# A multi-agent collaborative algorithm for task-oriented dialogue systems

**Jingtao Sun** ( sun2651@qq.com )
Xi’an University of Posts and Telecommunications
**Jiayin Kou**
Xi’an University of Posts and Telecommunications

**Research Article**

**Keywords:** Task-oriented dialogue systems, Reinforcement learning, Multi-agent collaboration, Integrated learning

**Posted Date:** March 24th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-2722056/v1

# Abstract

In recent years, reinforcement learning has been successfully applied to dialogue systems. However, in the face of task-oriented dialogue systems where policies are difficult to optimize, states are difficult to track, and tasks are multiple and compound, task-oriented dialogue systems based on reinforcement learning have problems such as poor collaboration, non-unique learning goals, and non-stationarity due to the lack of collaboration among agents. In this paper, we propose a multi-agent cooperative dialogue (MACD) algorithm for task-oriented dialogue systems, in which, for the information interaction between multi-agent in task-oriented dialogue systems, a deep neural network approach is used to integrate the observations of multiple single agents and obtain joint observations to achieve information sharing among single agents, to solve the non-stationarity caused by the lack of joint information of multi-agent. For multi-agent policy learning task-oriented dialogue systems, the multi-agent deep deterministic policy gradient (MADDPG) architecture is applied to the policy selection of task-oriented dialogue systems to solve the problem of lack of joint policy learning of multi-agent in task-oriented dialogue systems; the observation integration of single agents and multi-agent policy learning are effectively combined to solve the problem of poor multi-agent collaboration in task-oriented dialogue systems. By verifying and analyzing reinforcement learning algorithms such as MACD, REINFORCE, DQN, and QMIX in the MultiWOZ corpus, the experimental results show that the algorithms effectively improve the success rate of multi-agent working together to complete dialogue tasks, reduce the number of invalid dialogues in dialogue turns, and outperform common reinforcement learning algorithms in terms of agents' information interaction and joint policy learning in the composite task dialogue scenario.

# 1. Introduction

Dialogue systems, as an important means of human-computer interaction, are receiving increasing attention from academia and industry. From an application point of view, they can be broadly classified into open-domain dialogue systems and task-oriented dialogue systems. Currently, most of the open-domain dialogue systems on the market are in the form of companion robots (Li Q et al. 2022; Zhou L et al. 2020), and they are not as widely used in life as task-oriented dialogue systems. Commonly seen "Tmall Genie", "Apple Siri", and service robots in banks and supermarkets are all task-oriented dialogue systems (Sakata W et al. 2019; Paweł Budzianowski and Ivan Vulić 2019). These applications allow people to complete simple tasks such as querying, controlling, and operating through simple command language such as "Tmall Genie, please turn on the TV", but once they encounter a complex task, they are unable to effectively execute the command and get embarrassing responses such as "Please say it again, I didn't hear you! " These results are mostly because the dialogue is beyond the scope of the application's pre-made corpus of rules. The AI customer service such as "intelligent assistants" and "voice assistants" in major business systems are also representations of task-oriented dialogue systems, which minimize the workload of manual services and are on call around the clock, greatly reducing the labor costs of business operations. However, most of these systems can only solve basic problems and are unable to handle the diverse needs of users, identify their true intentions, or apply flexibly in a variety of business scenarios. These practical problems are common to current task-oriented dialogue systems.

To solve these practical problems and make task-oriented dialogue systems more relevant to the real world, researchers need to optimize the algorithmic architecture of task-oriented dialogue systems to improve the performance indicators of the systems. To improve the performance of task-oriented dialogue systems, researchers need to optimize the algorithm architecture of task-oriented dialogue systems. In the face of the difficulties in handling complex tasks, recognizing user intentions, tracking dialogue state, exceeding the scope of the corpus, and in the poor concordance of dialogue agents, the task-oriented dialogue system needs to be improved by using a variety of techniques.

Most of the existing task-oriented dialogue systems use deep learning methods to solve many of the key problems plaguing the systems by leveraging their powerful feature extraction and learning capabilities, such as building NBT models for dialogue state tracking using neural networks (Mrkšić N et al. 2016) and using Transformer XL neural network architecture to obtain contextual long-term dependencies as a solution to contextual fragmentation (Dai Z et al. 2019). However, deep learning, as a supervised machine learning method, requires a large number of pre-labeled samples for training, but a large library of pre-labeled samples increases the computation of the deep neural network and increases the training time of the system while improving the accuracy of the task; and in the task dialogue, pre-labeled samples of text are instead harder to obtain, and it is impossible to build a large library of pre-labeled samples, which will directly affect the This will directly affect the accuracy rate of task completion. To address these issues in the field of human-computer dialogue, the researcher wishes to introduce a new approach. Reinforcement learning is an unsupervised learning method that learns by interacting with the environment through trial and error to generate samples, which requires a smaller number of samples than deep learning and reduces the amount of model computation. Reinforcement learning can also solve some of the problems associated with deep learning methods in task-oriented dialogue systems, e.g., the use of DQN for dialogue policy learning, which solves the problem of the inability of an agent to explore learning on its own (Lipton Z et al. 2018), and the use of artificially set reward functions in the reinforcement learning environment, which makes dialogue content more informative, coherent and easy to answer (Li J et al. 2016). In task-oriented dialogue systems, reinforcement learning faces the problem of the irrational state of the agents, which makes it difficult to maximize the global reward, making it difficult to achieve optimal dialogue policy learning; the existence of complex game relationships between agents, which will lead to multiple agents working together to complete a predetermined dialogue task if there is a non-cooperative game relationship between agents in task-oriented dialogue systems. In task-oriented dialogue systems, the existence of non-cooperative game relations among agents will lead to a series of new problems, such as the lack of collaboration among multiple agents working together to complete a predefined dialogue task.

In response to the new problems associated with reinforcement learning methods in task-oriented dialogue systems, this paper investigates three main aspects: 1) Observation value integration. The initial state of each part of the environment is obtained through the initial observation of a single agent, and the initial state is processed by noise reduction using a neural network model to obtain the observed value of each part. The multilayer neural network model is used to integrate the partial observation values of multiple single agents to obtain joint observation values, which enables the joint sharing of information among single agents and solves the problem of non-stationarity among agents; 2) Joint policy learning of multiple agents. The MADDPG framework of "centralized training, distributed execution" is used to lay out the policy learning of multiple agents in the system, and a simple extension of the Actor-Critic policy gradient method is used for the policy learning of a single agent to achieve the joint policy learning of multiple agents in a task-oriented dialogue system, solving the problem of single agents learning policies independently; 3) MACD construction. By effectively combining the observation integration method of a single agent with the multi-agent policy learning method, the joint observation is used to centralize policy learning of multiple agents, and the information of other agents enhances the policy learning of a single agent, strengthening the collaboration between agents and enabling the task-oriented dialogue system to complete the dialogue task more efficiently. Comparative experiments on the dialogue corpus MultiWOZ show that the MACD can construct more effective multi-agent collaborative training policies and achieve higher task success rates in composite dialogue scenarios by interacting with each other and with the baseline policy, effectively reducing the number of dialogues with invalid dialogue.

The paper is organized as follows: Section 2 discusses the work related to the research methodology of this paper. Section 3 explores in detail how the multi-agent collaboration algorithm can enhance the collaborative nature of task-oriented dialogue systems in terms of building dialogue multi-agents, integration of observations, and MACD design. Section 4 presents the experiments and results analysis. Section 5 concludes and gives an outlook on the work of this paper.

## 2 Related Work

With the rapid development of deep learning, advances in dialogue systems have been greatly facilitated. In recent years, the end-to-end model has become one of the common base models used by researchers to solve dialogue system problems, which encapsulates the three modules of natural language understanding (NLU), natural language generation (NLG), and dialogue management (DM) as a way to reduce the workload of dialogue systems in processing text in multiple steps. With the help of this model (Ham D et al. 2020r et al. 2018), researchers have continuously optimized the performance of dialogue systems in terms of language comprehension, variety of generated utterances, and accuracy of response utterances. In terms of response generation, existing dialogue systems ensure a certain level of fluency and coherence, but still suffer from insufficient information richness. For this reason, Lin X et al. (2020) proposed a knowledge copy mechanism that uses a knowledge-aware pointer network to copy words from external knowledge according to knowledge attention distribution. The joint neural conversation model built in this study integrates recurrent knowledge-interaction and knowledge copy (KIC) and performs well in generating informative responses. Neural network models have emerged as one of the most important approaches to generating dialogue responses. However, they always tend to generate the most common and generic responses in the corpus. To address this issue, Wenchao Du and Alan W Black (2019) proposed a boosting-based iterative training process and integration method to combine different training and decoding paradigms to build a base dialogue model, which in turn improves the diversity and relevance of the responses generated by this model. In task-oriented dialogue systems, Li X et al. (2017) proposed a novel end-to-end learning framework for such dialogue systems, addressing the drawback of separate training of modular tasks in task-oriented dialogue systems. In Li Z et al. (2019), a task-oriented dialogue system based on document dialogue is constructed and a two-pass decoder (Deliberation Decoder) is designed to address the problems of poor contextual coherence and low knowledge correctness in generating dialogue responses when given document content.

Most of these studies focus on the use of deep learning methods to solve a range of problems in dialogue systems, but deep learning methods are subject to bottlenecks such as excessive computing power and inefficiency in practical applications. Therefore, the current research attempts to introduce reinforcement learning methods to optimize dialogue systems by maximizing the training of the agents with the help of the reward mechanism set in the reinforcement learning environment. Jaques N et al. (2020) proposed the use of offline reinforcement learning to identify implicit conversational cues and embed them in a variety of reward functions, addressing the problem that policy training can fail in an offline environment due to a lack of exploration capabilities, as well as making overly optimistic estimates of future reward rewards. Takanobu R et al. (2019) proposed a new algorithm based on adversarial inverse reinforcement learning to guide dialogue policy learning, addressing the problem of joint reward estimation in multi-domain task-oriented dialogue. Among others, the framework of adversarial inverse reinforcement learning can also be utilized, and Li Z et al. (2019) proposed a dialogue generation reward model that can provide more accurate and precise reward signals, resulting in higher-quality responses. The applications of reinforcement learning in dialogue systems are diverse, and Huang X et al. (2021) proposed a novel reinforcement learning network that solves the problem of detecting emotion for each utterance by tracking information about the gradual emotion change of each utterance during the dialogue and using this information. In addition to its direct application to dialogue systems, reinforcement learning can also improve existing dialogue system models. Le A C (2021) used an encoder-decoder integrated with an attention mechanism and introduced reinforcement learning to improve this model, solving the problem of less natural response utterances due to a weak degree of contextual coherence in dialogue content. In task-oriented dialogue systems, more researchers have used reinforcement learning as a basis for building multi-agent collaborative dialogue environments as a way to strengthen the connections between agents. Papangelis A et al. (2019) trained natural language understanding (NLU) and generation (NLG) networks for each agent and let the agents interact online, allowing each agent to learn how to operate optimally in an environment with multiple sources of uncertainty. Das A et al. (2017) used deep reinforcement learning to build an end-to-end model to train the agents on policy, enabling two agents to communicate using the grounded language in the context of completing a collaborative visual dialogue task scenario.

While the introduction of reinforcement learning has led to advances in system training efficiency and agent collaboration in dialogue systems, it has also brought about new problems due to agent policy training. To address the problem that reinforcement learning focuses on the training of the target agent's policy while neglecting the training of the relative agent's policy, Zhang Z et al. (2020) proposed an opposite behavior-aware framework that evaluates the opposite agent's policy based on the opposite agent's behavior and uses this estimation as part of the target policy to improve the target agent's policy. Wang H et al. (2021) proposed to decompose the dialogue actions of the centralized agents in a dialogue system, reducing the size of the operation space of each agent and solving the non-stationary problem of agents' policy evolution caused by dynamic changes in the environment.

In addition to these two approaches, some researchers also introduce other methods to solve problems that arise in dialogue systems. Kim H et al. (2020) proposed to assign different roles to multi-agent in a dialogue system to improve the consistency of roles and solve the problem of dialogue contradiction. Jia Q et al. (2020) proposed converting dialogue histories into threads and using dialogue dependencies to achieve multi-round responses, solving the problem of low dialogue counts. Lin Z et al. (2020) proposed Minimalist Transfer Learning (MinTL) to simplify the system design process of task-oriented dialogue systems, effectively alleviating the over-dependency on annotated data.

This paper proposes a multi-agent cooperative dialogue (MACD) algorithm, which combines the MADDPG "centralized training, distributed execution" multi-agent policy training framework with the simple extended Actor-Critic single-agent policy gradient method, and uses a deep neural network integration method to obtain joint observations, to enhance the problem of lack of joint information in centralized policy training for a single agent. The algorithm improves the task success rate of task-oriented dialogue systems and reduces the number of invalid dialogues in dialogue turns.

## 3 Construction Of Multi-agent Collaborative Algorithms

In this paper, a multi-agent cooperative dialogue (MACD) algorithm is proposed for task-oriented dialogue systems. This section firstly introduces the construction idea and basic framework of the algorithm secondly details the construction method of dialogue multi-agent and the use of the observation integration method to solve the dialogue multi-agent non-stationarity problem, and finally describes the design steps of the MACD in the form of pseudo-code.

# 3.1 Methodological ideas and basic framework

The analysis of task-oriented dialogue systems to achieve the introduction of reinforcement learning methods to optimize system performance needs to overcome the following difficulties: 1) The current task-oriented dialogue systems mainly use deep learning methods, there are difficulties such as difficulty to obtain dialogue text and difficult to match the common knowledge base; 2) In the dialogue task processing, reinforcement learning methods to achieve more stringent response constraints, generating more accurate response responses and other difficulties; (3) The difficulties of non-stationarity and low collaboration among multi-agent based on reinforcement learning methods. Therefore, this paper introduces a multi-agent reinforcement learning method into the task-oriented dialogue system and designs a novel algorithm to solve the above-mentioned difficulties.

Firstly, the approach of establishing dialogue with multiple agents is introduced into the task-oriented dialogue system. Given that dialogue is an interactive act between two or more people, single agents do not meet the needs of dialogue tasks well, this paper proposes to set up two dialogue agents with role identities for dialogue interaction acts.

Secondly, a deep neural network integration method is proposed. Given the non-stationarity problem caused by the lack of joint information between the dialogue multiple agents, this paper proposes to integrate the observations of each dialogue multiple agent for the environment to obtain a shared joint observation, to strengthen the lack of joint information between the dialogue multiple agents.

Finally, a new multi-agent collaborative algorithm is proposed. Given the low success rate of completing predefined tasks and the high number of invalid dialogues in task-oriented dialogue systems. In this paper, we propose to apply the multi-agent policy training architecture in the MADDPG and the single-agent policy gradient method in the Actor-Critic to effectively combine the two to build a dialogue policy learning framework for the new algorithm, to improve the dialogue multi-agent synergy and solve the above problems to a certain extent.

By describing the above design ideas and basic architecture, the MACD proposed in this paper is implemented, which effectively improves the main performance indicators of the task-oriented dialogue system and also achieves good dialogue effects in the interaction with human dialogue. The basic framework of the approach proposed in this paper is shown in Fig. 1.

## 3.2 Construction of dialogue multiple agents

Dialogue is a human activity in real life. To build dialogue systems, an artificial agent is used to simulating human dialogue behavior. In a task-oriented dialogue system, two dialogue agents are constructed to simulate humans to complete a dialogue task. First of all, the two agents are given specific roles in order to substitute realistic situations: the Requesterand the Service provider. Dialogue tasks are initiated by , andrespond by completing them together. For example, if a requester's objective is to find the address, postcode, and telephone number of a hospital gastroenterology department, the service provider will query the database and respond. The target task base (Goal) required by, and the database of response target tasks (DataBase) required by are provided by the corpus MultiWOZ, as shown in Fig. 2. There are certain constraints, onlyknows the target task and onlyhas access to the database, and the success of the task is calculated by using a multi-turn dialogue model from which the percentage of valid information is calculated, rather than a one-question-one-answer dialogue model of a question-and-answer dialogue system.

The agent is an important element of reinforcement learning, single agents follow Markov decision processes (MDPs) in reinforcement learning, and there are game relationships between multiple agents that are not present in single agents, the MDPs are extended by the corresponding number of agents, and after the extension both follow a partially observable Markov game together (Littman M L 1994). A framework of reinforcement learning is used to build two dialogue agents, which follow a Markov game defined by a set of states$S = (s^W, s^P)$, a set of actions$A = (a^W, a^P)$and a set of observations$O = (o^W, o^P)$describing the configuration of the two agents. To select an action, uses the random policy$\mu_\alpha : o^W \times a^W \rightarrow [0, 1]$ anduses the random policy$\pi_\beta : o^P \times a^P \rightarrow [0, 1]$. Both use the same state transfer functionto generate the respective next state:$s^W \times a^W \rightarrow s'^W$,$s^P \times a^P \rightarrow s'^P$. Both dialogue agents should receive a joint reward$r^G$for environmental feedback during training,$r^G$is only counted at the end of a round of dialogs, Both Policy$\mu$and Policy$\pi$objectives aim to maximize the accumulation of global rewards$E[\sum_t \gamma^t (r_t^G)]$, whereγis the discount factor andis the time horizon. The two dialogue agents then share joint information in a dialogue environment built by reinforcement learning and train their respective dialogue policies by continuous trial and error in an algorithmic framework.

## 3.3 Integration of observations

Currently, in multi-agent systems, multiple agents are very vulnerable to non-stationarity problems (Georgios Papoudakis et al. 2019), where the agents can only obtain local observations and not the overall state of the environment. This is because every single agent not only faces a changing environment but may also be affected by the changes and adaptation policies of other agents. If agents could share information (e.g. observations, intentions, or experiences) from other agents to stabilize learning, by communicating interactively through the shared information, agents would better understand the environment (or other agents) to be able to coordinate their behavior with each other.

This method solves the non-stationarity problem of conversational multi-agents by sharing information about their observationsand establishing a perfect information game-like relationship between them. First, the respective initial states$s^W$and$s^P$are obtained byandduring their initial observation of the environment. The initial states$s^W$and$s^P$ are pre-processed for data noise using a multilayer neural network Sequential model, which acts as a container that encapsulates the structure of the neural network with only one set of inputs and one set of outputs. The initial states$s^W$,$s^P$of andare used as inputs to the neural network and the outputs are the respective observations$O^W$,$O^P$of and , as shown in Eqs. (1), (2); Secondly,andshare information about each other's observations and form a collection of observations$\{O^W, O^P\}$, in a multi-agent system, policies get better training is based on global awareness, and joint shared information needs to be considered to enhance the training of jointly policies. The method uses the set of observations $\{O^W, O^P\}$to enhance the global information that is missing during policy training. In the DQN (Volodymyr Mnih et al. 2015) a deep neural network model is used to approximate a state value function value, and this same approach is used to obtain a joint state value function value, with the difference that the global set of observations $\{O^W, O^P\}$is processed by a multilayer neural network Sequential model to obtain a joint observation value$Q^G$, as shown in Eqs. (3), $Q^G$is used in section 3.4 for centralized training of policies.

$$O^R = f(Z_s^W(s^W))$$

1

$$O^P = f(Z_s^P(s^P))$$

2

$$Q^G(s) = Z_G([O^W; O^P])$$

3

where $Z(\cdot)$ denotes any neural network unit and $f(\cdot)$ denotes the activation function.

## 3.4 Multi-agent cooperation dialogue algorithm

To address the problem of low dialogue multi-agent concordance in task-oriented dialogue systems, this paper proposes a new multi-agent collaborative algorithm in this setting. The corpus MultiWOZ provides multi-domain, complex conversational tasks, and the dialogue multiple agents are built in section 3.2 using a framework of reinforcement learning, where the training of dialogue policies for the agents is almost impossible to explore and learn from scratch. When moving on to the real dialogue policies training of the agents, the dialogue policies need to be pre-trained by the corpus and then the pre-trained policies are improved using the proposed new algorithm. After the above preparatory work is completed, the real training of the new algorithm of the agents will be entered.

First, a new policy training method is proposed for the dialogue policy training of dialogue multiple agents in Section 3.2. The method extends the reinforcement learning Actor-Critic policy gradient approach (Bahdanau D et al. 2016). The difference is that both and are treated as Actors and the Critic in this algorithm is treated as a Global Critic. The actor's score is based on the change in the state caused by the action, and the actor modifies the probability of the selected action based on the Critic's score. In the new algorithm and select actions based on $\mu_\alpha$and $\pi_\beta$, respectively, and Global Critic judges the scores of the actions based on the changes in the joint observations caused by the actions of and . and modify $\mu_\alpha$and $\pi_\beta$, respectively, according to the Global Critic scores.

Secondly, for the W and P and Global Critic overall policy training frameworks, the MACD draws on the MADDPG (Lowe R et al. 2017) policy training framework "Centralized Training, Distributed Execution" (CTDE) (Bernstein D S et al. 2002), with a simple modification of the MADDPG policy training method to improve collaboration between dialogue multiple agents.

The MADDPG addresses the non-stationarity problem among multiple agents by only using additional information from other agents' policies to enhance the connection between them when training Critic centrally, and does not make good use of the joint information to solve the problem effectively. The MACD uses the shared joint observation information described in Section 3.3 to solve the non-stationarity problem for dialogue multiple agents and uses the integrated joint observations to better train Global Critic during centralized training. The centralized training in this method is the training of the Global Critic policy, and the distributed execution reflects the individualized differences in the selection of actions thatandcontinue to maintain after training their respective actor policies.

Within the algorithmic framework of CTDE, the Global Critic policy is trained centrally as shown in Eqs. (4).

$L^G(\theta) = [y - Q^G_\theta(s)]^2$ , $y = r^G + \gamma Q^G_{\theta'}(s')$ (4)

Where $L^G$is the loss function of Global Critic and the training objective is to minimize $L^G$. The goal is to minimize the squared error between the target$y = r^G + \gamma Q^G_{\theta'}(s')$and the estimate $Q^G_\theta(s)$, using the time-discrepancy (TD) method of reinforcement learning.$Q_\theta$is parameterised by$\theta$,$\theta'$is the weight of the target network, $Q^G_\theta(s) = Z_G([O^W_s; O^P_s])$, $Q^G_{\theta'}(s') = Z_G([O^W_{s'}; O^P_{s'}])$from Section 3.3. $\gamma$is the loss factor and $r^G$is the joint reward obtained from the environmental feedback.

Distributed execution of the trained versus policy, by using the log-likelihood ratio trick, with gradients generated as shown in Eqs. (5), (6).

$$\nabla_\alpha J_\mu(\alpha) = \nabla_\alpha \log \mu_\alpha(a^W|s^W)[A^G(s)]$$

5

$$\nabla_\beta J_\pi(\beta) = \nabla_\beta \log \pi_\beta(a^P|s^P)[A^G(s)]$$

6

Where $A^G(s) = r^G + \gamma Q^G(s') - Q^G(s)$is a dominance function calculated via Global Critic that evaluates the new state $s'$and the current state to determine whether the dialogue is becoming better or worse than expected. The policy $\mu_\alpha$is parameterised by $\alpha$ and the policy $\pi_\beta$is parameterised by$\beta$. In summary, a short script of the MACD is shown in Algorithm 1.

## 4 Experiments

This section compares the performance differences between the MACD proposed in this paper and the REINFORCE, DQN, PPO, QMIX, and IterDPL through simulation experiments as a way to verify the effectiveness and correctness of the MACD.

# 4.1 Experimental design and data

The experimental code was written using the Pycharm software platform and the parameters of the task-oriented dialogue system were set as follows: the policies of the dialogue agents (Requesterand the Service provider) were implemented using multilayer perceptron networks (MLPs) with action spaces of 100 and 200 respectively; the neural network units used in the Global Critic joint state value network, were implemented using MLPs. The activation functions were all Relu for MLPs, and the optimization algorithm used RMSprop with a batch size of 32; reinforcement learning training with learning rates of 1e-4 and 5e-5 for $\mu_\alpha$and $\pi_\beta$, respectively, and 3e-5 for the joint critic Global Critic policy, with a discount factorγof 0.9, and the target network updated every 300 training iteration once; for the reward design, the artificially set penalty of -4 for empty actions and − 2 for other types of penalties. The sub-goal completion reward is 5, if triggered, mission success and goal reward are 20, otherwise − 5.

To verify the performance differences of multiple methods in a task-oriented dialogue context, the task-oriented dialogue corpus MultiWOZ (Paweł Budzianowski et al. 2018) was selected for experimental evaluation, containing more than 10,000 annotated dialogues across 8 domains, which is at least one order of magnitude larger than all previous task-oriented dialogue corpora with annotations. To further enhance the reproducibility of the results, the corpus was randomly divided into a training set, a test set, and a development set, with a pre-training set of 1000, a training set of 300, a test set, and a development set of 1000. As all the dialogues were coherent, some of them did not follow the task description. Therefore, to make the comparisons more accurate, the validation and test sets only contain a fully successful dialogue corpus.

## 4.2 Baselines

In the MultiWOZ corpus test environment, experiments were conducted to compare and analyze the performance differences of different algorithms as follows.

**REINFORCE** The REINFORCE with two layers of fully connected policy networks (Ronald J Williams 1992).

**DQN** Traditional DQN (Volodymyr Mnih et al. 2015) with a 2-layer fully connected network for Q-functions.

**PPO** Proximity Policy Optimisation (John Schulman et al. 2017), a policy-based RL using a constant-limit mechanism.

**IterDPL** Iterative Dialog Policy Learning (Bing Liu and Ian Lane 2017) uses a single RL training iteration to update two agents to reduce the risk of non-stationarity when training two agents jointly.

**QMIX** Estimates the joint values as a complex non-linear combination of each agent's value conditional on local observations only, via a network (Rashid T et al. 2018).

## 4.3 Metrics

For the measurement of the experimental results, four measures are $\mathrm{Re}\,call = \frac{|Y_A \cap Y_B|}{|Y_B|}$$\mathrm{Re}\,call = \frac{|Y_A \cap Y_B|}{|Y_B|}$ selected in this paper: the number of dialogues (Turn), the F1 score (Inform F1), the match rate (Match Rate), and the task success rate (Success). The Requester discourse and the subsequent Service provider discourse are considered as several dialogues and Turn is the average number of dialogues in a round. In a task-oriented towards a collaborative goal, completing the task with lower-cost conversational rounds increases requester satisfaction. Two other metrics, Inform F1 and Match Rate, are also utilized to estimate Success. Inform F1 is used to assess whether the requester has been successfully notified of all slots in the requested entity, and is used because it considers both Precision and Recall, as detailed in Eqs. (7). Match Rate is used to assess whether the booked entity matches the target in all domains, with a score of 1 for the domain when and only when its entity is successfully booked.

$$\mathrm{Pr}\,ecision = \frac{|Y_A \cap Y_B|}{|Y_A|} \quad \mathrm{Pr}\,ecision = \frac{|Y_A \cap Y_B|}{|Y_A|}$$

(7)

$$F1 = 2 \times \frac{\mathrm{Re}\,call \times \mathrm{Pr}\,ecision}{\mathrm{Re}\,call + \mathrm{Pr}\,ecision} \quad F1 = 2 \times \frac{\mathrm{Re}\,call \times \mathrm{Pr}\,ecision}{\mathrm{Re}\,call + \mathrm{Pr}\,ecision}$$

Where: $Y_A$and$Y_B$refer to the set of non-stop words in the generated session and background knowledge.

## 4.4 Analysis of collaborative dialogue

The experiments used reinforcement learning methods to build a framework for a task-oriented dialogue system, in which the Requester , the Service providerand the Global Critic were used as the base network to interact with the Environment for training.

First, the MACD was trained separately with various comparison algorithms such as REINFORCE, DQN, PPO, QMIX, and IterDPL in a task-oriented dialogue system environment to obtain Turn, Inform F1, Match Rate and Success of the Requester-Service provider interaction, and the experimental results are shown in Table 1.

Table 1
Performance of the interaction between the Requester and the Service provider

| to | #Turn | Inform F1 | Match Rate | Success |
|---|---|---|---|---|
| REINFORCE | 9.470 | 82.12 | 66.13 | 57.20 |
| DQN | 10.110 | 76.21 | 62.34 | 52.40 |
| PPO | 9.250 | 82.23 | 70.02 | 59.96 |
| IterDPL | 11.560 | 74.10 | 70.25 | 63.60 |
| QMIX | 8.720 | 73.55 | 80.20 | 68.40 |
| MACD | **8.286** | **75.80** | **81.20** | **73.60** |

As can be seen from Table 1, the difference between the IterDPL and MACD is the largest in terms of the "Turn" metric, with the MACD decreasing by 3.274 compared to the IterDPL, while the MACD is the least different from the QMIX, with the MACD decreasing by only 0.434 compared to the QMIX. In comparison with other algorithms, the MACD has the lowest Turn metric. The MACD and QMIX performed better than the other four algorithms in this metric, mainly since the IterDPL, REINFORCE, DQN, and PPO are all single-agent RL algorithms. In a multi-agent environment, they do not create information connections between each other and have poorer synergy than the QMIX and MACD, which have this advantage because these two algorithms have consciously used information between agents to enhance synergy.

Under the "Inform F1" and "Match Rate" metrics, the Inform F1 of the PPO (82.23%) was slightly higher than most of the comparison algorithms, but the Match Rate was only 70.02%, while the Match Rate of the MACD (81.20%) was much higher than most of the comparison algorithms, but the Inform F1 (75.80%) decreased less. The Match Rate for the MACD (81.20%) was much higher than most of the comparison algorithms, but the Inform F1 (75.80%) decreased to a lesser extent, while the Inform F1 and the Match Rate were used together to estimate the Success. The MACD showed a 13.64% increase compared to the PPO for the Success metric, where it was found that the Match Rate had a greater impact on the Success compared to the Inform F1. By comparing the values of the other algorithms for these two metrics, it can be seen that the MACD outperforms the other algorithms in terms of task completion performance due to the significant improvement in Inform F1 and Match Rate.

Secondly, to further analyze the variety of different algorithms in the Success metrics, a training set of 300 was selected to compare the Success of MACD with five algorithms, including REINFORCE, DQN, PPO, QMIX, and IterDPL, and the specific experimental results are shown in Fig. 3.

As can be seen in Fig. 3, there is little difference in task completion performance between the REINFORCE and the MACD for the first 100 sets of iterations. At 150 training sets, the MACD achieves the highest Success compared to the other comparison algorithms at 53.3%, which is 1.3%, 2.8%, 5.1%, 7%, and 7.9% higher than the IterDPL, REINFORCE, DQN, QMIX and PPO respectively. As the number of iteration sets increases, the MACD has the advantage of being more collaborative between agents in completing dialogue tasks and improving Success compared to other comparison algorithms. The MACD achieved the highest Success (73.60%) when the number of training sets was 300, with the greatest variability compared to the other comparison algorithms. At the highest Success, the Turn was the lowest among the compared algorithms (8.286). The MACD shares joint information to enhance the synergy between dialogue agents, avoiding some lengthy interactions to complete the dialogue task more efficiently. Therefore, the MACD proposed in this paper is more helpful for dialogue multiple agents to complete the dialogue task collaboratively.

To further validate the performance of the MACD in a multi-domain dialogue environment, the Requester and the Service provider are confronted with a composite dialogue task. The MultiWOZ corpus was divided into three different domains (2, 4, and 6) and the five comparison algorithms mentioned above were still used to obtain the Turn, Success Reward of the Requester, and the Service provider interaction, and the experimental results are shown in Table 2.

Table 2
Performance of the Requester and the Service provider interacting between different domains

| | | **domains = 2** | | | **domains = 4** | | | **domains = 6** | |
|---|---|---|---|---|---|---|---|---|---|
| Method | Turn | Reward | Succ. | Turn | Reward | Succ. | Turn | Reward | Succ. |
| REINFORCE | 6.62 | 52.30 | 0.73 | 8.20 | 35.40 | 0.70 | 10.14 | 31.20 | 0.52 |
| DQN | 13.03 | -3.46 | 0.69 | 16.40 | -15.20 | 0.32 | 18.74 | -31.23 | 0.12 |
| PPO | 7.42 | 49.54 | 0.75 | 10.23 | 37.12 | 0.69 | 14.52 | 23.45 | 0.49 |
| IterDPL | 9.65 | 52.13 | 0.86 | 9.56 | 35.62 | 0.81 | 13.55 | 25.12 | 0.79 |
| QMIX | 7.32 | 48.26 | 0.87 | 10.12 | 42.20 | 0.83 | 11.20 | 20.14 | 0.82 |
| MACD | **6.30** | **68.20** | **0.90** | **8.45** | **54.12** | **0.89** | **10.56** | **40.98** | **0.85** |

Table 3
Chi-squared tests for different algorithms

| Index | Levene Statistics | Significance |
|---|---|---|
| Turn | 0.901 | 0.511 |
| Reward | 0.062 | 0.997 |
| Succ. | 4.231 | 0.019 |

To test the significance of the effect of different algorithms on the dialogue multiple agents collaborating to complete the dialogue task, this paper conducted ANOVA analysis on the data in Table 2. The results of the chi-square test for the different algorithms are shown in Table 3, and the results of the ANOVA test for the different algorithms are shown in Table 4.

Table 4
ANOVA analysis of different algorithms

| Index | F value | Significance |
|---|---|---|
| Turn | 4.166 | 0.021 |
| Reward | 10.015 | 0.001 |
| Succ. | 5.400 | 0.008 |

As can be seen from Table 3, the significance of all the indicators except Succ. was greater than 0.05, 0.511, and 0.997 respectively, indicating that the variance of the analyzed data was chi-square, and therefore, a one-way ANOVA analysis could be applied to the data in Table 2. As can be seen from Table 4, the significance of Turn, Reward, and Succ. are all less than 0.05, 0.021, 0.001, and 0.008 respectively, indicating that there are significant differences between the six different algorithms used in this experiment, i.e. the different algorithms have an impact on the dialogue multiple agents working together to complete the dialogue task in a task-oriented dialogue system.

Table 5
Chi-squared tests for different domains

| Index | Levene Statistics | Significance |
|---|---|---|
| Turn | 0.179 | 0.838 |
| Reward | 0.004 | 0.996 |
| Succ. | 2.376 | 0.127 |

Again, to test the significance of the effect of different domains on the dialogue multiple agents working together to complete the dialogue task, ANOVA analysis was conducted on the data in Table 2. The results of the chi-square test for different domains are shown in Table 5, and the results of the ANOVA test for different domains are shown in Table 6.

Table 6
ANOVA analysis for different domains

| Index | F value | Significance |
|---|---|---|
| Turn | 3.385 | 0.061 |
| Reward | 1.699 | 0.216 |
| Succ. | 1.425 | 0.271 |

As can be seen from Table 5, the significance of the Turn, Reward, and Succ. rubrics were 0.838, 0.996, and 0.127 respectively all greater than 0.05, indicating that the variance of the analyzed data was chi-square and could be analyzed using a one-way ANOVA. Table 6 shows the significance of Turn, Reward, and Succ. were 0.061, 0.216, and 0.271 respectively, all of which were greater than 0.05. This indicates that the division of the corpus into domains 2, 4, and 6 did not have a significant effect on the rubrics in this experiment, but that different algorithms in the same domain had some effect on the generated dialogue rubrics.

Combining the data from Tables 2–6 shows that the MACD has the highest average task success rate (88%), which is 23%, 50.4%, 23.7%, 4%, and 6% higher than the REINFORCE, DQN, PPO, QMIX, and IterDPL respectively, thus showing that the MACD is the most different from the DQN, with the MACD being 50.4% greater compared to the DQN. This indicates that the DQN is most impacted by the multi-domain dialogue environment, and performance continues to weaken as the number of domains increases, while the DQN model is built by a neural network, and the growth in space reduces the learning speed of the agents when faced with a multi-domain dialogue environment. The MACD is the least different from the QMIX, with the MACD increasing by only 4% compared to the QMIX. It fully illustrates that the QMIX and MACD use the idea of joint information to optimize dialogue policy training, which effectively alleviates the inevitable non-stationarity problem in multi-agent reinforcement learning, while the independent experience playback buffer reduces the gradient dependence. Both algorithms use the CTDE multi-agent policy training framework to achieve better learning performance for dialogue, multiple agents.

Finally, to compare the variability of the Success metrics between the different algorithms in a multi-domain dialogue environment. The MACD was used to generate the approximate direction of the Success with five comparison algorithms at domains 2, 4, and 6 with the number of training sets at 300, and the experimental results are shown in Fig. 4, 5, and 6.

As can be seen in Fig. 4, there is a significant difference in the "Success" metrics of the various algorithms as the system iterates through 200 sets. The MACD achieves the highest Success (74.5%), which is 11.7%, 33.2%, 24.4%, 23.3%, and 12.4% higher than the REINFORCE, DQN, PPO, QMIX, and IterDPL respectively. 33.2%, 24.4%, 23.3%, and 12.4%, respectively. As the number of iteration sets increases, the MACD maintains a higher Success compared to the other comparison algorithms. As can be seen in Fig. 5, at 50 training iterations, the MACD is already higher than the other comparison algorithms, by 16.7%, 22.3%, 19.8%, 16.9%, and 1.2% compared to REINFORCE, DQN, PPO, QMIX, and IterDPL, respectively, and has been in the lead ever since. As can be seen in Fig. 6, the performance of all the comparison algorithms weakened at a domain of 6, except for QMIX as well as the MACD, which increased by 2.4% compared to the QMIX when training iterations of 250 sets.

Combining the data from Fig. 4 - Fig. 6 it can be found that in the dialogue environment of different domains, the dialogue multiple agents differed in the synergy of completing the composite tasks provided in different domains after being trained by different algorithms for dialogue policies. From an overall perspective, each algorithm was impacted by the multi-domain dialogue environment and the composite dialogue task, and as a result, the Success tended to decrease. When compared from an individual algorithm perspective, the MACD proposed in this paper learns faster and performs better with statistically significant advantages. As the size of the domain continues to increase, the MACD shows less fluctuation in the Success measures compared to the other comparison algorithms, with a significant decrease of only 1% from domain 2 to domain 4 and 4% from domain 4 to domain 6. Thus, the MACD proposed in this paper is robust against a multi-domain environment and has even shows to be more effective in a multi-domain environment than a single-domain environment.

## 4.5 Human Evaluation

The MACD proposed in this paper requires a model conversion with real users to better verify the effectiveness of the algorithm through manual evaluation. In the evaluation process, the user interacts with the agents, and Turn and Success are used as evaluation metrics. The algorithm was tested with 1000 dialogues in a single domain environment of the MultiWOZ corpus. The experimental results are shown in Table 7.

Table 7
Evaluation metrics for different algorithms vs. human users

| Method | #Turn | Success |
|---|---|---|
| DQN | 8.21 vs 11.10 | 51.56 vs 51.3 |
| IterDPL | 6.10 vs 9.02 | 63.6 vs 54.7 |
| QMIX | 8.23 vs 10.23 | 60.6 vs 55.3 |
| MACD | **6.26 vs 9.12** | **69.46 vs 52.6** |

As can be seen from Table 7, under the "Turn" metric, the different algorithms outperform human users, with DQN, IterDPL, QMIX, and MACD decreasing by 2.89, 2.92, 2, and 2.86 compared to human users. In the "Success" metric, the DQN, IterDPL, QMIX, and MACD increased by 0.26%, 8.9%, 5.3%, and 16.86% compared to human users, and it can be seen that the MACD performed better than the other comparison algorithms when interacting with humans. From an overall perspective, all algorithms had a lower Success than the scores in Table 1 when interacting with humans in dialogue, and all results differed significantly from the original results. In particular, the MACD has a 4.14% lower Success when interacting with a human compared to interacting with an agent. This is mainly because real human users need to bond over time to have a sense of cooperation and team spirit, and they want to complete the task by getting as much information as possible without making compromises. For this reason, the Turn of human users is larger compared to even the comparison algorithm, and as the sessions become quite long, the Success decreases to some extent.

## 5 Conclusion

In this paper, we propose a reinforcement learning based multi-agent dialogue policy training algorithm—MACD, to solve the problem of lack of collaboration between agents in a multi-domain dialogue environment in a task-oriented dialogue system. The MACD firstly adopts a reinforcement learning framework to build dialogue multiple agents as a way to address the cost consumption of human-computer dialogue interaction, and secondly adopts the CTDE multi-agent policy training framework, combined with the simple extended Actor-Critic single-agent policy gradient method to train dialogue multiple agents policies, as a way to address the lack of joints in multi-agent policy training. The experimental results show that the MACD can effectively improve the success rate of multi-agent jointly completing dialogue tasks and reduce the number of invalid dialogues in dialogue turns in a composite task dialogue scenario. Future work will explore the confidence assignment problem caused by identical returns when dialogue multiple agents interact with the environment after dialogue policy training.

## Declarations

**Data Availability** The datasets generated during and/or analysed during the current study are available in the [MultiWOZ] repository, [✓https://doi.org/10.18653/v1/D18-1547]

**Conflicts of Interest** The authors declare that there are no conflicts of

interest regarding the publication of this paper.

## References

1. Li Q, Li P, Ren Z et al (2022) Knowledge bridging for empathetic dialogue generation[J]. arXiv preprint arXiv:2009.09708,
2. Zhou L, Gao J, Li D et al The design and implementation of xiaoice, an empathetic social chatbot.Computational Linguistics[J].2020, 46(1):53–93
3. Sakata W, Shibata T, Tanaka R et al (2019) FAQ retrieval using query-question similarity and BERT-based query-answer relevance[C].Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval: : 1113–1116
4. Paweł Budzianowski and Ivan Vulić (2019) Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. In Proceedings of the 3rd Workshop on Neural Generation and Translation, pages 15–22, Hong Kong. Association for Computational Linguistics

5. Mrkšić N, Séaghdha DO, Wen TH et al Neural belief tracker: Data-driven dialogue state tracking[J]. arXiv preprint arXiv:1606.03777, 2016.
6. Dai Z, Yang Z, Yang Y et al Transformer-xl: Attentive language models beyond a fixed-length context[J]. arXiv preprint arXiv:1901.02860, 2019.
7. Lipton Z, Li X, Gao J et al (2018) Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems[C]//Proceedings of the AAAI Conference on Artificial agent. 32(1)
8. Li J, Monroe W, Ritter A et al Deep reinforcement learning for dialogue generation[J]. arXiv preprint arXiv:1606.01541, 2016.
9. Ham D, Lee JG, Jang Y et al (2020) End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2[C].Proceedings of the 58th annual meeting of the association for computational linguistics: : 583–592
10. Gür I, Hakkani-Tür D, Tür G et al (2018) User modeling for task oriented dialogues[C].2018 IEEE Spoken Language Technology Workshop (SLT): : 900–906
11. Lin X, Jian W, He J et al (2020) Generating informative conversational response using recurrent knowledge-interaction and knowledge-copy[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. : 41–52
12. Du W, Alan W, Black (2019) Boosting Dialog Response Generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 38–43, Florence, Italy. Association for Computational Linguistics
13. Li X, Chen YN, Li L et al (2017) End-to-end task-completion neural dialogue systems[J]. arXiv preprint arXiv:1703.01008,
14. Li Z, Niu C, Meng F et al (2019) Incremental transformer with deliberation decoder for document grounded dialogues[J]. arXiv preprint arXiv:1907.08854,
15. Jaques N, Shen JH, Ghandeharioun A et al (2020) Human-centric dialog training via offline reinforcement learning[J]. arXiv preprint arXiv:2010.05848,
16. Takanobu R, Zhu H, Huang M Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog[J]. arXiv preprint arXiv:1908.10719, 2019.
17. Li Z, Kiseleva J, De Rijke M (2019) Dialogue generation: From imitation learning to inverse reinforcement learning[C]//Proceedings of the AAAI Conference on Artificial agent. 33(01): 6722–6729
18. Huang X, Ren M, Han Q et al (2021) Emotion Detection for dialogues Based on Reinforcement Learning Framework[J]. IEEE Multimedia 28(2):76–85
19. Le AC (2021) A Deep Reinforcement Learning Model using Long Contexts for Chatbots[C]//2021 International Conference on System Science and Engineering (ICSSE). IEEE, : 83–87
20. Papangelis A, Wang YC, Molino P et al Collaborative multi-agent dialogue model training via reinforcement learning[J]. arXiv preprint arXiv:1907.05507, 2019.
21. Das A, Kottur S, Moura JMF et al (2017) Learning cooperative visual dialog agents with deep reinforcement learning[C]//Proceedings of the IEEE international conference on computer vision. : 2951–2960

22. Zhang Z, Liao L, Zhu X et al Learning goal-oriented dialogue policy with opposite agent awareness[J]. arXiv preprint arXiv:2004.09731, 2020.
23. Wang H, Wong KF (2021) A Collaborative Multi-agent Reinforcement Learning Framework for Dialog Action Decomposition[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. : 7882–7889
24. Kim H, Kim B, Kim G Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness[J]. arXiv preprint arXiv:2004.05816, 2020.
25. Jia Q, Liu Y, Ren S et al (2020) Multi-turn response selection using dialogue dependency relations[J]. arXiv preprint arXiv:2010.01502,
26. Lin Z, Madotto A, Winata GI et al Mintl: Minimalist transfer learning for task-oriented dialogue systems[J]. arXiv preprint arXiv:2009.12005, 2020.
27. Littman ML (1994) Markov games as a framework for multi-agent reinforcement learning[M]//Machine learning proceedings 1994. Morgan Kaufmann, : 157–163
28. Georgios Papoudakis F, Christianos A, Rahman (2019) and Stefano V. Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning.CoRR, abs/1906.04737,
29. Volodymyr Mnih K, Kavukcuoglu D, Silver,Andrei A, Rusu J, Veness MG, Bellemare,Alex Graves M, Riedmiller, Andreas K, Fidjeland G et al (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529
30. Bahdanau D, Brakel P, Xu K et al An actor-critic algorithm for sequence prediction[J]. arXiv preprint arXiv:1607.07086, 2016.
31. Lowe R, Wu YI, Tamar A et al (2017) Multi-agent actor-critic for mixed cooperative-competitive environments[J].Advances in neural information processing systems,30
32. Bernstein DS, Givan R, Immerman N et al (2002) The complexity of decentralized control of Markov decision processes[J]. Math Oper Res 27(4):819–840

# Figures

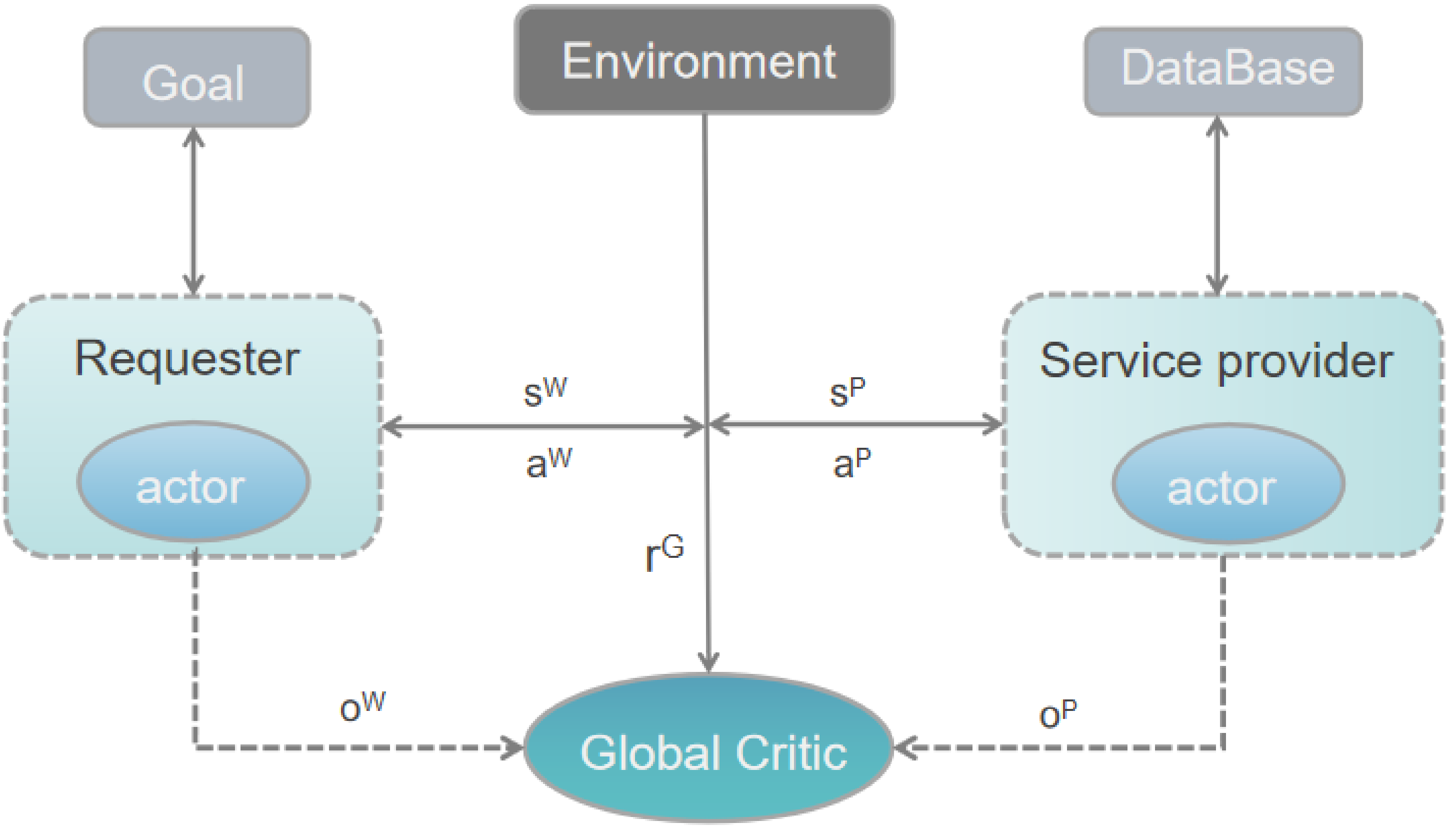


**Figure 1**

Basic framework of the MACD

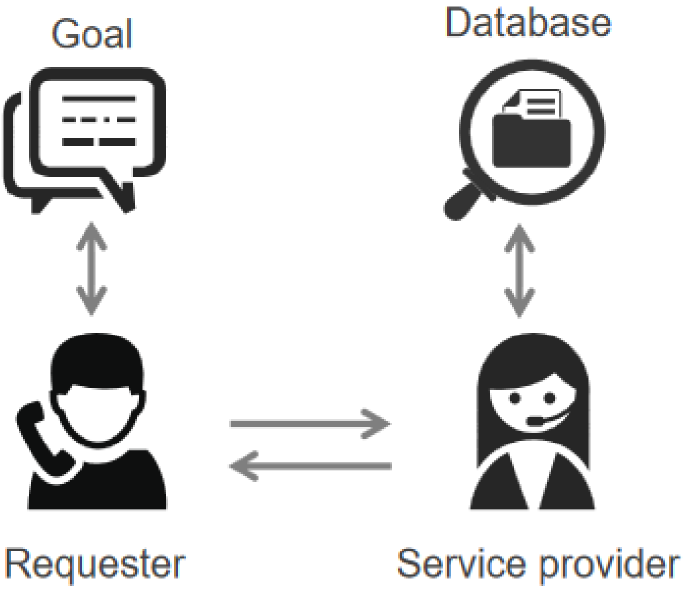


**Figure 2**

Dialogue multiple agents interaction

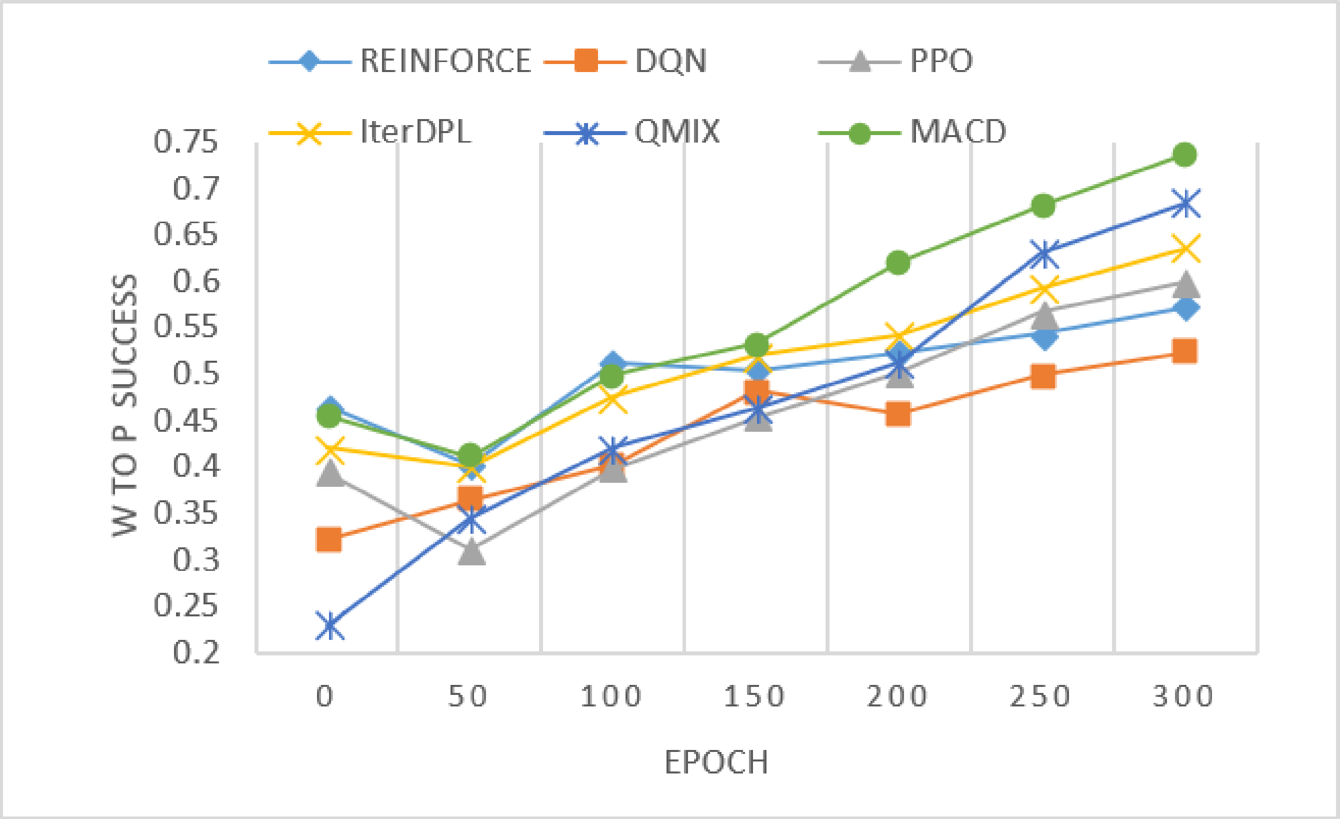
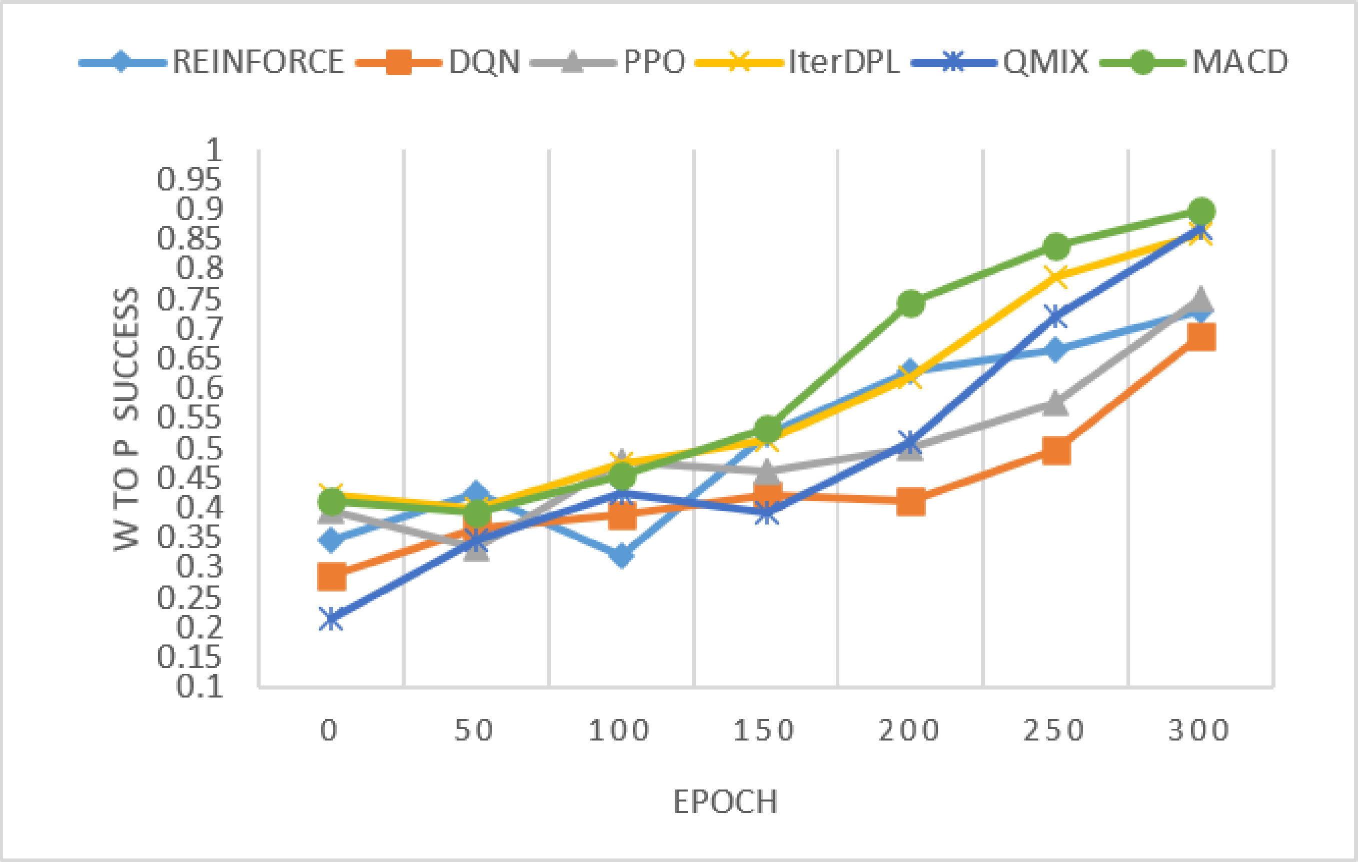


**Figure 3**

Success (%) for interactions between the Requester and the Service provider



### Figure 4

Success (%) for interactions between domains=2 the Requester and the Service provider

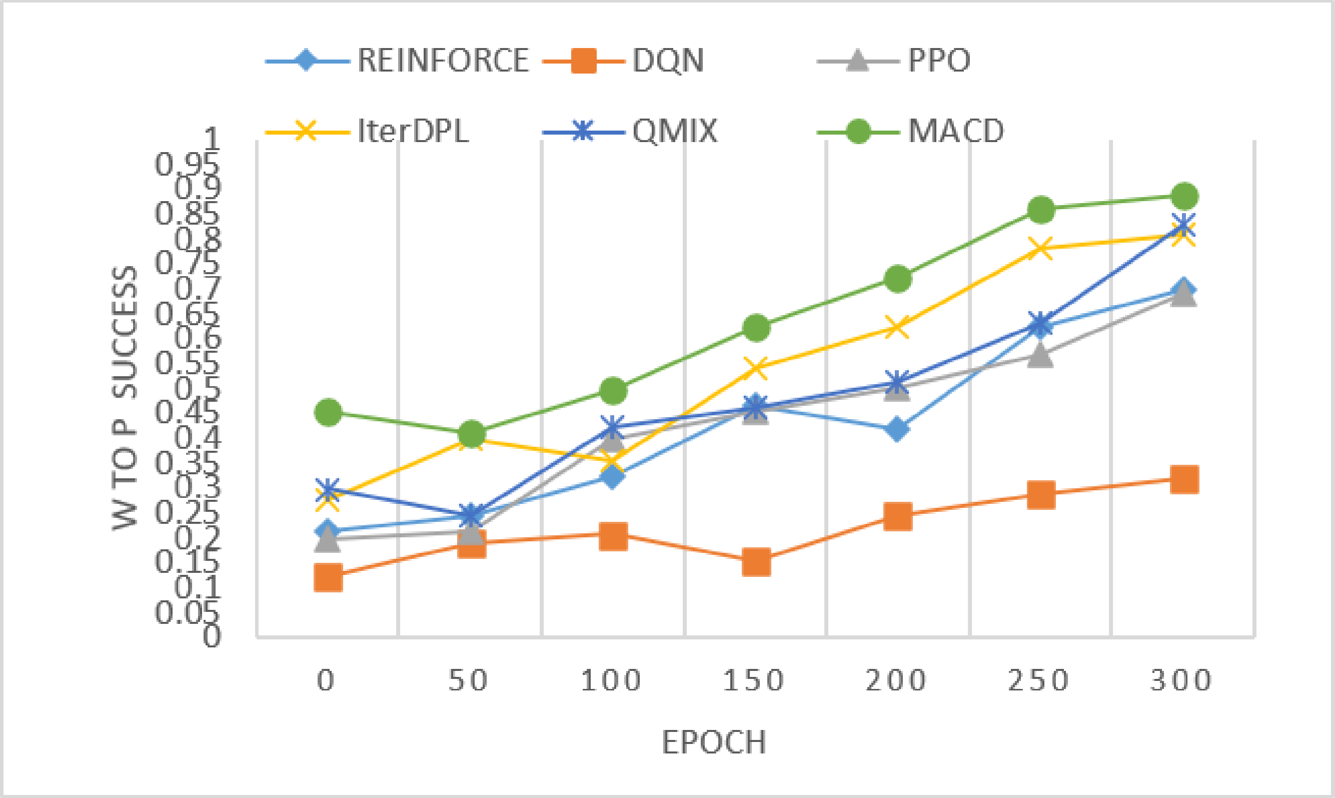


**Figure 5**

Success (%) for interactions between domains=4 the Requester and the Service provider

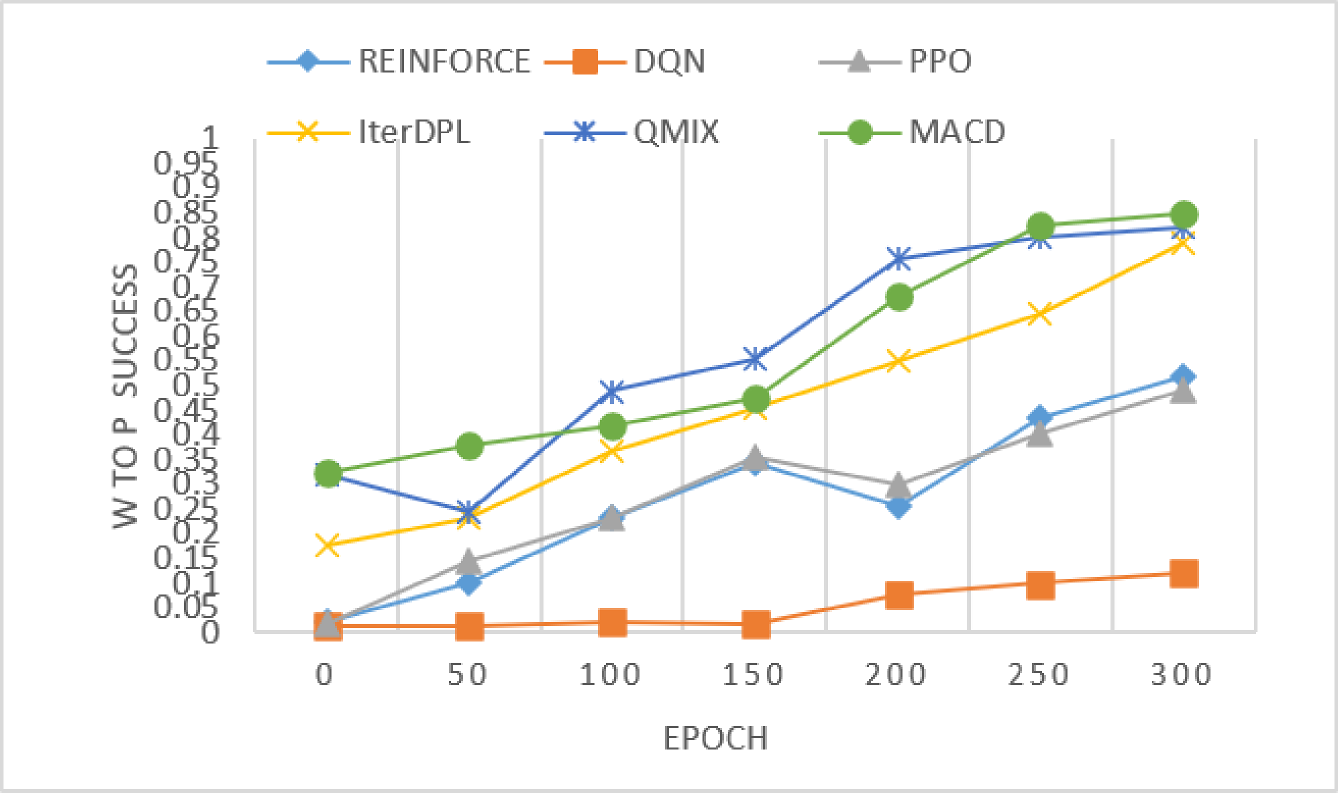


**Figure 6**

Success (%) for interactions between domains=6 the Requester and the Service provider