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

Matthieu Riou, Bassam Jabaian, Stéphane Huet, Fabrice Lefèvre

JOINT ON-LINE LEARNING OF A ZERO-SHOT SPOKEN SEMANTIC PARSER AND A REINFORCEMENT LEARNING DIALOGUE MANAGER

Matthieu Riou, Bassam Jabaian, Stéphane Huet, Fabrice Lefèvre

CERI-LIA, University of Avignon, France

firstname.lastname@univ-avignon.fr

ABSTRACT

Despite many recent advances for the design of dialogue systems, a true bottleneck remains the acquisition of data required to train its components. Unlike many other language processing applications, dialogue systems require interactions with users, therefore it is complex to develop them with pre-recorded data. Building on previous works, on-line learning is pursued here as a most convenient way to address the issue. Data collection, annotation and use in learning algorithms are performed in a single process. The main difficulties are then: to bootstrap an initial basic system, and to control the level of additional cost on the user side. Considering that well-performing solutions can be used directly off the shelf for speech recognition and synthesis, the study is focused on learning the spoken language understanding and dialogue management modules only. Several variants of joint learning are investigated and tested with user trials to confirm that the overall on-line learning can be obtained after only a few hundred training dialogues and can overstep an expert-based system.

Index Terms— on-line learning, adversarial bandit, reinforcement learning, zero-shot learning, spoken dialogue systems

1. INTRODUCTION

While a new avenue of research on end-to-end deep-learning-based dialogue systems has shown promising results lately [1, 2], a major hindrance remains the need of a huge quantity of data for these models to be trained efficiently. So far, in this case, it is not clear how some initial (low cost) knowledge can be leveraged for a warm start of the system development followed by on-line training with users as describe in [3, 4], although some recent works have proposed end-to-end architectures [5, 6].

In the experiments reported here our underlying goal is to develop a system intended to be used in a neuroscience experiment. From inside a fMRI, users interact with a robotic platform, vocally powered by our system, which is live-recorded and displayed inside the head-antenna. These experiments will be performed in French. Therefore, it is not possible to use the publicly available corpora since the vast majority is in English [7] and a new task is targeted (see section 5) for which no data are yet available.

As a consequence, this work still refers to a classical architecture, with proven capabilities, for goal-directed vocal interaction. It is basically a pipeline of modules dealing with the audio information from the user downstream; progressive processing aims to first extract the content (speech recognition), then the meaning (semantic parsing, SP), to finally combine it with previous information (including grounding status) from the dialogue history (belief tracking) so that a policy can decide upon this dialogue state representation the next best action to perform according to some global criteria (generally dialogue length and success in reaching the goal). This step of dialogue management (DM) is then followed by the following operations to convey back the information upstream to the user: conversion of the dialogue manager action into natural language (NLG) followed by speech synthesis. The HIS architecture [8] offers such a setup encompassing a global statistical framework to account for the relations between the data handled by the main modules of the system, allowing a reinforcement learning of the DM policy. This system can be implemented with sample-efficient learning algorithms [9] and can involve on-line learning through direct interactions with users [10]. More recently, on-line learning has been generalised to the input/output modules, SP and NLG, with protocols to control the cost of such operations during the system development (as in [11, 12, 13]). This work is a first attempt to combine the on-line learning of SP and DM in a single phase of development. Not only it is expected to help speed up and simplify the process, but also to benefit from intertwined improvements of the modules.

In dialogue systems, SP extracts a list of semantic concept hypotheses from an input sentence transcription of the user's query. State-of-the-art SPs are based on probabilistic approaches and trained with various machine learning methods to tag the user input with these semantic concepts [14, 15]. Dealing with supervised machine learning techniques requires a large amount of annotated data which are domain dependent and hardly available.

To deal with this limitation, Dauphin et al. [16] proposed a zero-shot learning algorithm for Semantic Utterance Clas-

sification (SUC). This method tries to find a sentence-wise link between categories and utterances in a semantic space. A deep neural network can be trained on a large amount of non-annotated and unstructured data to learn this semantic space. In the same line, in [11] was presented a zero-shot learning method for SP (ZSSP) based on word embeddings [17]. This approach requires neither annotated data nor in-context data and has recently been used for different dialogue systems' modules (as in [18, 19, 20]). Indeed, only the ontological description of the target domain and generic word embedding features (learned from freely available and general purpose data) are required to initiate the model. On top of that, an active learning strategy based on an adversarial bandit has been proposed [12] in order to train ZSSP with a light and controlled supervision from the users.

In the same line of ideas, thanks to the sample-efficient RL algorithm KTD [21], an active learning scheme has also been proposed for the DM training which uses reward shaping [22] to take into account local (turn-based) rewards from the user to offer a better control over the learning process and speed it up [10].

Since solutions exist for active on-line learning of both SP and DM subsystems, we now consider their joint application to address the issue of the overall training of the system. First, a direct application of existing techniques is presented and tested; both modules remain separated and the parameters of their on-line training are kept disjoint (a bandit algorithm for SP, a Q-learner for DM). Then a new possibility with shared parameters in a single Q-learner is also introduced and evaluated.

The remainder of this paper is organised as follows. After presenting the basis of the on-line learning versions of SP in Section 2 and DM in Section 3, we define the joint on-line learning strategies in Section 4. Section 5 provides an experimental study with human evaluations of the proposed approaches and we conclude in Section 6.

2. ON-LINE LEARNING FOR ZERO-SHOT SP

The SP model concerned by this study is the ZSSP model presented in [12]. This latter makes use of a semantic knowledge base K and a semantic feature space F. K contains some examples of lexical chunks associated with each targeted Dialogue Act (DA) and F is a word embedding representation learnt with neural network algorithms on large non-annotated open domain data [17, 23]. The ZSSP model builds a scored graph of hypotheses from user utterances. A best-path decoding is performed in order to find the best semantic tags hypothesis for the considered user utterance.

An on-line adaptation strategy (facilitated by the zero-shot approach) is adopted, as presented in [12] and briefly recalled here. In this approach, at each dialogue iteration, the system chooses an adaptation action $i_t \in \mathcal{I}$ and uses the user feedback to update K.

The system gain $g(i_t)$, the user effort $\phi(i_t)$ and the loss function $l(i_t)$ for performing each action are defined and can be estimated during on-line training.

Three possible actions are considered:

- **Skip**: Skip the adaptation process for this turn ($\phi(\text{skip}) = 0$).
- **AskConfirm**: A yes/no question is presented to the user about the correctness of the selected DAs in the best semantic hypothesis. If the whole sentence is accepted, $\phi(\text{YesNoQuestions}) = 1$. Otherwise, $\phi(\text{YesNoQuestions})$ is equal to $1+$ the number of DAs in the best semantic hypothesis (one yes/no confirmation request per DA).
- **AskAnnotation**: the user is asked to re-annotate the whole utterance. $\phi(\text{AskAnnotation}) = 1$ if the sentence is accepted straight away. Otherwise, the user will first inform the system about which chunks he wants to annotate ($+1$ per selected boundary), and then the system will sequentially ask for $acttype$, $slot$ and $value$ if necessary ($+1$ per interim question) for each DA.

An adversarial bandit algorithm is used in order to find $i_1, i_2, \ldots$, so that for every t, the system minimises the loss $l(i_t)$. The loss function $l(i) \in [0, 1]$ is calculated as follows:

$$l(i) := \underbrace{\gamma g(i)}_{\text{system improvement}} + \underbrace{(1-\gamma)\frac{\phi(i)}{\phi_{max}}}_{\text{user effort}},$$

where $\gamma \in [0, 1]$ balances the importance of information improvement and user effort for the system, and $\phi_{max} \in N^*$ is the maximum number of exchanges between the system and the user (in a same turn/round). In this work, γ has been set to 0.5 for example.

3. ON-LINE LEARNING FOR RL DIALOGUE MANAGER

The dialogue manager used in this paper adapts a system presented in [10]. It is based on a POMDP-based dialogue management framework, the Hidden Information State (HIS) [8]. In this setup, the system maintains a distribution over possible dialogue states (the belief state) and uses it to generate an adequate answer. A reinforcement learning (RL) algorithm is used to train the system by maximising an expected cumulative discounted reward.

At each turn, the dialogue manager generates several possible answers, depending on its belief state. It generates 11 dialogue acts, matching the 11 summary acts (Greet, Bye, Bold Request, Tentative Request, Confirm, Find Alternative, Split, Repeat, Offer, Inform and Request More). Some can be deemed impossible at some point if no conversion to full action is possible (for instance Inform if no entity is selected yet).

Subsequently, the dialogue manager chooses the best summary act according to the given context. To learn this policy, an RL approach is used: the KTDQ learning algorithm [24], derived from a Kalman-based Temporal Differences (KTD) framework. At each turn, the policy selects a summary act to answer the user, then a feedback is given by the users to score the response and update the policy. There are two types of feedback. The global feedback is given at the end of the dialogue by asking the user if the entire dialogue is a success or not. The social feedback s_i is given at each turn i to score the last response only. It is composed of two parts, the score given by the user to this last response (named additional feedback a_i) minus a Ψ function (which takes into account the history of annotations to smooth the local feedback), and the turn cost which penalises too long dialogues by adding a negative score (named feedback f_i) for each new turn : $s_i = f_i + (\theta a_i - \Psi)$

Here Ψ is the previous turn's additional-feedback a_{i-1} and $\theta = 0.95$. At the end of the dialogue, the policy is updated according to all the collected feedbacks.

In this work, the global feedback value is set to 20 in case of success, 0 otherwise. The feedback f_i is set to -1 for each turn and the additional-feedback $a_i \in \{-1, -0.5, 0, 0.5, 1\}$.

4. JOINT ON-LINE LEARNING

In order to effectively learn on-line the dialogue system, the user needs to be able to both improve the SP model and the dialogue manager. Two different joint learning protocols are proposed to achieve it.

The first one, referred to as **BR** hereafter, directly juxtaposes the bandit to learn the ZSSP and the Q-learner RL approaches to learn the dialogue manager policy. An adversarial bandit algorithm as described in Section 2 is applied for learning ZSSP and a Q-learner as mentioned in Section 3 is used to learn the DM policy. The knowledge base of the ZSSP as well as the DM policy are adapted after each dialogue turn.

The second protocol, referred to as **RR** hereafter, directly adds the ZSSP learning actions to the dialogue manager RL policy, and therefore combines the two learning processes into one single policy.

This variant of joint learning merges both policies in a single Q-learner. In that purpose the DM summary state vector was augmented with a ZSSP-related dimension. Let us note that only one dimension was added so as to limit the increase of the state size. This new dimension was evaluated from a set of quality indices of the annotations made by the ZSSP model. On a 3-point scale, five features were used:

1. **confidence**: confidence score of the semantic parser in $[0, 1]$.
2. **fertility**: ratio of concepts w.r.t. the utterance word length in $[0, 1]$, since ZSSP tends to produce an oversegmentation of the incoming utterances with inserted concepts.
3. **rare**: binary presence of rare concepts in the annotation. Rare concepts are "help", "repeat", "restart", "reqalts", "reqmore", "ack" or "thankyou", and are wrongly annotated in general.
4. **known chunks**: ratio of annotated chunks available in the semantic knowledge base K among the total number of annotated chunks in $[0, 1]$.
5. **gap**: the difference between the confidence scores of the 1-best and the 2-best annotations. Since those differences are very low (< 0.01), natural logarithm is applied to break out the data in order to have more readable values.

From these features, the ZSSP-related dimension is computed as:

- 0 *all clear*: rare $= 0$ and confidence $<= 0.499$ and fertility $<= 0.4$ and known chunks $>= 0.5$ and gap $>= -5.5$
- 1 *average condition*: rare $= 0$ and fertility $<= 0.5$ and known chunks $>= 0.15$ and gap $>= -6.5$ and (confidence > 0.499 or fertility > 0.4 or known chunks < 0.5 or gap < -5.5)
- 2 *alarming*: rare $= 1$ or fertility > 0.5 or known chunks < 0.15 or gap < -6.5

Under the RR protocol, the two ZSSP-annotation actions (AskConfirm and AskAnnotation, see Section 2) are also included inside the list of summary actions that can be picked up by the dialogue policy. In such case, the user is presented with the appropriate annotation window in the system's graphical interface and can correct the current annotation. Purely vocal interactions for this process are under study. Yet feasible, it remains a challenging task which could introduce errors of its own, so it seemed more appropriate to evaluate the whole process first with a graphical interface and no input errors. Once done, the turn is updated (i.e. the annotation process has taken the place of the normal user audio response) and the dialogue is pursued. Even though it might be possible that the policy learned it by itself, we chose to inhibit two Ask actions in a row (they are tagged as impossible in the next turn). Finally, these two ZSSP-annotation actions have a specific social-feedback: instead of -1, the feedback f_i uses the loss function $l(i)$ defined in Section 2 and rescaled to obtain a score $\in [-1, 1] : f_i = (1.0 - l_i) \times 2 - 1$.

5. EXPERIMENTAL STUDY

5.1. Task Description

Experiments presented in this paper concern a chit-chat dialogue system framed in a goal-oriented dialogue task. In this context, users discussed with the system about an image (out of a small predefined set of 6), and jointly tried to discover the message conveyed by the image, as described in [25]. In

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

Table 1. Evaluation of the different configurations of on-line learning

order to use a goal-oriented system for such a task, the principle which was followed was to construct, as the system's back-end, a database containing several hundreds of possible combinations of characteristics of the image, each associated with a hypothesis of the conveyed message. During its interaction with the system, it is expected that the user progressively provides elements from the image matching entities in the database. This makes the system select a small subset of possible entities from which it can pick both additional characteristics to inform the user with, or ultimately a pre-defined message to give as a plausible explanation for the image purpose. This allows the user to speak rather freely about the image for several tens of seconds before arguing briefly about the message. No argumentation is possible from the system's side, it can only propose a canned message and the discussion is expected to last around one minute at most.

The task-dependent knowledge base used in the experiments is derived from the INT task description [25], as well as from a generic dialogue information task. The semantics of the domain is represented by 16 different act types, 9 slots and 51 values. The 53 lexical forms used to model act types were manually elaborated.

5.2. Results

The evaluation of the two joint learning approaches is presented here. Two complementary systems are proposed in comparison: **ZH** is a baseline system without on-line learning using the initial ZSSP and a handcrafted dialogue manager policy, whereas the system **BH** combines the bandit online learning for ZSSP and the handcrafted dialogue manager policy.

For each system, an expert user communicated with the system to train a model. Then a group of 11 naive users tested each model. Two expert users also tested the **ZH** model for a total of 46 dialogues. At the end of each session, the users were asked to rate on a scale of 0 (worst) to 5 (best) the understanding and generation qualities of the system. The amount of training dialogues as well as the number of test sets for each configuration are given in Table 1.

Regarding the training phase, we observed that the success rate tends to highly variate: at the beginning of the learning process, the expert is inclined to use simple dialogues to build an efficient dialogue manager policy, leading to a large increase of the success rate. Then, when the system starts to be usable, more sophisticated dialogues are tested to teach more adaptability to the system. During the training, it drifts towards a decrease of the reward and success rates.

The user trials of the two training trials for each protocol are given in Table 1. The results show that the different configurations of the system display acceptable performance. The BR model trained with 140 dialogues shows the best success rate (89%) and significantly[1] over-performs all other models. Moreover, the ZH model leads to significantly[1] lower success values than all other models. The difference in performance between the ZH and the BH models (+41 points) shows the impact of the ZSSP adaptation on the overall success of the conversation, along with a better understanding (rates of 1.6 for ZH vs. 3.2 for BH). The average cumulated reward on the test is directly correlated with the success rate and confirms previous findings. Besides, due to a well-tuned template-based generation system, the system generation rate is high (≥ 3.8) for all configurations.

The RR protocol offers smaller success rates than BH and BR (65% for RR vs. 89% for BR). After analysing the training logs, it seems to be related to the very low triggering level of the ZSSP learning actions after the exploration steps during RR w.r.t the use of the bandit in BH and BR. To remedy this shortcoming, the policy state space should be modified to take a better account of the situations favourable to ZSSP actions, while preserving its capacities of discrimination for the dialogue actions. Anyhow, this approach remains to be developed further and improved as it is based on a unique framework for joint learning, which simplifies the system elaboration from a programming point of view.

6. CONCLUSION

After proposing methods to interactively train both semantic parsing and dialogue management on-line, this paper proposed and evaluated ways to combine them in a joint learning process. Experiments have been carried out in real conditions and are therefore scarce. Yet it has been possible to show that joint learning can be operated, and that after only a hundred dialogues the performance of the various configurations tested were generally good enough compared to a handcrafted system.

Based on these results, we now investigate the possibility of merging the resulting policies between trials, so as to be able to pile up training data coming from different users and save even more time to the system developers.

[1] Statistical significances were analysed with a two-tailed Welch's t-test. Results were considered statistically significant with a p-value < 0.001.

7. REFERENCES

[1] T. Wen, D. Vandyke, N. Mrkšić, M. Gasic, L. M. Rojas Barahona, P.-H. Su, S. Ultes, and S. Young, "A network-based end-to-end trainable task-oriented dialogue system," in *ACL*, 2017, pp. 438–449.

[2] X. Li, Y.-N. Chen, L. Li, J. Gao, and A. Celikyilmaz, "End-to-end task-completion neural dialogue systems," in *IJCNLP*, 2017, vol. 1, pp. 733–743.

[3] E. Ferreira and F. Lefèvre, "Expert-based reward shaping and exploration scheme for boosting policy learning of dialogue management," in *ASRU*, 2013.

[4] M. Gašić, F. Jurčíček, S. Keizer, F. Mairesse, B. Thomson, K. Yu, and S. Young, "Gaussian processes for fast policy optimisation of POMDP-based dialogue managers," in *SIGDIAL*, 2010.

[5] B. Dhingra, L. Li, X. Li, J. Gao, Y.-N. Chen, F. Ahmed, and L. Deng, "Towards end-to-end reinforcement learning of dialogue agents for information access," in *ACL*, 2017, vol. 1, pp. 484–495.

[6] P. Shah, D. Hakkani-Tur, B. Liu, and G. Tur, "Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning," in *ACL*, 2018, vol. 3, pp. 41–51.

[7] I. Casanueva, P. Budzianowski, P.-H. Su, N. Mrkšić, T.-H. Wen, S. Ultes, L. Rojas-Barahona, S. Young, and M. Gašić, "A benchmarking environment for reinforcement learning based task oriented dialogue management," *arXiv preprint arXiv:1711.11023*, nov 2017.

[8] S. Young, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu, "The hidden information state model: A practical framework for pomdp-based spoken dialogue management," *Computer Speech and Language*, vol. 24, no. 2, pp. 150–174, 2010.

[9] L. Daubigney, M. Geist, S. Chandramohan, and O. Pietquin, "A comprehensive reinforcement learning framework for dialogue management optimization," *Selected Topics in Signal Processing*, vol. 6, no. 8, pp. 891–902, 2012.

[10] E. Ferreira and F. Lefèvre, "Reinforcement-learning based dialogue system for human-robot interactions with socially-inspired rewards," *Computer Speech & Language*, vol. 34, no. 1, pp. 256–274, 2015.

[11] E. Ferreira, B. Jabaian, and F. Lefèvre, "Online adaptative zero-shot learning spoken language understanding using word-embedding," in *ICASSP*, 2015.

[12] E. Ferreira, A. Reiffers-Masson, B. Jabaian, and F. Lefèvre, "Adversarial bandit for online interactive active learning of zero-shot spoken language understanding," in *ICASSP*, 2016.

[13] M. Riou, B. Jabaian, S. Huet, and F. Lefèvre, "Online adaptation of an attention-based neural network for natural language generation," in *INTERSPEECH*, 2017.

[14] S. Hahn, M. Dinarelli, C. Raymond, F. Lefèvre, P. Lehnen, R. De Mori, A. Moschitti, H. Ney, and G. Riccardi, "Comparing stochastic approaches to spoken language understanding in multiple languages," *IEEE TASLP*, vol. 19, no. 6, pp. 1569–1583, 2010.

[15] A. Deoras and R. Sarikaya, "Deep belief network based semantic taggers for spoken language understanding," in *INTERSPEECH*, 2013.

[16] Y. Dauphin, G. Tur, D. Hakkani-Tur, and L. Heck, "Zero-shot learning and clustering for semantic utterance classification," *arXiv preprint arXiv:1401.0509*, 2014.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[18] S. Upadhyay, M. Faruqui, G. Tür, D.. Hakkani-Tur, and L. Heck, "(almost) zero-shot cross-lingual spoken language understanding," in *2018 ICASSP*. IEEE, 2018, pp. 6034–6038.

[19] T. Zhao and M. Eskenazi, "Zero-shot dialog generation with cross-domain latent actions," *arXiv preprint arXiv:1805.04803*, 2018.

[20] A. Bapna, G. Tur, D. Hakkani-Tur, and Larry Heck, "Towards zero-shot frame semantic parsing for domain scaling," *arXiv preprint arXiv:1707.02363*, 2017.

[21] M. Geist and O. Pietquin, "Kalman temporal differences," *Artificial Intelligence Research*, vol. 39, no. 1, pp. 483–532, Sept. 2010.

[22] A. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *ICML*, 1999.

[23] J. Bian, B. Gao, and T. Liu, "Knowledge-powered deep learning for word embedding," in *ECML*, 2014.

[24] M. Geist and O. Pietquin, "Managing uncertainty within value function approximation in reinforcement learning," in *Active Learning and Experimental Design workshop (AISTATS 2010)*, 2010, vol. 92.

[25] T. Chaminade, "An experimental approach to study the physiology of natural social interactions," *Interaction Studies*, vol. 18, no. 2, pp. 254–276, 2017.