Feudal Reinforcement Learning for Dialogue Management in Large Domains

Inigo Casanueva1∗, Paweł Budzianowski1, Pei-Hao Su2, Stefan Ultes1, Lina Rojas-Barahona1, Bo-Hsiang Tseng1 and Milica Gašić1
1Department of Engineering, University of Cambridge, UK
2PolyAI Limited, London, UK
ic340@cam.ac.uk

Abstract

Reinforcement learning (RL) is a promising approach to solve dialogue policy optimisation. Traditional RL algorithms, however, fail to scale to large domains due to the curse of dimensionality. We propose a novel Dialogue Management architecture, based on Feudal RL, which decomposes the decision into two steps; a first step where a master policy selects a subset of primitive actions, and a second step where a primitive action is chosen from the selected subset. The structural information included in the domain ontology is used to abstract the dialogue state space, taking the decisions at each step using different parts of the abstracted state. This, combined with an information sharing mechanism between slots, increases the scalability to large domains. We show that an implementation of this approach, based on Deep-Q Networks, significantly outperforms previous state of the art in several dialogue domains and environments, without the need of any additional reward signal.

1 Introduction

Task-oriented Spoken Dialogue Systems (SDS), in the form of personal assistants, have recently gained much attention in both academia and industry. One of the most important modules of a SDS is the Dialogue Manager (DM) (or policy), the module in charge of deciding the next action in each dialogue turn. Reinforcement Learning (RL) (Sutton and Barto, 1999) has been studied for several years as a promising approach to model dialogue management (Levin et al., 1998; Henderson et al., 2008; Pietquin et al., 2011; Young et al., 2013; Casanueva et al., 2015; Su et al., 2016). However, as the dialogue state space increases, the number of possible trajectories needed to be explored grows exponentially, making traditional RL methods not scalable to large domains.

Hierarchical RL (HRL), in the form of temporal abstraction, has been proposed in order to mitigate this problem (Cuayáhuitel et al., 2010, 2016; Budzianowski et al., 2017; Peng et al., 2017). However, proposed HRL methods require that the task is defined in a hierarchical structure, which is usually handcrafted. In addition, they usually require additional rewards for each subtask. Space abstraction, instead, has been successfully applied to dialogue tasks such as Dialogue State Tracking (DST) (Henderson et al., 2014b), and policy transfer between domains (Gašić et al., 2013, 2015; Wang et al., 2015). For DST, a set of binary classifiers can be defined for each slot, with shared parameters, learning a general way to track slots. The policy transfer method presented in (Wang et al., 2015), named Domain Independent Parametrisation (DIP), transforms the belief state into a slot-dependent fixed size representation using a handcrafted feature function. This idea could also be applied to large domains, since it can be used to learn a general way to act in any slot.

In slot-filling dialogues, a HRL method that relies on space abstraction, such as Feudal RL (FRL) (Dayan and Hinton, 1993), should allow RL scale to domains with a large number of slots. FRL divides a task spatially rather than temporally, decomposing the decisions in several steps and using different abstraction levels in each sub-decision. This framework is especially useful in RL tasks with large discrete action spaces, making it very attractive for large domain dialogue management.

In this paper, we introduce a Feudal Dialogue Policy which decomposes the decision in each turn into two steps. In a first step, the policy decides if it takes a slot independent or slot dependent action. Then, the state of each slot sub-policy is abstracted to account for features related to that slot,
and a primitive action is chosen from the previously selected subset. Our model does not require any modification of the reward function and the hierarchical architecture is fully specified by the structured database representation of the system (i.e. the ontology), requiring no additional design.

2 Background

Dialogue management can be cast as a continuous MDP (Young et al., 2013) composed of a continuous multivariate belief state space \( B \), a finite set of actions \( A \) and a reward function \( R(b_t, a_t) \).

At a given time \( t \), the agent observes the belief state \( b_t \in B \), executes an action \( a_t \in A \) and receives a reward \( r_t \in \mathbb{R} \) drawn from \( R(b_t, a_t) \). The action taken, \( a_t \), is decided by the policy, defined as the function \( \pi(b) = a \). For any policy \( \pi \) and \( b \in B \), the Q-value function can be defined as the expected (discounted) return \( R \), starting from state \( b \), taking action \( a \), and then following policy \( \pi \) until the end of the dialogue at time step \( T \):

\[
Q^\pi(b, a) = \mathbb{E}\{R | b_t = b, a_t = a\} \tag{1}
\]

where \( R = \sum_{\tau=1}^{T-1} \gamma^{\tau} r_\tau \) and \( \gamma \) is a discount factor, with \( 0 \leq \gamma \leq 1 \).

The objective of RL is to find an optimal policy \( \pi^* \), i.e. a policy that maximizes the expected return in each belief state. In Value-based algorithms, the optimal policy can be found by greedily taking the action which maximises \( Q^\pi(b, a) \).

In slot-filling SDSs the belief state space \( B \) is defined by the ontology, a structured representation of a database of entities that the user can retrieve by talking to the system. Each entity has a set of properties, referred to as slots \( S \), where each of the slots can take a value from the set \( \mathcal{V}_s \).

The belief state \( b \) is then defined as the concatenation of the probability distribution of each slot, plus a set of general features (e.g. the communication function used by the user, the database search method...) (Henderson et al., 2014a). The set \( A \) is defined as a set of summary actions, where the actions can be either slot dependent (e.g. request(food), confirm(area)...) or slot independent\(^1\) (e.g. hello(), inform(...)).

The belief space \( B \) is defined by the ontology, therefore belief states of different domains will have different shapes. In order to transfer knowledge between domains, Domain Independent Parametrisation (DIP) (Wang et al., 2015) proposes to abstract the belief state \( b \) into a fixed size representation. As each action is either slot independent or dependent on a slot \( s \), a feature function \( \phi_{dip}(b, s) \) can be defined, where \( s \in S \cup S_i \) and \( s_i \) stands for slot independent actions. Therefore, in order to compute the policy, \( Q(b, a) \) can be approximated as \( Q(\phi_{dip}(b, s), a) \), where \( s \) is the slot associated to action \( a \).

Wang et al. (2015) presents a handcrafted feature function \( \phi_{dip}(b, s) \). It includes the slot independent features of the belief state, a summarised representation of the joint belief state, and a summarised representation of the belief state of the slot \( s \). Section 4 gives a more detailed description of the \( \phi_{dip}(b, s) \) function used in this work.

3 Feudal dialogue management

FRL decomposes the policy decision \( \pi(b) = a \) in each turn into several sub-decisions, using different abstracted parts of the belief state in each sub-decision. The objective of a task oriented SDS is to fulfill the users goal, but as the goal is not observable for the SDS, the SDS needs to gather enough information to correctly fulfill it. Therefore, in each turn, the DM can decompose its decision in two steps: first, decide between taking an action in order to gather information about the user goal (information gathering actions) or taking an action to fulfill the user goal or a part of it (information providing actions) and second, select a (primitive) action to execute from the previously selected subset. In a slot-filling dialogue, the set of

\[ \sum^{T-1}_{\tau=1} \gamma^{\tau} r_\tau \]

Figure 1: Feudal dialogue architecture used in this work. The sub-policies surrounded by the dashed line have shared parameters. The simple lines show the data flow and the double lines the sub-policy decisions.

\[ \sum^{T-1}_{\tau=1} \gamma^{\tau} r_\tau \]
information gathering actions can be defined as the set of slot dependent actions, while the set of information providing actions can be defined as the remaining actions.

The architecture of the feudal policy proposed by this work is represented schematically in Figure 1. The (primitive) actions are divided between two subsets; slot independent actions \(A_i\) (e.g. hello(), inform()) and slot dependent actions \(A_d\) (e.g. request(), confirm())\(^2\). In addition, a set of master actions \(A_m = (a_{m}^{i}, a_{m}^{d})\) is defined, where \(a_{m}^{i}\) corresponds to taking an action from \(A_i\) and \(a_{m}^{d}\) to taking an action from \(A_d\). Then, a feature function \(\phi_s(b) = b_s\) is defined for each slot \(s \in S\), as well as a slot independent feature function \(\phi_i(b) = b_i\) and a master feature function \(\phi_m(b) = b_m\). These feature functions can be handcrafted (e.g. the DIP feature function introduced in section 2) or any function approximator can be used (e.g. neural networks trained jointly with the policy).

Finally, a master policy \(\pi_m(b_m) = a^m\), a slot independent policy \(\pi_i(b_i) = a^i\) and a set of slot specific policies \(\pi_s(b_s) = a^s\), one for each \(s \in S\), are defined, where \(a^m \in A_m\), \(a^i \in A_i\) and \(a^s \in A_s\). Contrary to other feudal policies, the slot specific sub-policies have shared parameters, in order to generalise between slots (following the idea used by Henderson et al. (2014b) for DST). The differences between the slots (size, value distribution...) are accounted by the feature function \(\phi_s(b)\). Therefore \(\pi_m(b_m)\) is defined as:

\[
\pi_m(b_m) = \arg\max_{a^m \in A_m} Q^m(h_m, a^m) \tag{2}
\]

If \(\pi_m(b_m) = a_{m}^{m}\), the sub-policy run is \(\pi_i\):

\[
\pi_i(b_i) = \arg\max_{a^i \in A_i} Q^{i}(b_i, a^i) \tag{3}
\]

Else, if \(\pi_m(b_m) = a_{m}^{d}\), \(\pi_d\) is selected. This policy runs each slot specific policy, \(\pi_s\), for all \(s \in S\), choosing the action-slot pair that maximises the Q function over all the slot sub-policies.

\[
\pi_d(b_s | \forall s \in S) = \arg\max_{a^d \in A_d, s \in S} Q^s(b_s, a^d) \tag{4}
\]

Then, the summary action \(a\) is constructed by joining \(a^d\) and \(s\) (e.g. if \(a^d=\text{request}()\) and \(s=\text{food}\), then the summary action will be \(\text{request(food)}\)). A pseudo-code of the Feudal Dialogue Policy algorithm is given in Appendix A.

\(^2\)Note that the actions of this set are composed just by the communication function of the slot dependent actions, thus reducing the number of actions compared to \(A\).

The implementation of the models can be obtained in www.pydial.org

<table>
<thead>
<tr>
<th>Domain</th>
<th>Code</th>
<th># constraint slots</th>
<th># requests</th>
<th># values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cambridge Restaurants</td>
<td>CR</td>
<td>3</td>
<td>9</td>
<td>268</td>
</tr>
<tr>
<td>San Francisco Restaurants</td>
<td>SFR</td>
<td>6</td>
<td>11</td>
<td>636</td>
</tr>
<tr>
<td>Laptops</td>
<td>LAP</td>
<td>11</td>
<td>21</td>
<td>257</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Env. 1</th>
<th>Env. 2</th>
<th>Env. 3</th>
<th>Env. 4</th>
<th>Env. 5</th>
<th>Env. 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>SER</td>
<td>0%</td>
<td>0%</td>
<td>15%</td>
<td>15%</td>
<td>15%</td>
<td>30%</td>
</tr>
<tr>
<td>Masks</td>
<td>on</td>
<td>off</td>
<td>on</td>
<td>off</td>
<td>on</td>
<td>on</td>
</tr>
<tr>
<td></td>
<td>Std.</td>
<td>Std.</td>
<td>Std.</td>
<td>Std.</td>
<td>Unf.</td>
<td>Std.</td>
</tr>
</tbody>
</table>

Table 1: Sumarised description of the domains and environments used in the experiments. Refer to (Casanueva et al., 2017) for a detailed description.

4 Experimental setup

The models used in the experiments have been implemented using the PyDial toolkit (Ultes et al., 2017)\(^3\) and evaluated on the PyDial benchmarking environment (Casanueva et al., 2017). This environment presents a set of tasks which span different size domains, different Semantic Error Rates (SER), and different configurations of action masks and user model parameters (Standard (Std.) or Unfriendly (Unf.)). Table 1 shows a summarised description of the tasks. The models developed in this paper are compared to the state-of-the-art RL algorithms and to the handcrafted policy presented in the benchmarks.

4.1 DIP-DQN baseline

An implementation of DIP based on Deep-Q Networks (DQN) (Mnih et al., 2013) is implemented as an additional baseline (Papangelis and Stylianou, 2017). This policy, named DIP-DQN, uses the same hyperparameters as the DQN implementation released in the PyDial benchmarks. A DIP feature function based in the description in (Wang et al., 2015) is used, \(\phi_{dip}(b, s) = \psi_0(b) \oplus \psi_j(b) \oplus \psi_d(b, s)\), where:

- \(\psi_0(b)\) accounts for general features of the belief state, such as the database search method.
- \(\psi_j(b)\) accounts for features of the joint belief state, such as the entropy of the joint belief.
- \(\psi_d(b, s)\) accounts for features of the marginal distribution of slot \(s\), such as the entropy of \(s\).

Appendix B shows a detailed description of the DIP features used in this work.

4.2 Feudal DQN policy

A Feudal policy based on the architecture described in sec. 3 is implemented, named FDQN. Each sub-policy is constructed by a DQN policy.

\(^3\)The implementation of the models can be obtained in www.pydial.org
717

Table 2: Success rate and reward for Feudal-DQN and DIP-DQN in the 18 benchmarking tasks, compared with the reward of the best performing algorithm in each task (Bnch.) and the handcrafted policy (Hdc.) presented in (Casanueva et al., 2017)

(Su et al., 2017). These policies have the same hyperparameters as the baseline DQN implementation, except for the two hidden layer sizes, which are reduced to 130 and 50 respectively. As feature functions, subsets of the DIP features are used:

$$\phi_m(b) = \phi_i(b) = \psi_0(b) + \psi_j(b)$$

$$\phi_s(b) = \psi_0(b) + \psi_j(b) + \psi_d(b, s) \forall s \in S$$

The original set of summary actions of the benchmarking environment, $\mathcal{A}$, has a size of $5 + 3 \times |S|$, where $|S|$ is the number of slots. This set is divided in two subsets: $\mathcal{A}_i$ of size 6 and $\mathcal{A}_d$ of size 4. Each sub-policy (including $\pi_m$) is trained with the same sparse reward signal used in the baselines, getting a reward of 20 if the dialogue is successful or 0 otherwise, minus the dialogue length.

5 Results

The results in the 18 tasks of the benchmarking environment after 4000 training dialogues are presented in Table 2. The same evaluation procedure of the benchmarks is used, presenting the mean over 10 different random seeds and testing every seed for 500 dialogues. The FDQN policy substantially outperforms every other other policy in all the environments except Env. 1. The performance increase is more considerable in the two largest domains (SFR and LAP), with gains up to 5 points in accumulated reward in the most challenging environments (e.g. Env. 4 LAP), compared to the best benchmarked RL policies (Bnch.). In addition, FDQN consistently outperforms the handcrafted policy (Hdc.) in environments 2 to 6, which traditional RL methods could not achieve. In Env. 1, however, the results for FDQN and DIP-DQN are rather low, specially for DIP-DQN. Surprisingly, the results in Env. 2, which only differs from Env. 1 in the absence of action masks (thus, in principle, is a more complex environment), outperform every other algorithm. Analysing the dialogues individually, we could observe that, in this environment, both policies are prone to “overfit” to an action $^5$. The performance of FDQN and DIP-DQN in Env. 4 is also better than in Env. 3, while the difference between these environments also lies in the masks. This suggests that an specific action mask design can be helpful for some algorithms, but can harm the performance of others. This is especially severe in the DIP-DQN case, which shows good performance in some challenging environments, but it is more unstable and prone to overfit than FDQN.

However, the main purpose of action masks is to reduce the number of dialogues needed to train a policy. Observing the learning curves shown in Figure 2, the FDQN model can learn a near-optimal policy in large domains in about 1500 dialogues, even if no additional reward is used, making the action masks unnecessary.

$^5$The model overestimates the value of an incorrect action, continuously repeating it until the user runs out of patience.
6 Conclusions and future work

We have presented a novel dialogue management architecture, based on Feudal RL, which substantially outperforms the previous state of the art in several dialogue environments. By defining a set of slot dependent policies with shared parameters, the model is able to learn a general way to act in slots, increasing its scalability to large domains.

Unlike other HRL methods applied to dialogue, no additional reward signals are needed and the hierarchical structure can be derived from a flat ontology, substantially reducing the design effort.

A promising approach would be to substitute the handcrafted feature functions used in this work by neural feature extractors trained jointly with the policy. This would avoid the need to design the feature functions and could be potentially extended to other modules of the SDS, making text-to-action learning tractable. In addition, a single model can be potentially used in different domains (Papangelis and Stylianou, 2017), and different feudal architectures could make larger action spaces tractable (e.g., adding a third sub-policy to deal with actions dependent on 2 slots).

Acknowledgments

This research was funded by the EPSRC grant EP/M018946/1 Open Domain Statistical Spoken Dialogue Systems.

References


### A Feudal Dialogue Policy algorithm

**Algorithm 1** Feudal Dialogue Policy

1: for each dialogue turn do
2:    observe $b$
3:    $b_m = \phi_m(b)$
4:    $a^m = \text{argmax } Q^m(b_m, a^m)$
5:    if $a^m = a^m_i$ then▷ drop to $\pi_i$
6:        $b_i = \phi_i(b)$
7:        $a = \text{argmax } Q_i(b_i, a^i)$
8:            $a^i \in A_i$
9:    else $a^m = a^m_i$ then▷ drop to $\pi_d$
10:        $b_s = \phi_s(b) \forall s \in S$
11:        $\text{slot, act } = \text{argmax } Q^s(b_s, a^d)$
12:        $s \in S, a^d \in A_d$
13:        $a = \text{join} (\text{slot, act})$
14:    end if
15:    execute $a$
16: end for

### B DIP features

This section gives a detailed description of the DIP feature functions $\phi_{dip}(b, s) = \psi_0(b) \oplus \psi_j(b) \oplus \psi_d(b, s)$ used in this work. The differences with the features used in (Wang et al., 2015) and (Papangelis and Stylianou, 2017) are the following:

- No priority or importance features are used.
- No Potential contribution to DB search features are used.
- The joint belief features $\psi_j(b)$ are extended to account for large-domain aspects.

#### Table 3: List of features composing the DIP features.

<table>
<thead>
<tr>
<th>Feature function</th>
<th>Feature description</th>
<th>Feature size</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\psi_0(b)$</td>
<td>last user dialogue act (bin) *</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>DB search method (bin) *</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td># of requested slots (bin)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>offer happened *</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>last action was Inform no venue *</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>normalised # of slots (1/# of slots)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>normalised avg. slot length (1/avg. # of values)</td>
<td>1</td>
</tr>
<tr>
<td>$\psi_j(b)$</td>
<td>prob. of the top 3 values of $b_j$</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>prob. of <em>NONE</em> value of $b_j$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>entropy of $b_j$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>diff. between top and 2nd value probs. (bin)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td># of slots with top value not <em>NONE</em> (bin)</td>
<td>5</td>
</tr>
<tr>
<td>$\psi_d(b, s)$</td>
<td>prob. of the top 3 values of $s$</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>prob. of <em>NONE</em> value of $s$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>diff. between top and 2nd value probs. (bin)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>entropy of $s$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td># of values of $s$ with prob. $&gt; 0$ (bin)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>normalised slot length (1/# of values)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>slot length (bin)</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>entropy of the distr. of values of $s$ in the DB</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3: List of features composing the DIP features. The tag (bin) denotes that a binary encoding is used for this feature. Some of the joint features $\psi_j(b)$ are extracted from the joint belief $b_j$, computed as the Cartesian product of the beliefs of the individual slots. * denotes that these features exist in the original belief state $b$. 

719