Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 256–274

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Reinforcement-learning based dialogue system for human–robot interactions with socially-inspired rewards☆

Emmanuel Ferreira*, Fabrice Lefèvre

LIA-CERI, University of Avignon, Avignon, France

Received 27 May 2014; received in revised form 11 February 2015; accepted 25 March 2015
Available online 2 April 2015

Abstract

This paper investigates some conditions under which polarized user appraisals gathered throughout the course of a vocal interaction between a machine and a human can be integrated in a reinforcement learning-based dialogue manager. More specifically, we discuss how this information can be cast into socially-inspired rewards for speeding up the policy optimisation for both efficient task completion and user adaptation in an online learning setting. For this purpose a potential-based reward shaping method is combined with a sample efficient reinforcement learning algorithm to offer a principled framework to cope with these potentially noisy interim rewards. The proposed scheme will greatly facilitate the system's development by allowing the designer to teach his system through explicit positive/negative feedbacks given as hints about task progress, in the early stage of training. At a later stage, the approach will be used as a way to ease the adaptation of the dialogue policy to specific user profiles. Experiments carried out using a state-of-the-art goal-oriented dialogue management framework, the Hidden Information State (HIS), support our claims in two configurations: firstly, with a user simulator in the tourist information domain (and thus simulated appraisals), and secondly, in the context of man–robot dialogue with real user trials.

Keywords: Human–robot interaction; POMDP-based dialogue management; Reinforcement learning; Reward shaping

1. Introduction

In a goal-oriented vocal interaction between a machine and a human, the dialogue manager (DM) is responsible for making appropriate dialogue decisions to fulfil the user goal based on uncertain dialogue contexts. The Partially Observable Markov Decision Process (POMDP) framework (Kaelbling et al. (1998)) has been successfully employed in the Spoken Dialogue System (SDS) field (Young et al., 2010; Thomson and Young, 2010; Pinault and Lefèvre, 2011) as well as in human robot interaction (HRI) context (Roy et al., 2000; Lucignano et al., 2013), due to its capacity to explicitly handle parts of the inherent uncertainty of the information which the system has to deal with (e.g. erroneous speech recognizer, falsely recognised gestures, etc.). In this setup, the agent maintains a distribution over possible dialogue states, referred to as the belief state in the literature, and interacts with its perceived environment using a

☆ This paper has been recommended for acceptance by Roger K. Moore.

* Corresponding author. Tel.: +33 490843500.
E-mail addresses: emmanuel.ferreira@univ-avignon.fr (E. Ferreira), fabrice.lefevre@univ-avignon.fr (F. Lefèvre).

http://dx.doi.org/10.1016/j.csl.2015.03.007
0885-2308/

reinforcement learning (RL) algorithm so as to maximise some expected cumulative discounted reward (Sutton and Barto, 1998).

Recent studies in SDS have shown the possibility to learn a dialogue policy from scratch with a limited number (several hundreds) of interactions (Gašić et al., 2010; Sungjin and Eskenazi, 2012; Daubigney et al., 2012) and the potential benefit of this approach compared to the classical use of a Wizard-of-Oz or developing a well-calibrated user simulator (Gašić et al., 2010). Following this idea, sample-efficient learning algorithms, as for instance the Kalman Temporal Differences (KTD) framework (Geist and Pietquin, 2010; Daubigney et al., 2012), can be employed to learn and adapt a system behaviour in an online setup, i.e. while interacting with users. The main shortcoming of this approach is the very poor initial performance. Lowering the length of this warm-up learning phase, defined as the phase when the system can hardly interact with real users due to a high level of exploration and poor performance, is still an open problem when such systems are to be applied to real-world domains. Some solutions can be to introduce some initial expert knowledge (Williams, 2008) or to find ways to collect more hints from the environment which will accelerate the policy learning.

Moreover, problems addressed by RL generally introduce non-stationa-rity at several levels. Indeed, as in many real-world machine learning applications, adaptation to non-stationary environments is a desired feature. In the DM case, users with various levels of expertise (from novice to advanced) and characteristics (restless, bad pronunciations, bad hearing, etc.) can interact with the system. So, the latter should be able to cope with a wide range of behaviours, which may also change over time (switch to new users but also user self-adaptation to the system). Another source of non-stationarity arises when the policy iteration scheme (Sutton and Barto, 1998) is adopted. Policy iteration is an iterative procedure which aims at discovering the optimal policy by generating a sequence of monotonically improving policies. Each iteration consists of two stages: policy evaluation which computes the value function of a given policy and policy improvement which defines the improved policy over the value function. The fact that the value function changes together with the policy makes it non-stationary. In all non-stationary contexts (e.g. environment, optimization method) tracking the value function instead of converging to it seems preferable. A more detailed discussion about the advantages of tracking versus converging, even in stationary environments, can be found in Sutton et al. (2007). Most existing RL algorithms assume stationarity of the problem at hand and aim at converging to a fixed solution. Actually, few attempts to handle non-stationarity can be found in the literature. Among them, we can mention a class of methods which combine RL and planning paradigms such as the Dyna-Q algorithm (Sutton and Barto, 1998).

In most works, the reward function used to learn the dialogue agent is exclusively based on objective features, such as duration and full completion of the user goal. The overall quality of such a function plays a crucial role in finding the optimal solution. However, recent studies have shown that such features, although objective, could not be collected with entire reliability from users (Gašić et al., 2010; Sungjin and Eskenazi, 2012). Anyhow, if the user's point of view is totally ignored or reduced to a rather simple satisfaction questionnaire, naturalness of the overall system can be impacted. In the PARADISE evaluation paradigm (Walker et al., 1997), subjective and objective features are correlated through linear regression. It is worth noting that subjective information is more easily produced by the user. Therefore, it may be interesting to gather some subjective features during the course of the dialogue in order to accelerate the policy learning instead of relying exclusively on an imprecise final appraisal.

This idea could be linked to some works in social and human sciences (e.g. psychology, anthropology) which have shown to which extent acts as simple and spontaneous as facial expressions or gestures can convey social meaning affecting our perception and shaping our daily interactions (Richmond et al., 1991; Kunda, 1999; Custers and Henk, 2005). Also Vinciarelli et al. (2009) present a wide coverage survey of an emerging domain aiming to endow computers with social intelligence abilities. And among these abilities some of the most important are correct perception, accurate interpretation and appropriate generation of social signals. So in the same line of thought, in this proof of concept study, we are focusing on the potential interest of considering a subclass of these social signals with which a user conveys some raw assessments of the current situation during the course of the interaction. Indeed, we claim that positive/negative user appraisals gathered during the course of the dialogue can be used to partially address the two aforementioned bootstrap and tracking problems.

By the fact that user appraisals can be gathered all along the dialogue, we intend to directly exploit them in a socially-inspired diffuse and interim reward function employed in online learning strategy. In that sense, the formulated problem can be closely related to the reward shaping one. In RL, reward shaping consists in supplying meaningful diffuse rewards to a learning agent with the objective to speed up the learning towards the same optimal policy than the one that we could reach with a sparse reward function, giving the meaningful reward only at the end of an episode (e.g. task