Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 51 (2018) 24–43

www.elsevier.com/locate/csl

Reward estimation for dialogue policy optimisation☆

Pei-Hao Su*, Milica Gašić, Steve Young

University of Cambridge, Engineering Department, Cambridge CB2 1TP, United Kingdom

Received 23 March 2017; received in revised form 5 October 2017; accepted 19 February 2018
Available online 24 February 2018

Abstract

Viewing dialogue management as a reinforcement learning task enables a system to learn to act optimally by maximising a reward function. This reward function is designed to induce the system behaviour required for the target application and for goal-oriented applications, this usually means fulfilling the user's goal as efficiently as possible. However, in real-world spoken dialogue system applications, the reward is hard to measure because the user's goal is frequently known only to the user. Of course, the system can ask the user if the goal has been satisfied but this can be intrusive. Furthermore, in practice, the accuracy of the user's response has been found to be highly variable. This paper presents two approaches to tackling this problem. Firstly, a recurrent neural network is utilised as a task success predictor which is pre-trained from off-line data to estimate task success during subsequent on-line dialogue policy learning. Secondly, an on-line learning framework is described whereby a dialogue policy is jointly trained alongside a reward function modelled as a Gaussian process with active learning. This Gaussian process operates on a fixed dimension embedding which encodes each varying length dialogue. This dialogue embedding is generated in both a supervised and unsupervised fashion using different variants of a recurrent neural network. The experimental results demonstrate the effectiveness of both off-line and on-line methods. These methods enable practical on-line training of dialogue policies in real-world applications.

Keywords: Dialogue systems; Reinforcement learning; Deep learning; Reward estimation; Gaussian process; Active learning

1. Introduction

Spoken Dialogue Systems (SDS) allow human–computer interaction using natural speech. They can be broadly divided into two categories: chat-based systems which converse with users with the aim of providing contextually relevant responses in broad domains (Vinyals and Le, 2015; Serban et al., 2015), and task-oriented systems designed to assist users to achieve specific goals (e.g. find hotels, movies or bus schedules) (Young et al., 2013). The latter are typically designed on top of a structured *ontology* (or a database *schema*), which defines the domain that the system can talk about. The development of such systems traditionally requires a substantial amount of hand-crafted rules combined with various statistical components. These include a spoken language understanding module (Henderson

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

* Corresponding author.
E-mail address: phs26@cam.ac.uk (P.-H. Su), mg436@cam.ac.uk (M. Gašić), sjy11@cam.ac.uk (S. Young).

http://dx.doi.org/10.1016/j.csl.2018.02.003

et al., 2012; Chen et al., 2016), a dialogue belief state tracker (Henderson et al., 2014) to predict user intents and track the dialogue history, a dialogue policy (Gašić and Young, 2014) to determine the dialogue flow, and a natural language generator (Wen et al., 2015) to convert abstract system responses into natural language. Teaching such a system how to respond appropriately in all situations is non-trivial. Traditionally, the *dialogue management* component has been designed manually using flow charts to directly specify system behaviour. More recently, it has been formulated as a planning problem and solved using reinforcement learning (RL) to enable automatic optimisation during interaction with users (Gašić and Young, 2014; Levin and Pieraccini, 1997; Roy et al., 2000; Williams and Young, 2007; Jurčíček et al., 2011; Li et al., 2016; Su et al., 2016a; Dhingra et al., 2016; Su et al., 2017). In this framework, the system learns by a *trial and error* process governed by a potentially delayed learning objective, a *reward function*.

A typical approach to defining the reward function in a task-oriented dialogue system is to apply a small per-turn penalty to encourage short dialogues and to give a large positive reward at the end of each successful interaction. Fig. 1 is an example of a dialogue task which is typically set when bootstrapping or evaluating a system using paid subjects. When paid subjects are primed with a specific task to complete, dialogue success can be determined from both subjective user ratings (*Subj*), and from an objective measure (*Obj*) based on whether or not the pre-specified task was completed (Walker et al., 1997; Gašić et al., 2013). However, when operating with real users, prior knowledge of the user's goal is not normally available and hence it is not possible to compute an objective reward.

Furthermore, objective ratings must necessarily be quite strict to ensure that paid subjects properly exercise the system. As a consequence, tasks often fail because the paid subject forgot to ask for a required piece of information. For example, in Fig. 1, the paid subject forgot to ask for the phone number. This results in a mismatch between the *Obj* and *Subj* ratings.

Unfortunately, relying on subjective ratings alone is also problematic since paid subjects frequently give inaccurate responses partly because they forget the complete task as in the example above, and also because they are concerned that their answers will affect their payment. Real users are often unwilling to extend the interaction in order to give feedback, and even when they do, their feedback can also be unreliable (Gašić et al., 2013), for example due to sociological effects such as not wishing to be impolite.

All of the above can result in unstable learning (Zhao et al., 2011; Gašić et al., 2011). When bootstrapping a system using paid users, an effective albeit inefficient solution is to ignore all dialogues for which the objective success assessment differs from the subjective success assessment (referred to as the $Obj = Subj$ check below) (Gašić et al., 2013). However, as well as being inefficient, this solution does not help address the real problem which is how to train systems on-line with real users when the user's goal is generally unknown and difficult to infer.

To deal with the above issues, this paper investigates the use of an independent reward estimator which can be trained to monitor a dialogue and accurately estimate task success independently of the specific goal set, or of whatever goal is in the user's mind. The investigation is in two parts.

We first describe a recurrent neural network (RNN) designed to estimate objective success *Obj*, which is trained from off-line simulated dialogue data (Su et al., 2015a; Vandyke et al., 2015). The resulting policy is found to be as

Fig. 1. An example of a task-oriented dialogue with a pre-defined task and the evaluation results.