Knowing What You Know: Calibrating Dialogue Belief State Distributions via Ensembles

Carel van Niekerk, Michael Heck, Christian Geishauser
Hsien-Chin Lin, Nurul Lubis, Marco Moresi, Milica Gašić
Heinrich Heine University Düsseldorf, Germany
niekerk,heckmi,geishaus,linh,lubis,moresi,gasic@hhu.de

Abstract

The ability to accurately track what happens during a conversation is essential for the performance of a dialogue system. Current state-of-the-art multi-domain dialogue state trackers achieve just over 55% accuracy on the current go-to benchmark, which means that in almost every second dialogue turn they place full confidence in an incorrect dialogue state. Belief trackers, on the other hand, maintain a distribution over possible dialogue states. However, they lack in performance compared to dialogue state trackers, and do not produce well calibrated distributions. In this work we present state-of-the-art performance in calibration for multi-domain dialogue belief trackers using a calibrated ensemble of models. Our resulting dialogue belief tracker also outperforms previous dialogue belief tracking models in terms of accuracy.

1 Introduction

Task-oriented dialogue systems aim to act as assistants to their users, solving tasks such as finding a restaurant, booking a train, or providing information about a tourist attraction. They have become very popular with the introduction of virtual assistants such as Siri and Alexa.

Two tasks are fundamental to such a system. The first is the ability to track what happened in the conversation, referred to as **tracking**. Based on the result of tracking, the system needs to conduct the conversation towards the fulfilment of the user goal, referred to as **planning**. The tracking component summarises the dialogue history, or the past, while the planning component manages the dialogue and concerns the future. In this work we focus on the first component.

Early approaches to statistical dialogue modelling view dialogue as a Markov decision process (Levin et al., 1998) and define a set of dialogue states that the conversation can be in at any given dialogue turn. The tracking component tracks the **dialogue state**. In recent years discriminative models achieve state-of-the-art dialogue state tracking (DST) results (Kim et al., 2019; Zhang et al., 2019; Heck et al., 2020). Still, in a multi-domain setting such as MultiWOZ (Eric et al., 2019; Budzianowski et al., 2018), they achieve an accuracy of just over 55%. This means that in approximately 45% of cases they make a wrong prediction and, even worse, they have full confidence in that wrong prediction.

In the wake of statistical dialogue modeling, the use of partially observable Markov decision processes has been proposed to address this issue. The idea is to model the probability over all possible dialogue states in every dialogue turn (Williams and Young, 2007). This probability distribution is referred to as the **belief state**. The advantages of belief tracking are probably best illustrated by an excerpt from a dialogue with a real user in (Metallinou et al., 2013): even though the dialogue state predicted with the highest probability is not the true one, the system is able to provide a valid response because the true dialogue state also has assigned a non-zero probability.

> A model is considered well **calibrated** if its confidence estimates are aligned with the empirical likelihood of its predictions (Desai and Durrett, 2020).

The belief state can be modelled by deep learning-based approaches such as the neural belief tracker (Mrkšić et al., 2017), the multi-domain belief tracker (Ramadan et al., 2018), the globally conditioned encoder belief tracker (Nouri and Hosseini-Asl, 2018) and the slot utterance matching belief tracker (SUMBT) (Lee et al., 2019) models. None of these models however address the issue of calibrating the probability distribution that

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3096–3102
November 16 - 20, 2020.

they provide, resulting in them being more confident than they should be. In a dialogue setting, overconfidence can lead to bad decisions and unsuccessful dialogues.

In this work, we present methods for learning well-calibrated belief distributions. Our contributions are the following:

- We present the state-of-the-art performance in calibration for dialogue belief trackers using a calibrated ensemble of models, called the calibrated ensemble belief state tracker (CEBST).
- Our model achieves best overall joint goal accuracy among the state-of-the-art **belief** tracking models.

Such a well-calibrated belief tracking model is essential for the planning component to successfully conduct dialogue.

2 Related Work

Since no other belief tracking methods that we are aware of have achieved success in producing well-calibrated confidence, we look towards methods used in other language tasks. Natural language inference is a related task that also benefits from well-calibrated confidence in predictions. Desai and Durrett (2020) introduce the use of post-processing techniques such as temperature scaling to produce better-calibrated confidence estimates.

Additionally, there have been recent advances in the construction of more adequate loss functions. These methods, including Bayesian matching and prior networks, aim to learn well-calibrated models without the burden of requiring many extra parameters. These methods achieve good calibration in computer vision tasks such as CIFAR (Joo et al., 2020; Malinin and Gales, 2018; Szegedy et al., 2016).

When the limitations of a single model still inhibit us from producing more accurate and better-calibrated models, a popular alternative is to use an ensemble of models. Recently Malinin and Gales (2020) showed the success of using an ensemble of models for machine translation, and in particular utilising accurate confidence predictions for analysing translation quality.

3 Calibration Techniques

In this section we explain the details of three calibration techniques that we apply to dialogue belief tracking.

3.1 Loss Functions

The loss function can have a great impact on the calibration and accuracy of models. The most commonly used loss function in belief tracking is the standard softmax cross entropy loss. However, it tends to cause overconfident predictions where most of the probability is placed on the top class.

Label smoothing cross entropy (Szegedy et al., 2016) aims to resolve this problem by replacing the one-hot targets of cross entropy with a smoothed target distribution. That is, for label y_i and smoothing parameter $\alpha \in \left(0, \frac{1}{K}\right]$, the target distribution will be:

$$t(c|\alpha, y_i) = \begin{cases} 1-(K-1)\alpha & c = y_i, \\ \alpha & \text{otherwise}, \end{cases} \quad (1)$$

where K is the number of possible values of c. The loss for a model with parameters $\boldsymbol{\theta}$ and a set of N output logits $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2, ..., \hat{\mathbf{z}}_N$ with true labels $y_1, y_2, ..., y_N$ is defined as:

$$\mathcal{L}(\boldsymbol{\theta}, \alpha) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{KL}\left[\text{Softmax}(\hat{\mathbf{z}}_i)||t(c_i|\alpha, y_i)\right], \quad (2)$$

where $\mathbf{KL}$ is the Kullback–Leibler divergence between two distributions (Kullback and Leibler, 1951).

Alternatively, Bayesian matching loss (Joo et al., 2020) uses a Dirichlet distribution as the final activation function. The target is constructed using the Bayes rule, where we assume the observed label y_i to be an observation from a categorical distribution $y_i|\boldsymbol{\pi}_i \sim \text{Cat}(\boldsymbol{\pi}_i)$ and $\boldsymbol{\pi}_i$ is the true underlying distribution of the label. To introduce uncertainty into the target distribution we assume that the prior of $\boldsymbol{\pi}_i$ is a Dirichlet distribution, $\text{Dir}(\mathbf{1})$. In this way, we have a highly uncertain prior distribution. From this it can be shown that the posterior will be $\boldsymbol{\pi}_i|y_i \sim \text{Dir}(\mathbf{1} + \mathbf{I}(y_i))$, where $\mathbf{I}(y_i)$ is the one-hot representation of y_i. The loss function is then constructed using the negative log likelihood of the true label given the predicted distribution $\hat{\boldsymbol{\pi}}_i \sim \text{Dir}(\hat{\mathbf{z}}_i)$, penalised by the KL divergence from the the uncertain $\text{Dir}(\mathbf{1})$ distribution:

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) = \sum_{i=1}^{N}\{\lambda \mathbf{KL}\left[\hat{\boldsymbol{\pi}}_i||\text{Dir}(\mathbf{1})\right] - \mathbb{E}_{\hat{\boldsymbol{\pi}}_i}\left[\log(p(y_i|\hat{\boldsymbol{\pi}}_i))\right]\}, \quad (3)$$

where $\lambda > 0$ is the penalisation parameter.

3.2 Ensemble Distribution Estimation

From a Bayesian viewpoint, the probability of observing an outcome given the observed examples can be broken down into two components: the predictive distribution of the model and the posterior of the model given the observed examples. The posterior of the model given the data is an unknown distribution which can be estimated in various ways. One method is to use an ensemble of models, where the ensemble acts as an estimator for the posterior distribution of the parameters, $p(\boldsymbol{\theta}|\mathcal{D})$, where $\mathcal{D}$ represents the observed examples. Let $q(\boldsymbol{\theta})$ represent the distribution over all possible members of an ensemble. This distribution could be seen as the ensemble estimate of the posterior, $p(\boldsymbol{\theta}|\mathcal{D})$, (Malinin et al., 2019; Malinin and Gales, 2020). Hence,

$$\hat{p}(y|\mathbf{x}, \mathcal{D}) = \int p(y|\mathbf{x}, \boldsymbol{\theta}) q(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \qquad (4)$$

Since this integral is still intractable we need to estimate it using Monte Carlo. To sample from the ensemble distribution $q(\boldsymbol{\theta})$ we consider two approaches: using dropout during inference to collect an ensemble of N equally likely models (Gal and Ghahramani, 2016), or alternatively bootstrap sampling N equally likely subsets of the data to train N equally likely ensemble members. Let these N members be $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ..., \boldsymbol{\theta}^{(N)}\}$. The estimated predictive distribution can then be calculated as follows:

$$\hat{p}(y|\mathbf{x}, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} p(y|\mathbf{x}, \boldsymbol{\theta}^{(i)}) \qquad (5)$$

3.3 Temperature Scaling

Temperature scaling is a post-processing technique which scales the logits of the model by a scaling factor $\beta > 1$ (Guo et al., 2017), resulting in better-calibrated estimates. The temperature scaling parameter β can be trained on a development set.

4 Experimental Setup

We seek to build a well-calibrated dialogue belief tracker. For our baseline belief tracker, we use the SUMBT model architecture (Lee et al., 2019), which uses BERT (Devlin et al., 2018) as a turn encoder and multi-head attention for slot candidate matching. We perform all experiments on the MultiWOZ 2.1 dataset (Eric et al., 2019), the current standard dataset for multi-domain dialogue. When training using Bayesian matching, we use a scaling coefficient of $\lambda = 0.003$, and for label smoothing, a smoothing coefficient of $\alpha = 0.05$. For the ensemble belief tracker, we train 10 identical independent models, each with a sub-sample of 7500 dialogues. All hyper-parameters are obtained using a parameter search based on validation set performance. For all training, we use the BERT-base-uncased model from PyTorch Transformers (Wolf et al., 2019) for turn embedding. We use a gated recurrent unit with a hidden dimension 300 for latent tracking and Euclidean distance for value candidate scoring. During training, we use a learning rate of $5e-5$ in combination with a linear learning rate scheduler, the warm-up proportion is set to 0.1. A dropout rate of 0.3 is used, and training is performed for 100 epochs.[1]

5 Evaluation Metrics

5.1 Joint Goal Accuracy

The joint goal accuracy (JGA) is the percentage of turns for which the model predicts the complete user goal correctly. We further propose the introduction of an adjusted top 3 JGA, which considers a user goal prediction correct if the true label for each slot is among the top 3 predicted candidates for that slot in the belief state given there are at least 5 possible candidates.

5.2 L2 Norm Error

The L2 norm error is the L2 norm of the difference between the true labels and the predicted distributions. To form the user goals and belief states we concatenate all the slot labels and slot distributions. This error measure does not only consider the accuracy of the predictions but also the uncertainty.

5.3 Joint Goal Calibration Error

A well-calibrated model is one where the accuracy is aligned with the confidence predictions. The expected calibration error (ECE) evaluates the calibration by measuring the difference between the model's confidence and accuracy (Guo et al., 2017), meaning a lower ECE indicates better calibration. Hence:

$$\mathrm{ECE} = \sum_{k=1}^{B} \frac{b_k}{N} |\mathrm{acc}(k) - \mathrm{conf}(k)|, \qquad (6)$$

[1] Our code will be made available at https://gitlab.cs.uni-duesseldorf.de/general/dsml/calibrating-dialogue-belief-state-distributions.

where B is the number of bins, b_k are the bin sizes, N the number of observations, $\text{acc}(k)$ and $\text{conf}(k)$ the accuracy and confidence measures of bin k. We also propose an adapted ECE, called the expected joint goal calibration error (EJCE), which uses the joint goal accuracy for bin k as $\text{acc}(k)$, and the following metric as confidence:

$$\text{conf(k)} = \frac{1}{b_k} \sum_{i=1}^{b_k} \min_{s \in \text{slots}} \max_{v \in \text{values}} \hat{p}_i(v|s), \quad (7)$$

where $\hat{p}_i(v|s)$ is the predicted probability of value v for slot s given the i^{th} observation in bin k.

6 Results

Model	JGA	Top 3 JGA	EJCE
Cross entropy	46.78%	69.97%	1.996
Label smoothing	46.32%	74.57%	1.292
Bayesian matching	31.03%	45.16%	4.922
Temperature scaling			
Cross entropy (1.73*)	46.78%	69.97%	4.758
Label smoothing (1.00*)	46.32%	74.57%	1.292
Dropout ensembles			
Cross entropy (35**)	47.18%	71.14%	2.909
Label smoothing (35**)	46.36%	76.12%	2.217
Bootstrap model ensembles			
Label smoothing (10**)	**48.41%**	**84.08%**	**0.841**

Table 1: Calibration strategy performance. *temperature scaling coefficient **ensemble size.

Model	JGA	L2 Norm
SUMBT (Lee et al., 2019)	46.78%	1.1075
CE-BST (ours)	48.41%	**1.1041**
SOTA DST	$<$ **56.0%**	$>$ 1.2445

Table 2: MultiWOZ 2.1 performance.

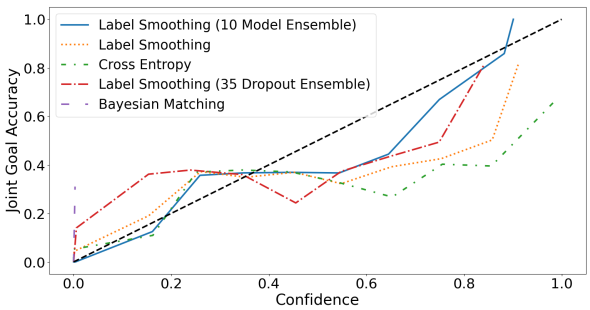

Figure 1: Reliability Diagram.

All of the calibration techniques presented above can be combined. Here, we focus on the most important combinations and present the results in Table 1. We make the following observations. First, cross entropy on its own leads to a high EJCE, as expected. Second, label smoothing reduces EJCE while leading to a negligible drop in accuracy. Third, Bayesian matching underperformed in our experiments, suggesting a difficulty in choosing the right priors. Fourth, temperature scaling is not an effective way of calibrating uncertainty, as the same calibration is applied to each observation. Finally, the ensemble methods produce very promising results for both accuracy and calibration of the model. In particular, if we look at the Top 3 JGA, our method achieves an improvement of 14.11 percentage points over the baseline, in the Appendix we include a comprehensive set of Top n JGA results. In Figure 1 we plot JGA as a function of confidence. The best calibrated model is the one that is closest to the diagonal, i.e. the one whose confidence for each dialogue state is closest to the achieved accuracy. From this reliability diagram we see that both the dropout and model ensembles improve model calibration and do not produce over-confident output as the cross entropy baseline does. In Table 2 we compare our model to some of the best performing belief and state tracking models. Here we see that we outperform the best performing **belief** tracker but the state-of-the-art (SOTA) **state** trackers (Heck et al., 2020; Chen et al., 2020; Hosseini-Asl et al., 2020) have a significantly higher JGA. However, when analysing the L2 norm[2] we see that the uncertainty estimates of **belief** tracking models compensate for the lower joint goal accuracy. This corroborates our premise that it is important to have well calibrated confidence estimates and not just a high JGA.

7 Conclusion

We applied a number of calibration techniques to a baseline dialogue belief tracker. We showed that a label smoothed trained ensemble provides state-of-the-art calibration of the belief state distributions and has the best accuracy among the available **belief** trackers. Although it does not compete with **state** trackers in terms of JGA, when considering top 3 predictions it achieves 84.08% accuracy

[2]For a model with a given JGA we can calculate the minimum L2 that such a model can possibly achieve by assuming that it never predicts more than one slot incorrectly.

(Top 3 JGA), almost 30 percentage points above state-of-the art state trackers. We also find that our model has the best L2 norm performance, which suggests that the quality of predicted uncertainty is as important as the average JGA.

It is important to note that the proposed calibration methods can be applied to any neural dialogue belief tracking method. The uncertainty estimates predicted by this model could improve the success of dialogue systems because this model can provide the dialogue manager with a good measure of confidence. This could allow the system to ask questions in moments of confusion. In the Appendix we include example dialogues to illustrate this. In future, we aim to combine the state-of-the-art dialogue state tracking and belief tracking methods to create a method that can achieve both states-of-the-art joint goal accuracy and well-calibrated belief states.

Acknowledgements

C. van Niekerk, M. Heck and N. Lubis are supported by funding provided by the Alexander von Humboldt Foundation in the framework of the Sofja Kovalevskaja Award endowed by the Federal Ministry of Education and Research, while C. Geishauser, H-C. Lin and M. Moresi are supported by funds from the European Research Council (ERC) provided under the Horizon 2020 research and innovation programme (Grant agreement No. STG2018 804636).

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. *AAAI 2020*.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained Transformers. *arXiv preprint arXiv:2003.07892*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur. 2019. MultiWOZ 2.1: Multi-domain dialogue state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Y Gal and Z Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on Machine Learning, ICML 2016*, volume 3, pages 1651–1660.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. TripPy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 35–44, 1st virtual meeting. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

Taejong Joo, Uijung Chung, and Min-Gwan Seo. 2020. Being Bayesian about categorical probability. *arXiv preprint arXiv:2002.07965*.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2019. Efficient dialogue state tracking by selectively overwriting memory. *arXiv preprint arXiv:1911.03906*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. SUMBT: slot-utterance matching for universal and scalable belief tracking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5478–5483.

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 1998. Using Markov decision process for learning dialogue strategies. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 1, pages 201–204. IEEE.

Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7047–7058.

Andrey Malinin and Mark Gales. 2020. Uncertainty in structured prediction. *arXiv preprint arXiv:2002.07650*.

Andrey Malinin, Bruno Mlodozeniec, and Mark Gales. 2019. Ensemble distribution distillation. *arXiv preprint arXiv:1905.00076*.

Angeliki Metallinou, Dan Bohus, and Jason D. Williams. 2013. Discriminative state tracking for spoken dialog systems. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), Sofia, Bulgaria*. Association for Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Elnaz Nouri and Ehsan Hosseini-Asl. 2018. Toward scalable neural dialogue state tracking model. *arXiv preprint arXiv:1812.00899*.

Osman Ramadan, Paweł Budzianowski, and Milica Gašić. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 432–437.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Jason D Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Jian-Guo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wan, Philip S Yu, Richard Socher, and Caiming Xiong. 2019. Find or classify? Dual strategy for slot-value predictions on multi-domain dialog state tracking. *arXiv preprint arXiv:1910.03544*.

A Appendices

A.1 Joint Goal Accuracy Analysis

In Table 3 we compare SUMBT and our CE-BST method using 5 different top n joint goal accuracy's.

Model	Top 1	Top 2	Top 3	Top 4	Top 5
SUMBT	46.78%	64.61%	69.97%	72.10%	73.70%
CE-BST	48.41%	77.25%	84.08%	85.84%	86.93%

Table 3: Top n joint goal accuracy comparison.

A.2 Example Dialogues

In Figures 2 - 9 we present some example dialogues together with an extract from their belief state distributions. These examples show situations where a well-calibrated belief state distribution could be beneficial for decision making.

User: I need a place to stay.
System: Sure. I'll need a little more information. Is there an area you are interested in?
User: No specific area. I would like it to be in the moderate price range and it should have free parking. I would also like it to have 4 stars.

Figure 2: Dialogue *PMUL3364* from the MultiWOZ 2.1 corpus.

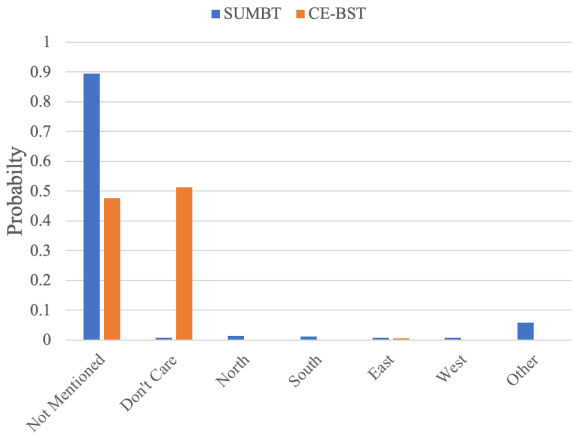

Figure 3: *PMUL3364* Hotel - Location belief state distribution.

User: Can you help me find a place to go in the centre?
System: I can help you with that. Is there a certain kind of attraction that you would like to visit?
User: Surprise me! Give me the postcode as well.
System: Would you prefer the castle galleries is a museum in the centre of town. Their post code is cb23bj.
User: Great! I am also looking for a place to eat in the same area. Something not too expensive, but not cheap.

Figure 4: Dialogue *PMUL4258* from the MultiWOZ 2.1 corpus.

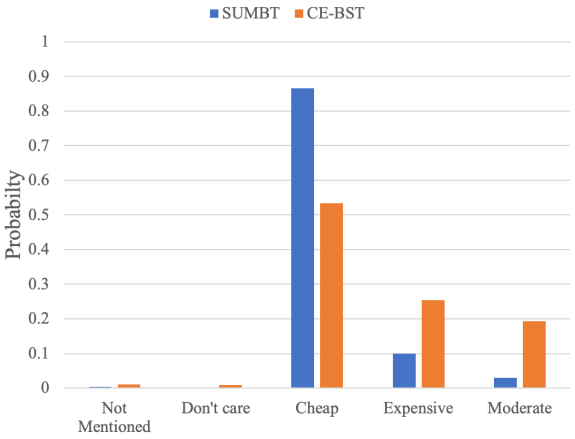

Figure 5: *PMUL4258* Restaurant - Price Range belief state distribution.

User: Hi, I am looking for a hotel by the name of Acorn guest house.
System: Sure, what would you like to know about it?
User: I would like to know if it is available for 8 people for 4 nights starting Saturday?

Figure 6: Dialogue *PMUL4605* from the MultiWOZ 2.1.

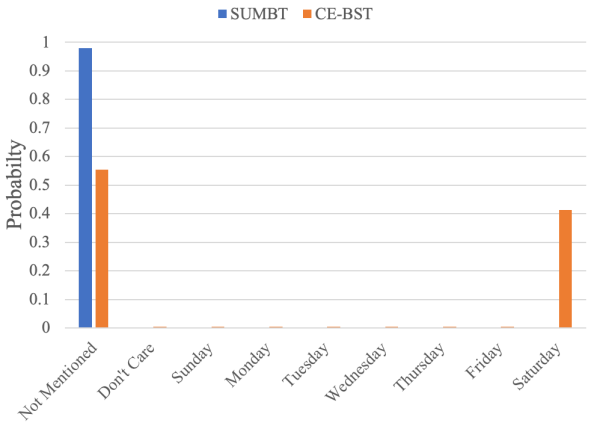

Figure 7: *PMUL4605* Hotel - Book Day belief state distribution.

User: I need a train to Stansted airport that leaves on Sunday.
System: Did you have a time you would like to arrive or leave?
User: I need to arrive by 14:30.
System: tr1668 will arrive at 14:08, would that work for you?
User: That is perfect. I would like to make a booking for 6 people please.
System: Booking was successful, the total fee is 48.48 gbp payable at the station. Your reference number is hf03ug02. Do you need assistance with anything else?
User: I need to eat too.
System: What type of restaurant and price range are you looking for?
User: I would like Catalan food. It needs to be in the centre and be expensive.

Figure 8: Dialogue *PMUL3625* from the MultiWOZ 2.1 corpus.

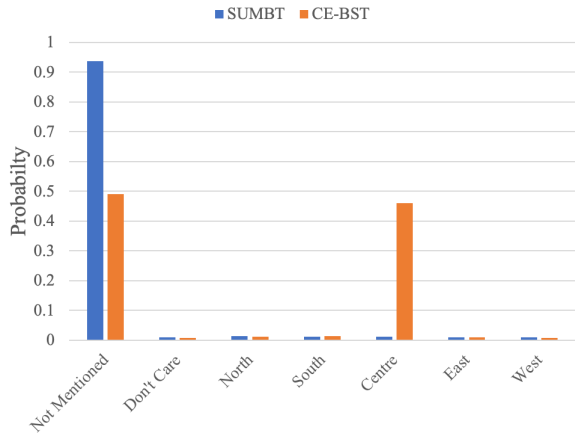

Figure 9: *PMUL3625* Restaurant - Location belief state distribution.