomson, B., and Williams, J. D. (2014).
The second dialog state tracking challenge.
In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A. Association for Computational Linguistics.


Raux, A., Langner, B., Bohus, D., Black, A., and Eskenazi, M. (2005).
Let's go public! taking a spoken dialog system to the real world.
In *in Proc. of Interspeech*, pages 885–888.

References IV

Su, P.-H., Gašić, M., Mrkšić, N., Rojas-Barahona, L. M., Ultes, S., Vandyke, D., Wen, T.-H., and Young, S. (2016).
On-line active reward learning for policy optimisation in spoken dialogue systems.
In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2431–2441, Berlin, Germany. Association for Computational Linguistics.

Ultes, S., Rojas Barahona, L. M., Su, P.-H., Vandyke, D., Kim, D., Casanueva, I. n., Budzianowski, P., Mrkšić, N., Wen, T.-H., Gasic, M., and Young, S. (2017).
PyDial: A Multi-domain Statistical Dialogue System Toolkit.
In *Proceedings of ACL 2017, System Demonstrations*, pages 73–78, Vancouver, Canada. Association for Computational Linguistics.

References V


Walker, M. A., Litman, D. J., Kamm, C. A., and Abella, A. (1997).
PARADISE: A framework for evaluating spoken dialogue agents.
In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain. Association for Computational Linguistics.


Williams, J., Raux, A., Ramachandran, D., and Black, A. (2013).
The dialog state tracking challenge.
In *SIGDIAL Conference*.

References VI


Zhu, Q., Zhang, Z., Fang, Y., Li, X., Takanobu, R., Li, J., Peng, B., Gao, J., Zhu, X., and Huang, M. (2020).
Convlab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.