

# Towards Task-Oriented Dialogue in Mixed Domains

Tho Luong Chi  
 FPT Technology Research Institute  
 FPT University, Hanoi, Vietnam  
*tholc2@fpt.com.vn*

Phuong Le-Hong  
 FPT Technology Research Institute  
 Vietnam National University, Hanoi, Vietnam  
*phuonglh@vnu.edu.vn*

**Abstract**—This work investigates the task-oriented dialogue problem in mixed-domain settings. We study the effect of alternating between different domains in sequences of dialogue turns using two related state-of-the-art dialogue systems. We first show that a specialized state tracking component in multiple domains plays an important role and gives better results than an end-to-end task-oriented dialogue system. We then propose a hybrid system which is able to improve the belief tracking accuracy of about 28% of average absolute point on a standard multi-domain dialogue dataset. These experimental results give some useful insights for improving our commercial chatbot platform FPT.AI, which is currently deployed for many practical chatbot applications.

**Index Terms**—task-oriented dialogue; multi-domain belief tracking; mixed-domain belief tracking; natural language processing

## I. INTRODUCTION

In this work, we investigate the problem of task-oriented dialogue in mixed-domain settings. Our work is related to two lines of research in Spoken Dialogue System (SDS), namely *task-oriented dialogue system* and *multi-domain dialogue system*. We briefly review the recent literature related to these topics as follows.

Task-oriented dialogue systems are computer programs which can assist users to complete tasks in specific domains by understanding user requests and generating appropriate responses within several dialogue turns. Such systems are useful in domain-specific chatbot applications which help users find a restaurant or book a hotel. Conventional approach for building a task-oriented dialogue system is concerned with building a quite complex pipeline of many connected components. These components are usually independently developed which include at least four crucial modules: a natural language understanding module, a dialogue state tracking module, a dialogue policy learning module, and a answer generation module. Since these systems components are usually trained independently, their optimization targets may not fully align with the overall system evaluation criteria [1]. In addition, such a pipeline system often suffers from error propagation where error made by upstream modules are accumulated and got amplified to the downstream ones.

To overcome the above limitations of pipeline task-oriented dialogue systems, much research has focused recently in designing end-to-end learning systems with neural network-based models. One key property of task-oriented dialogue model is that it is required to reason and plan over multiple di-

logue turns by aggregating useful information during the conversation. Therefore, sequence-to-sequence models such as the encoder-decoder based neural network models are proven to be suitable for both task-oriented and non-task-oriented systems. Serban et al. proposed to build end-to-end dialogue systems using generative hierarchical recurrent encoder-decoder neural network [2]. Li et al. presented persona-based models which incorporate background information and speaking style of interlocutors into LSTM-based seq2seq network so as to improve the modeling of human-like behavior [3]. Wen et al. designed an end-to-end trainable neural dialogue model with modularly connected components [4]. Bordes et al. [5] proposed a task-oriented dialogue model using end-to-end memory networks. At the same time, many works explored different kinds of networks to model the dialogue state, such as copy-augmented networks [6], gated memory networks [7], query-regression networks [8]. These systems do not perform slot-filling or user goal tracking; they rank and select a response from a set of response candidates which are conditioned on the dialogue history.

One of the significant effort in developing end-to-end task-oriented systems is the recent Seqicity framework [9]. This framework also relies on the sequence-to-sequence model and can be optimized with supervised or reinforcement learning. The Seqicity framework introduces the concept of *belief span* (bspan), which is a text span that tracks the dialogue states at each turn. In this framework, the task-oriented dialogue problem is decomposed into two stages: bspan generation and response generation. This framework has been shown to significantly outperform state-of-the-art pipeline-based methods.

The second line of work in SDS that is related to this work is concerned with multi-domain dialogue systems. As presented above, one of the key components of a dialogue system is dialogue state tracking, or belief tracking, which maintains the states of conversation. A state is usually composed of user’s goals, evidences and information which is accumulated along the sequence of dialogue turns. While the user’s goal and evidences are extracted from user’s utterances, the useful information is usually aggregated from external resources such as knowledge bases or dialogue ontologies. Such knowledge bases contain slot type and slot value entries in one or several predefined domains. Most approaches have difficulty scaling up with multiple domains due to the dependency of their model parameters on the underlying knowledge bases. Recently, Ramadan et al. [10] has introduced a novel approach which

Turn	Dialogue
$H_1$	are there any chinese restaurants nearby?
Match	<inf> chinese </ inf> <Req> </ Rep>
$M_1$	the closest is POI_SLOT restaurant , and POI_SLOT is DISTANCE_SLOT away with TRAFFIC_INFO
$H_2$	what's the weather like today?
Match	<inf> chinese </ inf> <inf> weather </ inf> <Rep> </ Rep>
$M_2$	what city are you wanting to know the weather for ?
$H_3$	for Newyork city please?
Match	<inf> chinese </ inf> <inf> weather </ inf> <Rep> Newyork city </ Rep>
$M_3$	Today is WEATHER_ATTRIBUTE_SLOT, and WEATHER_ATTRIBUTE_SLOT.

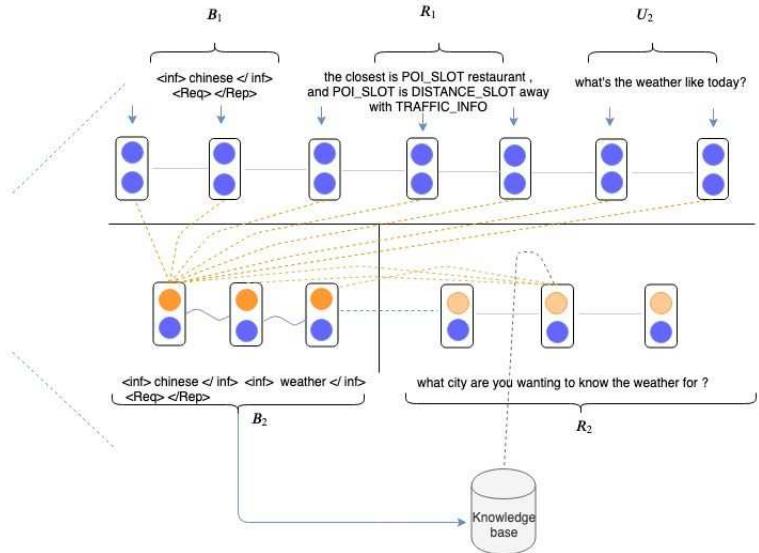


Fig. 1. Sequicity architecture.

utilizes semantic similarity between dialogue utterances and knowledge base terms, allowing the information to be shared across domains. This method has been shown not only to scale well to multi-domain dialogues, but also outperform existing state-of-the-art models in single-domain tracking tasks.

The problem that we are interested in this work is task-oriented dialogue in mixed-domain settings. This is different from the multi-domain dialogue problem above in several aspects, as follows:

- First, we investigate the phenomenon of alternating between different dialogue domains in subsequent dialogue turns, where each turn is defined as a pair of user question and machine answer. That is, the domains are mixed between turns. For example, in the first turn, the user requests some information of a restaurant; then in the second turn, he switches to the a different domain, for example, he asks about the weather at a specific location. In a next turn, he would either switch to a new domain or come back to ask about some other property of the suggested restaurant. This is a realistic scenario which usually happens in practical chatbot applications in our observations. We prefer calling this problem mixed-domain dialogue rather than multiple-domain dialogue.
- Second, we study the effect of the mixed-domain setting in the context of multi-domain dialogue approaches to see how they perform in different experimental scenarios.

The main findings of this work include:

- A specialized state tracking component in multiple domains still plays an important role and gives better results than a state-of-the-art end-to-end task-oriented dialogue system.
- A combination of specialized state tracking system and an end-to-end task-oriented dialogue system is beneficial in mix-domain dialogue systems. Our hybrid system is able to improve the belief tracking accuracy of about 28%

of average absolute point on a standard multi-domain dialogue dataset.

- These experimental results give some useful insights on data preparation and acquisition in the development of the chatbot platform FPT.AI<sup>1</sup>, which is currently deployed for many practical chatbot applications.

The remainder of this paper is structured as follows. First, Section II discusses briefly the two methods in building dialogue systems that our method relies on. Next, Section III presents experimental settings and results. Finally, Section IV concludes the paper and gives some directions for future work.

## II. METHODOLOGY

In this section, we present briefly two methods that we use in our experiments which have been mentioned in the previous section. The first method is the Sequicity framework and the second one is the state-of-the-art multi-domain dialogue state tracking approach.

### A. Sequicity

Figure 1 shows the architecture of the Sequicity framework as described in [9]. In essence, in each turn, the Sequicity model first takes a bspan ( $B_1$ ) and a response ( $R_1$ ) which are determined in the previous step, and the current human question ( $U_2$ ) to generate the current bspan. This bspan is then used together with a knowledge base to generate the corresponding machine answer ( $R_2$ ), as shown in the right part of Figure 1.

The left part of that figure shows an example dialogue in a mixed-domain setting (which will be explained in Section III).

<sup>1</sup><http://fpt.ai/>

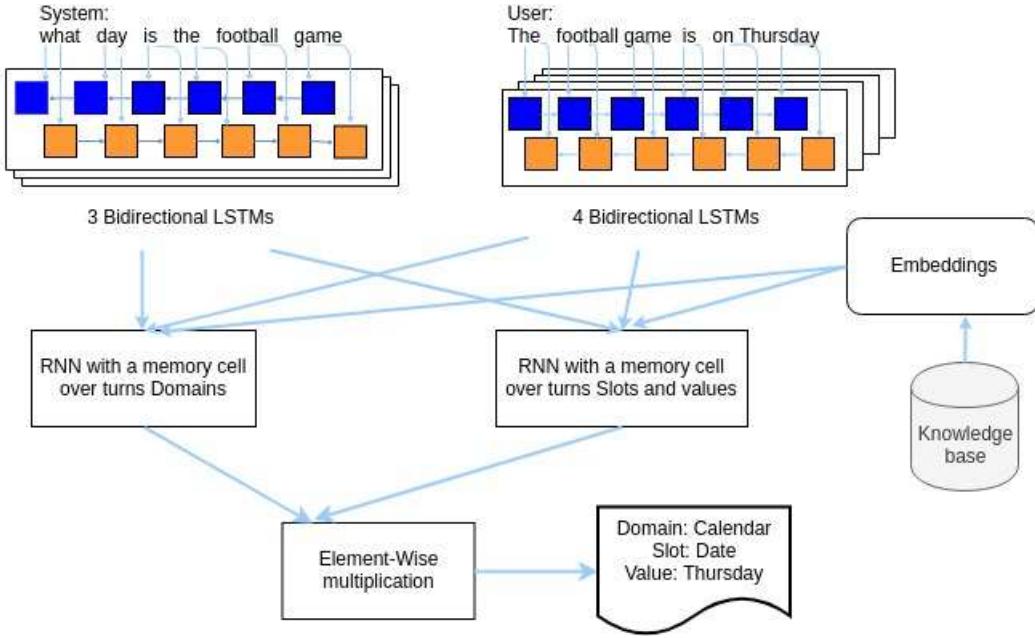


Fig. 2. Multi-domain belief tracking with knowledge sharing.

### B. Multi-domain Dialogue State Tracking

Figure 2 shows the architecture of the multi-domain belief tracking with knowledge sharing as described in [10]. This is the state-of-the-art belief tracker for multi-domain dialogue.

This system encodes system responses with 3 bidirectional LSTM network and encodes user utterances with 3+1 bidirectional LSTM network. There are in total 7 independent LSTMs. For tracking domain, slot and value, it uses 3 corresponding LSTMs, either for system response or user utterance. There is one special LSTM to track the user affirmation. The semantic similarity between the utterances and ontology terms are learned and shared between domains through their embeddings in the same semantic space.

### III. EXPERIMENTS

In this section, we present experimental settings, different scenarios and results. We first present the datasets, then implementation settings, and finally obtained results.

#### A. Datasets

We use the publicly available dataset KVRET [6] in our experiments. This dataset is created by the Wizard-of-Oz method [11] on Amazon Mechanical Turk platform. This dataset includes dialogues in 3 domains: calendar, weather, navigation (POI) which is suitable for our mix-domain dialogue experiments. There are 2,425 dialogues for training, 302 for validation and 302 for testing, as shown in the upper half of Table I.

In this original dataset, each dialogue is of a single domain where all of its turns are on that domain. Each turn is composed of a sentence pair, one sentence is a user utterance, the other sentence is the corresponding machine response. A dialogue is a sequence of turns. To create mix-domain

dialogues for our experiments, we make some changes in this dataset as follows:

- We keep the dialogues in the calendar domain as they are.
- We take a half of dialogues in the weather domain and a half of dialogues in the POI domain and mix their turns together, resulting in a dataset of mixed weather-POI dialogues. In this mixed-domain dialogue, there is a turn in the weather domain, followed by a turn in POI domain or vice versa.

We call this dataset *the sequential turn dataset*. Since the start turn of a dialogue has a special role in triggering the learning systems, we decide to create another and different mixed-domain dataset with the following mixing method:

- The first turn and the last turn of each dialogue are kept as in their original.
- The internal turns are mixed randomly.

We call this dataset *the random turn dataset*. Some statistics of these mixed-domain datasets are shown in the lower half of the Table I.

#### B. Experimental Settings

For the task-oriented Sequicity model, we keep the best parameter settings as reported in the original framework, on the same KVRET dataset [9]. In particular, the hidden size of GRU unit is set to 50; the learning rate of Adam optimizer is 0.003. In addition to the original GRU unit, we also re-run this framework with simple RNN unit to compare the performance of different recurrent network types. The Sequicity tool is freely available for download.<sup>2</sup>

<sup>2</sup><https://github.com/WING-NUS/sequicity>

TABLE I

SOME STATISTICS OF THE DATASETS USED IN OUR EXPERIMENTS. THE ORIGINAL KVRET DATASET IS SHOWN IN THE UPPER HALF OF THE TABLE. THE MIXED DATASET IS SHOWN IN THE LOWER HALF OF THE TABLE.

Dataset	<b>KVRET</b>		
Dialogues	Train: 2,425 ; Test: 302 ; Dev.: 302		
Domains	calendar	weather	POI

Dataset	<b>MIXED DOMAINS</b>		
Domains	calendar	weather	POI
Train	828	398	400
Test	102	50	50
Dev.	102	50	50
			Mixed weather – POI

For the multi-domain belief tracker model, we set the hidden size of LSTM units to 50 as in the original model; word embedding size is 300 and number of training epochs is 100. The corresponding tool is also freely available for download.<sup>3</sup>

### C. Results

Our experimental results are shown in Table II. The first half of the table contains results for task-oriented dialogue with the Sequicity framework with two scenarios for training data preparation. For each experiment, we run our models for 3 times and their scores are averaged as the final score. The *mixed training* scenario performs the mixing of both the training data, development data and the test data as described in the previous subsection. The *non-mixed training* scenario performs the mixing only on the development and test data, keeps the training data unmixed as in the original KVRET dataset. As in the Sequicity framework, we report entity match rate, BLEU score and Success F1 score. **Entity match rate** evaluates task completion, it determines if a system can generate all correct constraints to search the indicated entities of the user. **BLEU** score evaluates the language quality of generated responses. **Success F1** balances the recall and precision rates of slot answers. For further details on these metrics, please refer to [9].

In the first series of experiments, we evaluate the Sequicity framework on different mixing scenarios and different recurrent units (GRU or RNN), on two mixing methods (sequential turn or random turn), as described previously. We see that when the training data is kept unmixed, the match rates are better than those of the mixed training data. It is interesting to note that the GRU unit is much more sensitive with mixed data than the simple RNN unit with the corresponding absolute point drop of about 10%, compared to about 3.5%. However, the entity match rate is less important than the Success F1 score, where the GRU unit outperforms RNN in both sequential turn and random turn by a large margin. It is logical that if the test data are mixed but the training data are unmixed, we get lower scores than when both the training data and test data are mixed. The GRU unit is also better than the RNN unit on response generation in terms of BLEU scores.

We also see that the task-oriented dialogue system has difficulty running on mixed-domain dataset; it achieves only

about 75.62% of Success F1 in comparison to about 81.1% (as reported in the Sequicity paper, not shown in our table). Appendix A shows some example dialogues generated automatically by our implemented system.

In the second series of experiments, we evaluate the belief tracking components of two systems, the specialized multi-domain belief tracker and the Sequicity bspan component. As shown in the lower half of the Table II, Sequicity capability of belief tracking is much worse than that of the multi-domain belief tracker. The slot accuracy gap between the tools is about 21.6%, the value accuracy gap is about 34.4%; that is a large average gap of 28% of accuracy. This result suggests a future work on combining a specialized belief tracking module with an end-to-end task-oriented dialogue system to improve further the performance of the overall dialogue system.

### D. Error Analysis

In this subsection, we present an example of erroneous mixed dialogue with multiple turns. Table III shows a dialogue in the test set where wrong generated responses of the Sequicity system are marked in bold font.

In the first turn, the system predicts incorrectly the bspan, thus generates wrong slot values (`heavy traffic` and `Pizza Hut`). The word `Pizza Hut` is an arbitrary value selected by the system when it cannot capture the correct value `home` in the bspan. In the second turn, the machine is not able to capture the value `this_week`. This failure does not manifest immediately at this turn but it is accumulated to make a wrong answer at the third turn (`monday` instead of `this_week`).

The third turn is of domain `weather` and the fourth turn is switched to domain `POI`. The bspan value `cleveland` is retained through cross domain, resulting in an error in the fourth turn, where `cleveland` is shown instead of `home`. This example demonstrates a weakness of the system when being trained on a mixed-domain dataset. In the fifth turn, since the system does not recognize the value `fastest` in the bspan, it generates a random and wrong value `moderate traffic`. Note that the generated answer of the sixth turn is correct despite of the wrong predicted bspan; however, it is likely that if the dialogue continues, this wrong bspan may result in more answer mistakes. In such situations, multi-domain belief tracker usually performs better at bspan prediction.

<sup>3</sup><https://github.com/osmanio2/multi-domain-belief-tracking>

TABLE II

OUR EXPERIMENTAL RESULTS. **MATCH.** AND **SUCC.** **F1** ARE ENTITY MATCH RATE AND SUCCESS F1. THE UPPER HALF OF THE TABLE SHOWS RESULTS OF TASK-ORIENTED DIALOGUE WITH THE SEQUICITY FRAMEWORK. THE LOWER HALF OF THE TABLE SHOWS RESULTS OF MULTI-DOMAIN BELIEF TRACKER.

		Case 1 - sequential turn			Case 2 - random turn		
Sequicity		Match.	BLEU	Succ. F1	Match.	BLEU	Succ. F1
mixed training	GRU	0.6367	<b>0.1930</b>	<b>0.7358</b>	0.6860	0.1862	<b>0.7562</b>
	RNN	0.7354	0.1847	0.7129	0.6591	0.1729	0.7105
non-mixed training	GRU	0.7399	0.1709	0.7055	<b>0.7488</b>	0.1820	0.7173
	RNN	0.7706	0.1453	0.6156	0.6995	0.1580	0.6633
		Domain - accuracy	Slot - accuracy	Value - accuracy			
<b>Belief tracker</b>	Multi-domain	0.8253	<b>0.9329</b>	<b>0.9081</b>			
	Sequicity		0.7171	0.5644			

TABLE III

A MIXED DIALOGUE EXAMPLE IN THE TEST SET WITH ERRONEOUS GENERATED RESPONSES. THE LAST TWO COLUMNS SHOW RESPECTIVELY THE SYSTEM'S GENERATED BSPAN AND THE GOLD BSPAN OR BELIEF TRACKER.

Turn	Dialogue	Domain	Generated bspan	Gold bspan
1.Human	car please send me the fastest route to go back home	POI	EOS_Z1	home fastest EOS_Z1
1.Machine	there is <b>heavy traffic</b> on the way to <b>Pizza Hut</b> .	POI		
2.Human	what is the weather like this week	Weather info	EOS_Z1	this_week EOS_Z1
2.Machine	what city are you interested in ?	Weather info		
3.Human	Cleveland .	Weather info	cleveland EOS_Z1	cleveland EOS_Z1
3.Machine	<b>it will be clear sky on monday</b>	Weather info		
4.Human	what is the address ?	POI	cleveland EOS_Z1	home fastest EOS_Z1
4.Machine	<b>cleveland</b> is located at <b>5672 barringer street</b>	POI		
5.Human	please send me the fastest route on screen and let 's go there , thank you !	POI	cleveland EOS_Z1	home fastest EOS_Z1
5.Machine	there is <b>moderate traffic</b> on our way but i sent the route to your screen , drive carefully !	POI		
6.Human	thank .		cleveland EOS_Z1	EOS_Z1
6.Machine	you are welcome .			

#### IV. CONCLUSION

We have presented the problem of mixed-domain task-oriented dialogue and its empirical results on two datasets. We employ two state-of-the-art, publicly available tools, one is the Sequicity framework for task-oriented dialogue, and another is the multi-domain belief tracking system. The belief tracking capability of the specialized system is much better than that of the end-to-end system. We also show the difficulty of task-oriented dialogue systems on mixed-domain datasets through two series of experiments. These results give some useful insights in combining the approaches to improve the performance of a commercial chatbot platform which is under active development in our company. We plan to extend this current research and integrate its fruitful results into a future version of the platform.

#### REFERENCES

- [1] Bing Liu, Gokhan Tur, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of NAACL*, 2018.
- [2] Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of AAAI*, 2016.
- [3] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In *Proceedings of ACL*, 2016.
- [4] Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of EACL*, 2017.
- [5] Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialogue. In *Proceedings of ICLR*, 2017.
- [6] Mihail Eric and Christopher D. Manning. A copy-augmented sequence-to-sequence architecture gives good performance on task-oriented dialogue. In *Proceedings of EACL*, 2017.
- [7] Fei Liu and Julien Perez. Gated end-to-end memory networks. In *Proceedings of EACL*, 2017.
- [8] Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. Query-regression networks for machine comprehension. In *Preprint ArXiv*, 2016.
- [9] Wenqiang Lei, Xisen Jin, Zhaochun Ren, Xiangnan He, Min-Yen Kan, and Dawei Yin. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of ACL*, 2018.
- [10] Milica Gasic Osman Ramadan, Paweł Budzianowski. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of ACL*, 2018.
- [11] John F Kelley. Iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems*, 2(1):26–41, 1984.

## APPENDIX

The following is three example dialogues generated by our system. The first dialogue is in single-domain.

<b>Turn</b>	<b>Dialogue 1</b>	<b>Domain</b>
Human	What time is my next dinner event and who will be attending?	Calendar
Machine	your dinner is on MONDAY at 5 PM .	Calendar
Human	That will do, thanks.	
Machine	you are welcome	

The next two dialogues are in mixed-domains.

<b>Turn</b>	<b>Dialogue 2</b>	<b>Domain</b>
Human	Please give me the address and directions to home within 5 miles of here, avoiding all heavy traffic.	POI
Machine	HOME is located at 5671 BARRINGER STREET	POI
Human	what is the forecast for the next 7 days?	weather info
Machine	what city are you interested in ?	weather info
Human	Cleveland, will it be windy?	weather info
Machine	it will not be windy in Cleveland this week .	weather info
Human	Thank you.	
Machine	you're welcome.	

<b>Turn</b>	<b>Dialogue 3</b>	<b>Domain</b>
Human	are there any chinese restaurants nearby?	POI
Machine	the closest is TAI PAN restaurant , and TAI PAN is 5 miles away with NO TRAFFIC	POI
Human	what's the weather like today?	weather info
Machine	what city are you wanting to know the weather for ?	weather info
Human	for Newyork city please?weather info	
Machine	Today is CLEAR SKIES, and WARM.	weather info
Human	ok, good, set the direction to there	POI
Machine	the address is 830 Almanor Ln , i sent it on your screen	POI
Human	ok thanks	
Machine	you're welcome !	