Joint On-line Learning of a Zero-Shot Spoken Semantic Parser and a Reinforcement Learning Dialogue Manager

Matthieu Riou, Bassam Jabaian, Stéphane Huet and Fabrice Lefèvre
matthieu.riou@alumni.univ-avignon.fr,
{bassam.jabaian, stephane.huet, fabrice.lefevre}@univ-avignon.fr
CERI-LIA, Avignon Université

Context and goals

Context of Spoken Dialogue Systems

- Development of stochastic-based approaches, uncertainty management
  - Better overall performance (correct hit rate, dialogue length)
  - More robust to variability in users' inputs and errors in the process chain
- But requires a large amount of annotated data

Goals

- Use on-line learning to address data issue
- Develop on-line learning models for semantic parser and dialogue manager
- Define joint on-line learning strategies for overall training

Joint On-Line Learning

Zero-Shot Semantic Parser

- Extracts a list of semantic concept hypotheses from an input sentence transcription of the user's query
- Is trained on-line using an adversarial bandit algorithm, by balancing the information improvement and the user effort of three possible actions:
  - **Skip**: skip the adaptation process for this turn
  - **AskConfirm**: a yes/no question is presented to the user about the correctness of the selected concepts in the best semantic hypothesis.
  - **AskAnnotation**: the user is asked to re-annotate the whole utterance.

Reinforcement Learning Paradigm for Dialogue Manager (DM)

- Dialogue management based on POMDP, **Hidden Information State (HIS)**
- Trained on-line with a Q-learner RL algorithm, **KTDQ learning algorithm**, by maximising an expected reward composed of:
  - the global feedback: the entire dialogue is a success or not
  - the social feedback given at each turn: only the last response is scored on a 5-point Likert scale

Two different joint learning protocols are proposed:

- **BR** juxtaposes a bandit model to learn the ZSSP and the Q-learner RL approaches to learn the dialogue manager policy
- **RR** adds the ZSSP learning actions to the dialogue manager RL policy, therefore combining the two learning processes into one single policy
  - The DM summary state vector is then augmented with a ZSSP-related dimension, evaluated from a set of quality indices:
    1. **confidence**: confidence score of the semantic parser
    2. **fertility**: ratio of concepts w.r.t. the utterance word length
    3. **rare**: binary presence of rare concepts in the annotation
    4. **known chunks**: ratio of annotated chunks available in the semantic knowledge base K among the total number of annotated chunks
    5. **gap**: the difference between the confidence scores of the 1-best and the 2-best annotations

Experimental Study

On-line training

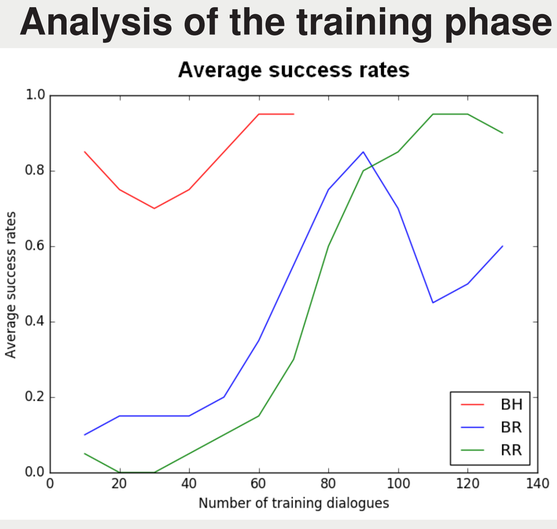
- The systems were trained by 3 expert users
- Two complementary systems in comparison:
  - **ZH**: a baseline system without on-line learning using the initial ZSSP and a handcrafted dialogue manager policy
  - **BH**: combines the on-line bandit learning for ZSSP and the handcrafted dialogue manager policy

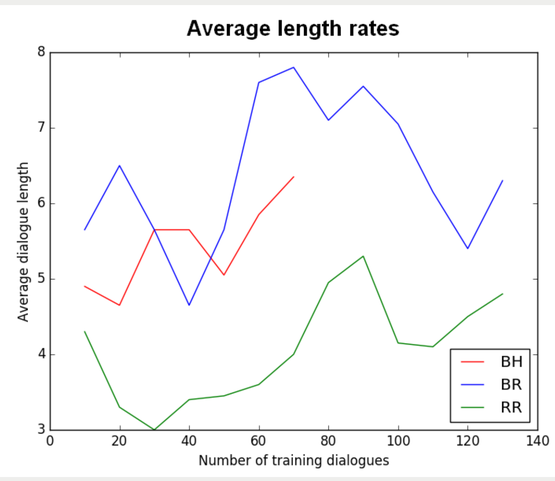
Evaluation of the systems

- Made in real conditions
- With 11 naive users and 2 expert users
- Based on an **evaluation survey**:
  - Success: "Was the task successful?" (0/1)
  - System Understandability: "Was the system easy to understand?" [0,5]
  - System Understanding: "Did the system understand you well?" [0,5]

Results

Analysis of the training phase





Evaluation of the different configurations of on-line learning

(13 participants)

Model	Train (#dial)	Test (#dial)	Success (%)	Avg cum. Reward	Sys. Underst. Rate	Sys. Gener. Rate
ZH	0	142	29	-1.9	1.6	4.0
BH	80	96	70	7.0	3.2	4.6
BR	140	96	89	10.9	3.3	4.6
RR	140	96	65	4.4	2.9	3.8

Experiments confirm that joint on-line learning:

- can be operated
- generally obtains good enough performances compared with a handcrafted system after only a hundred dialogues

Future work: investigation of the possibility of **merging the resulting policies** between trials