Safe transfer learning for dialogue applications
Nicolas Carrara, Romain Laroche, Jean-Léon Bouraoui, Tanguy Urvoyn, Olivier Pietquin

To cite this version:
Nicolas Carrara, Romain Laroche, Jean-Léon Bouraoui, Tanguy Urvoyn, Olivier Pietquin. Safe transfer learning for dialogue applications. SLSP 2018 - 6th International Conference on Statistical Language and Speech Processing, Oct 2018, Mons, Belgium. 2018. <hal-01928102>
Safe transfer learning for dialogue applications.

Nicolas Carrara\textsuperscript{1,2}, Romain Laroche\textsuperscript{3}, Jean-Léon Bouraoui\textsuperscript{2}, Tanguy Urvoy\textsuperscript{2}, and Olivier Pietquin\textsuperscript{1}

\textsuperscript{1} Univ. Lille, CNRS, Centrale Lille, INRIA UMR 9189 - CRIStAL, Lille, France
name.surname@inria.fr

\textsuperscript{2} Orange Labs, Lannion, France
name.surname@orange.com

\textsuperscript{3} Microsoft Maluuba, Montréal, Canada
name.surname@microsoft.com

Abstract. In this paper, we formulate the hypothesis that the first dialogues with a new user should be handled in a very conservative way, for two reasons: avoid user dropout; gather more successful dialogues to speedup the learning of the asymptotic strategy. To this extend, we propose to transfer a safe strategy to initiate the first dialogues.

Keywords: Transfer Learning · Dialogue · Safety

1 Introduction

During its early steps of learning, a Reinforcement Learning (RL) \cite{8} based dialogue agent does a lot of exploration that may lead to penalizing behavior. However, the first interactions between a user and a dialogue system are crucial to gain his trust. To improve jump-start performance of an RL agent, one can transfer a strategy \cite{9,7}. In dialogue, the strategy focuses on the success of the dialogue while minimizing its length \cite{3,5,2,4}. Rushing the dialogue may be problematic with some users and induce premature dialogue hangups. In one hand, it could lead to the lose of this user once for all. In an other hand, the lack of succeeding dialogues may affect the learning speed of the RL agent. To this extend, we introduce a novel algorithm: $\epsilon$-safe. It transfers a safe strategy which avoid any critical dialogue act to avoid the aforementioned problems.

2 $\epsilon$-safe

$\epsilon$-safe (Fig 1) is a $Q$-learning algorithm \cite{6} where each action is decided by a randomly chosen policy among the greedy policy, an exploratory policy and the transferred safe policy. It may be an handcrafted policy, an RL policy (with large reward penalty on the catastrophic event) or even a safe-RL policy \cite{1}.
3 Experiment

We test our algorithm on a simple slot-filling application. The agent asks for slot values, slot by slot in a fixed order. Several acts are available:

- ask_next: ask next slot (with NLU errors).
- repeat_oral: repeat current slot (with NLU errors).
- repeat_numpad: repeat using numeric pad (without NLU errors).
- summarize_inform: summarize slots values and return the form result. If values are correct, the dialogue ends successfully; if not, the slot values are reset and the dialogue continues from the first slot.

repeat_numpad is an unsafe action: the user hangs up with probability \( p \). For each new user, \( p \) is randomly generated. An histogram of \( p \) values is displayed Fig. 2.

![Histogram of p values](image_url)

Fig. 2: Half-gaussian distribution of \( p \) values.
We define 2 handcrafted dialogue systems. The **safe** system uses `repeat_oral` if the recognition score is below 0.5, otherwise `ask_next`, or `summarize_inform` after the last slot; The **unsafe** system uses `repeat_numpad` instead. We compare two $\epsilon$-safe agents, **safe_on** uses **safe** as transferred policy while **unsafe_on** transfers the **unsafe** policy.

![Graphs showing dialogue score and success frequency](image)

Fig. 3: Performance of the greedy policies.

We test **safe_on** and **unsafe_on** with 10 randomly generated users, for 1000 dialogues. We repeat the experiment 10 times. We display the performance of the greedy policies. The dialogue score (reward penalized by dialogue length) is plotted on Fig 3a while the dialogue success (reward only) is plotted on Fig 3b. We see that despite a slightly difference on the dialogue success, the dialogue score is the same. That means $\epsilon$-safe doesn’t improve the learning speed of the greedy policy even if it is conservative enough to keep the user in the dialogue by avoiding catastrophic acts.

### 4 Future work

We believe that the hangup-model is too simple, so we plan to design the hangup-model with a Poisson distribution. We also want to replace the handcrafted policies by actual RL policies learned on source users. Finally, a real application on DSTC2 may be considered.
References