On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems

introduction

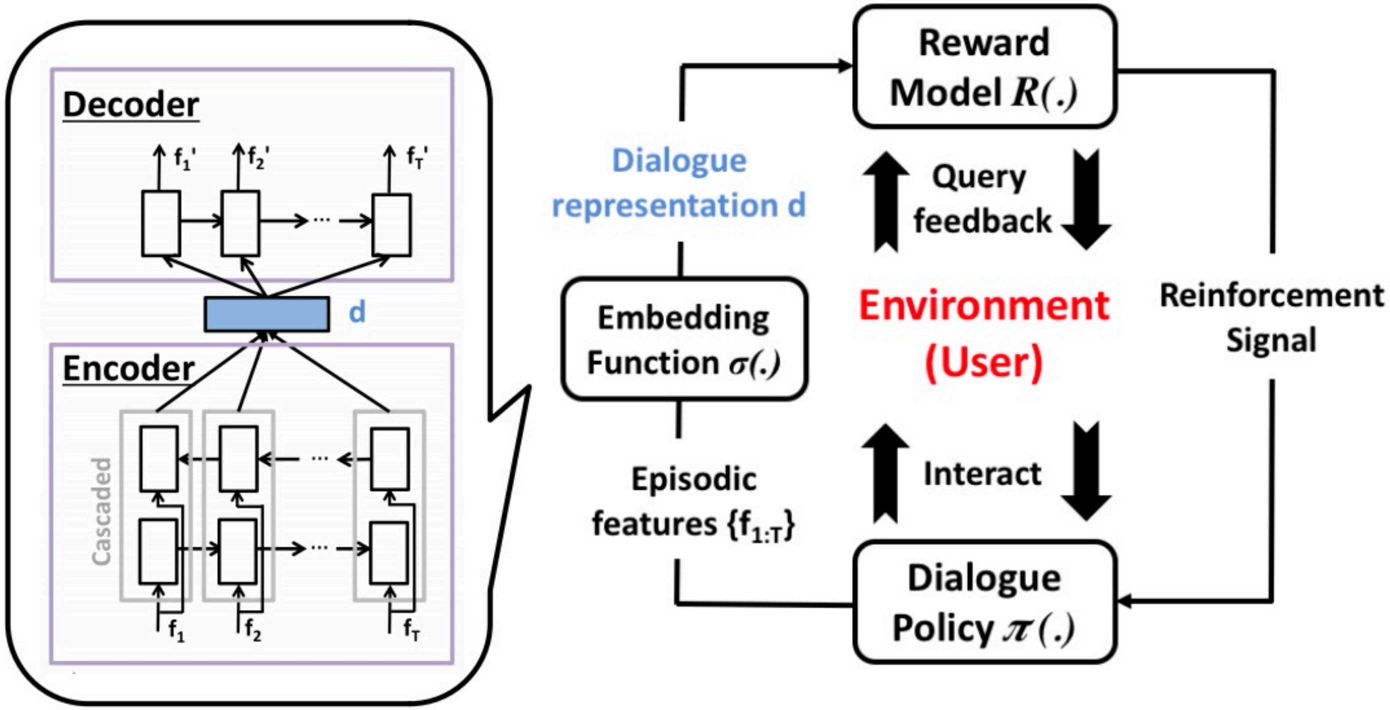
1. an on-line learning framework whereby the dialogue policy is jointly trained alongside the reward model via active learning with a Gaussian process model.

2.This Gaussian process operates on a continuous space dialogue representation generated in an unsupervised fashion using a recurrent neural network encoder-decoder.

3.reduce data annotation costs and mitigate noisy user feedback in dialogue policy learning.

Three main parts of propose framework

1. A dialogue embedding function
2. An active reward model of user feedback
3. A dialogue policy
4. the key contribution here is to learn the noise robust reward model and the dialogue policy simultaneo-usly on-line, using the user as a ‘supervisor’.



Unsupervised Dialogue Embeddings

$$\overrightarrow{\mathbf{h_t}} = LSTM(\mathbf{f_t}, \overrightarrow{\mathbf{h}}_{t-1})$$
$$\overleftarrow{\mathbf{h_t}} = LSTM(\mathbf{f_t}, \overleftarrow{\mathbf{h}}_{t+1})$$

$$\mathbf{d} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h_t}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}||\mathbf{f}_t - \mathbf{f}'_t||^2$$

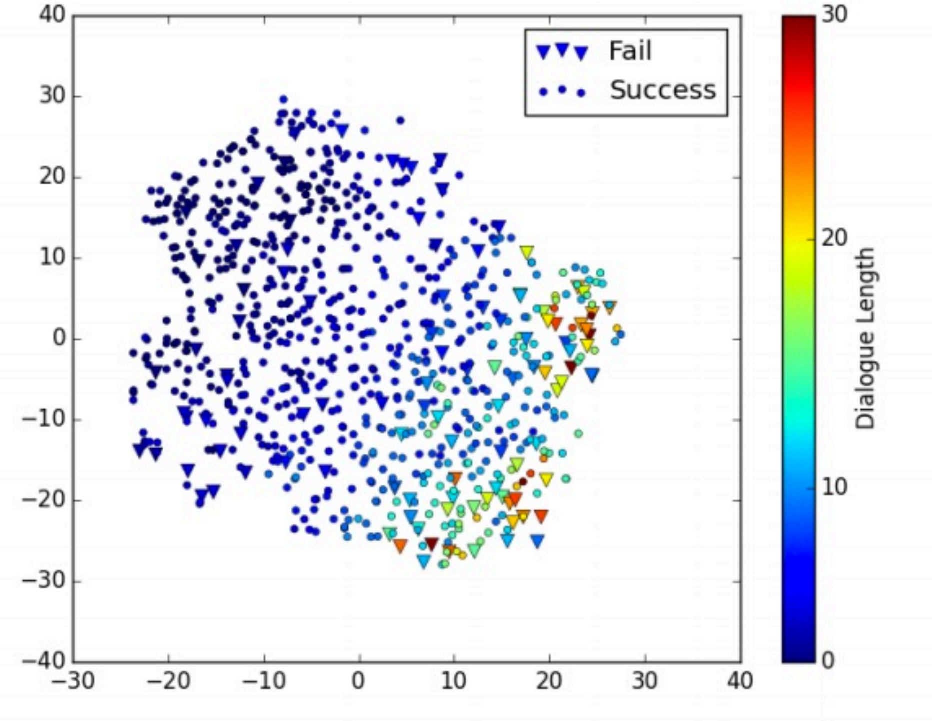
user intention determined

by the semantic decoder, the distribution over each

concept of interest defined in the ontology, a one-hot encoding of the system's reply action, and the

turn number normalised by the maximum number

of turns (here 30)

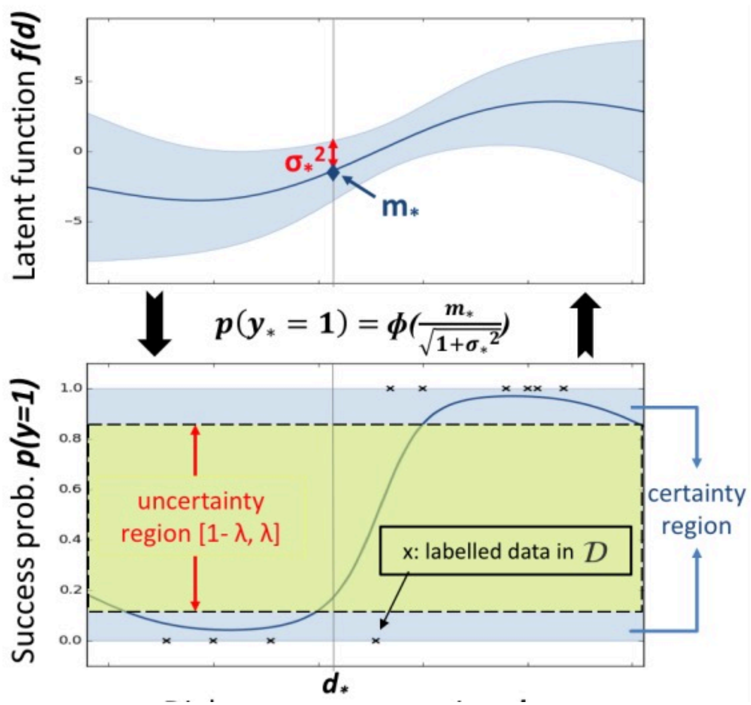


Active Reward Learning

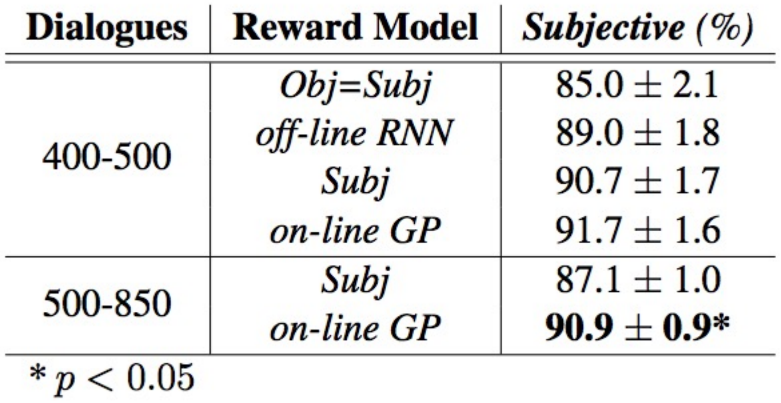
$$p(y = 1|\mathbf{d}, \mathcal{D}) = \phi(f(\mathbf{d}|\mathcal{D})).$$

$$k(\mathbf{d}, \mathbf{d}') = p^2 \exp(-\frac{||\mathbf{d} - \mathbf{d}'||^2}{2l^2}) + \sigma_n^2$$

a latent function $f(\mathbf{d}|\mathcal{D}) : \mathcal{R}^{dim(\mathbf{d})} \rightarrow \mathcal{R}$

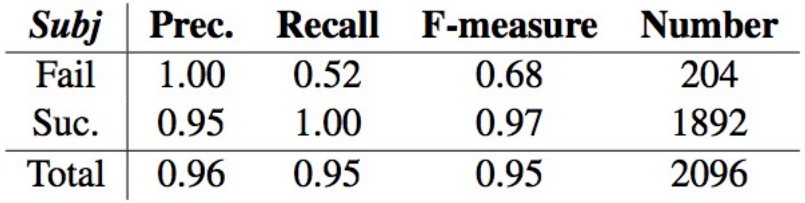

Latent function *f(d)*
5
0
-5
σ_*^2
m_*
$p(y_* = 1) = \phi(\frac{m_*}{\sqrt{1+\sigma_*^2}})$
Success prob. *p(y=1)*
1.0
0.8
0.6
0.4
0.2
0.0
uncertainty region [1- λ, λ]
x: labelled data in $\mathcal{D}$
certainty region
d_*

results



Dialogues	**Reward Model**	***Subjective*** **(%)**
400-500	*Obj=Subj*	85.0 ± 2.1
	off-line RNN	89.0 ± 1.8
	Subj	90.7 ± 1.7
	on-line GP	91.7 ± 1.6
500-850	*Subj*	87.1 ± 1.0
	on-line GP	**90.9 ± 0.9***

* $p < 0.05$



Subj	**Prec.**	**Recall**	**F-measure**	**Number**
Fail	1.00	0.52	0.68	204
Suc.	0.95	1.00	0.97	1892
Total	0.96	0.95	0.95	2096

Center for Speech and Language Technologies

Thank you!

Presented by Jiyuan Zhang