SUMBT: Slot-Utterance Matching for Universal and Scalable Belief Tracking

Hwaran Lee* **Jinsik Lee*** **Tae-Yoon Kim**
SK T-Brain, AI Center, SK telecom
{hwaran.lee, jinsik16.lee, oceanos}@sktbrain.com

Abstract

In goal-oriented dialog systems, belief trackers estimate the probability distribution of slot-values at every dialog turn. Previous neural approaches have modeled domain- and slot-dependent belief trackers, and have difficulty in adding new slot-values, resulting in lack of flexibility of domain ontology configurations. In this paper, we propose a new approach to universal and scalable belief tracker, called *slot-utterance matching belief tracker* (SUMBT). The model learns the relations between domain-slot-types and slot-values appearing in utterances through attention mechanisms based on contextual semantic vectors. Furthermore, the model predicts slot-value labels in a non-parametric way. From our experiments on two dialog corpora, WOZ 2.0 and MultiWOZ, the proposed model showed performance improvement in comparison with slot-dependent methods and achieved the state-of-the-art joint accuracy.

1 Introduction

As the prevalent use of conversational agents, goal-oriented systems have received increasing attention from both academia and industry. The goal-oriented systems help users to achieve goals such as making restaurant reservations or booking flights at the end of dialogs. As the dialog progresses, the system is required to update a distribution over dialog states which consist of users' intent, informable slots, and requestable slots. This is called belief tracking or dialog state tracking (DST). For instance, for a given domain and slot-types (e.g., 'restaurant' domain and 'food' slot-type), it estimates the probability of corresponding slot-value candidates (e.g., 'Korean' and 'Modern European') that are pre-defined in a domain ontology. Since the system uses the predicted outputs of DST to choose the next action based on a dialog policy, the accuracy of DST is crucial to improve the overall performance of the system. Moreover, dialog systems should be able to deal with newly added domains and slots[1] in a flexible manner, and thus developing scalable dialog state trackers is inevitable. Regarding to this, Chen et al. (2016) has proposed a model to capture relations from intent-utterance pairs for intent expansion.

Traditional statistical belief trackers (Henderson et al., 2014) are vulnerable to lexical and morphological variations because they depend on manually constructed semantic dictionaries. With the rise of deep learning approaches, several neural belief trackers (NBT) have been proposed and improved the performance by learning semantic neural representations of words (Mrkšić et al., 2017; Mrkšić and Vulić, 2018). However, the scalability still remains as a challenge; the previously proposed methods either individually model each domain and/or slot (Zhong et al., 2018; Ren et al., 2018; Goel et al., 2018) or have difficulty in adding new slot-values that are not defined in the ontology (Ramadan et al., 2018; Nouri and Hosseini-Asl, 2018).

In this paper, we focus on developing a "scalable" and "universal" belief tracker, whereby only a single belief tracker serves to handle any domain and slot-type. To tackle this problem, we propose a new approach, called *slot-utterance matching belief tracker* (SUMBT), which is a domain- and slot-independent belief tracker as shown in Figure 1. Inspired by machine reading comprehension techniques (Chen et al., 2017; Seo et al., 2017), SUMBT considers a domain-slot-

*Hwaran Lee and Jinsik Lee equally contributed to this work.

[1] For example, as reported by Kim et al. (2018), hundreds of new skills are added per week in personal assistant services.

type (e.g., 'restaurant-food') as a question and finds the corresponding slot-value in a pair of user and system utterances, assuming the desirable answer exists in the utterances. SUMBT encodes system and user utterances using recently proposed BERT (Devlin et al., 2018) which provides the contextualized semantic representation of sentences. Moreover, the domain-slot-types and slot-values are also literally encoded by BERT. Then SUMBT learns *the way where to attend* that is related to the domain-slot-type information among the utterance words based on their contextual semantic vectors. The model predicts the slot-value label in a non-parametric way based on a certain metric, which enables the model architecture not to structurally depend on domains and slot-types. Consequently, a single SUMBT can deal with any pair of domain-slot-type and slot-value, and also can utilize shared knowledge among multiple domains and slots.

We will experimentally demonstrate the efficacy of the proposing model on two goal-oriented dialog corpora: WOZ 2.0 and MultiWOZ. We will also qualitatively analyze how the model works. Our implementation is open-published.[2]

2 SUMBT

The proposed model consists of four parts as illustrated in Figure 1: BERT encoders for encoding slots, values, and utterances (the grey and blue boxes); a slot-utterance matching network (the red box); a belief tracker (the orange box); and a non-parametric discriminator (the dashed line on top).

2.1 Contextual Semantic Encoders

For sentence encoders, we employed a pre-trained BERT model (Devlin et al., 2018) which is a deep stack of bi-directional Transformer encoders. Rather than a static word vector, it provides effective contextual semantic word vectors. Moreover, it offers an aggregated representation of a word sequence such as a phrase and sentence, and therefore we can obtain an embedding vector of slot-types or slot-values that consist of multiple words.

The proposed method literally encodes words of domain-slot-types s and slot-values v_t at turn t as well as the system and user utterances. For the pair of system and user utterances, $\mathbf{x}_t^{sys} = (w_1^{sys}, ..., w_n^{sys})$ and $\mathbf{x}_t^{usr} = (w_1^{usr}, ..., w_m^{usr})$, the pre-trained BERT encodes each word w into a

[2] https://github.com/SKTBrain/SUMBT

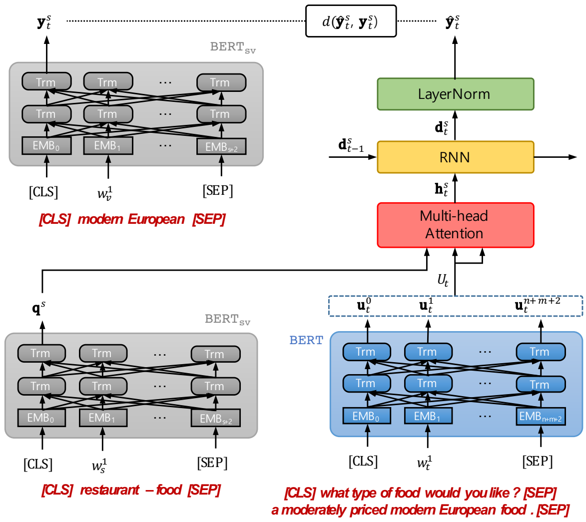

Figure 1: The architecture of slot-utterance matching belief tracker (SUMBT). An example of system and user utterances, a domain-slot-type, and a slot-value is denoted in red.

contextual semantic word vector $\mathbf{u}$, and the encoded utterances are represented in the following matrix representation:

$$U_t = \text{BERT}\left([\mathbf{x}_t^{sys}, \mathbf{x}_t^{usr}]\right). \quad (1)$$

Note that the sentence pairs are concatenated with a separation token [SEP], and BERT will be fine-tuned with the loss function (Eq. 7).

For the domain-slot-type s and slot-value v_t, another pre-trained BERT which is denoted as BERT$_{sv}$ encodes their word sequences $\mathbf{x}^s$ and $\mathbf{x}_t^v$ into contextual semantic vectors $\mathbf{q}^s$ and $\mathbf{y}_t^v$, respectively.

$$\begin{aligned} \mathbf{q}^s &= \text{BERT}_{sv}(\mathbf{x}^s), \\ \mathbf{y}_t^v &= \text{BERT}_{sv}(\mathbf{x}_t^v). \end{aligned} \quad (2)$$

We use the output vectors corresponding to the classification embedding token [CLS] that summarizes the whole input sequence.

Note that we consider $\mathbf{x}^s$ as a phrase of domain and slot words (e.g., $\mathbf{x}^s$ = "restaurant – price range") so that $\mathbf{q}^s$ represents both domain and slot information. Moreover, fixing the weights of BERT$_{sv}$ during training allows the model to maintain the encoded contextual vector of any new pairs of domain and slot-type. Hence, simply by forwarding them into the slot-value encoder, the proposed model can be scalable to the new domains and slots.

2.2 Slot-Utterance Matching

In order to retrieve the relevant information corresponding to the domain-slot-type from the utterances, the model uses an attention mechanism. Considering the encoded vector of the domain-slot-type $\mathbf{q}^s$ as a query, the model matches it to the contextual semantic vectors $\mathbf{u}$ at each word position, and then the attention scores are calculated.

Here, we employed multi-head attention (Vaswani et al., 2017) for the attention mechanism. The multi-head attention maps a query matrix Q, a key matrix K, and a value matrix V with different linear h projections, and then the scaled dot-product attention is performed on those matrices. The attended context vector $\mathbf{h}_t^s$ between the slot s and the utterances at t is

$$\mathbf{h}_t^s = \text{MultiHead}(Q, K, V), \tag{3}$$

where Q is Q^s and K and V are U_t.

2.3 Belief Tracker

As the conversation progresses, the belief state at each turn is determined by the previous dialog history and the current dialog turn. The flow of dialog can be modeled by RNNs such as LSTM and GRU, or Transformer decoders (i.e., left-to-right uni-directional Transformer).

In this work, the attended context vector $\mathbf{h}_t$ is fed into an RNN,

$$\mathbf{d}_t^s = \text{RNN}(\mathbf{d}_{t-1}^s, \mathbf{h}_t^s). \tag{4}$$

It learns to output a vector that is close to the target slot-value's semantic vector.

Since the output of BERT is normalized by layer normalization (Ba et al., 2016), the output of RNN $\mathbf{d}_t$ is also fed into a layer normalization layer to help training convergence,

$$\hat{\mathbf{y}}_t^s = \text{LayerNorm}(\mathbf{d}_t^s). \tag{5}$$

2.4 Training Criteria

The proposed model is trained to minimize the distance between outputs and target slot-value's semantic vectors under a certain distance metric. The probability distribution of a slot-value v_t is calculated as

$$p\left(v_t|\mathbf{x}_{\leq t}^{sys}, \mathbf{x}_{\leq t}^{usr}, s\right) = \frac{\exp\left(-d(\hat{\mathbf{y}}_t^s, \mathbf{y}_t^v)\right)}{\sum_{v' \in \mathcal{C}_s} \exp\left(-d(\hat{\mathbf{y}}_t^s, \mathbf{y}_t^{v'})\right)}, \tag{6}$$

where d is a distance metric such as Euclidean distance or negative cosine distance, and $\mathcal{C}_s$ is a set of the candidate slot-values of slot-type s which is defined in the ontology. This discriminative classifier is similar to the metric learning method proposed in Vinyals et al. (2016), but the distance metric is measured in the fixed space that BERT represents rather than in a trainable space.

Finally, the model is trained to minimize the log likelihood for all dialog turns t and slot-types $s \in \mathcal{D}$ as following:

$$\mathcal{L}(\theta) = -\sum_{s \in \mathcal{D}} \sum_{t=1}^{T} \log p(v_t|\mathbf{x}_{\leq t}^{sys}, \mathbf{x}_{\leq t}^{usr}, s). \tag{7}$$

By training all domain-slot-types together, the model can learn general relations between slot-types and slot-values, which helps to improve performance.

3 Experimental Setup

3.1 Datasets

To demonstrate the performance of our approach, we conducted experiments over WOZ 2.0 (Wen et al., 2017) and MultiWOZ (Budzianowski et al., 2018) datasets. WOZ 2.0 dataset[3] is a single 'restaurant reservation' domain, in which belief trackers estimate three slots (area, food, and price range). MultiWOZ dataset[4] is a multi-domain conversational corpus, in which the model has to estimate 35 slots of 7 domains.

3.2 Baselines

We designed three baseline models: `BERT+RNN`, `BERT+RNN+Ontology`, and a slot-dependent SUMBT. 1) The `BERT+RNN` consists of a contextual semantic encoder (`BERT`), an RNN-based belief tracker (`RNN`), and a linear layer followed by a softmax output layer for slot-value classification. The contextual semantic encoder in this model outputs aggregated output vectors like those of BERT_{sv}. 2) The `BERT+RNN+Ontology` consists of all components in the `BERT+RNN`, an ontology encoder (`Ontology`), and an ontology-utterance matching network which performs element-wise multiplications between the encoded ontology and

[3] Downloaded from https://github.com/nmrksic/neural-belief-tracker

[4] Downloaded from http://dialogue.mi.eng.cam.ac.uk/index.php/corpus. Before conducting experiments, we performed data cleansing such as correcting misspelled words.

utterances as in Ramadan et al. (2018). Note that two aforementioned models `BERT+RNN` and `BERT+RNN+Ontology` use the linear layer to transform a hidden vector to an output vector, which depends on a candidate slot-value list. In other words, the models require re-training if the ontology is changed, which implies that these models have lack of scalability. 3) The slot-dependent SUMBT has the same architecture with the proposed model, but the only difference is that the model is individually trained for each slot.

3.3 Configurations

We employed the pre-trained BERT model that has 12 layers of 784 hidden units and 12 self-attention heads.[5] We experimentally found the best configuration of hyper-parameters in which search space is denoted in the following braces. For slot and utterance matching, we used the multi-head attention with $\{4, 8\}$ heads and 784 hidden units. We employed a single-layer $\{\mathrm{GRU}, \mathrm{LSTM}\}$ with $\{100, 200, 300\}$ hidden units as the RNN belief tracker. For distance measure, both Euclidean and negative cosine distances were investigated. The model was trained with Adam optimizer in which learning rate linearly increased in the warm-up phase then linearly decreased. We set the warm-up proportion to be $\{0.01, 0.05, 0.1\}$ of $\{300, 500\}$ epochs and the learning rate to be $\{1 \times 10^{-5}, 5 \times 10^{-5}\}$. The training stopped early when the validation loss was not improved for 20 consecutive epochs. We report the mean and standard deviation of joint goal accuracies over 20 different random seeds. For reproducibility, we publish our PyTorch implementation code and the preprocessed MultiWOZ dataset.

4 Experimental Results

4.1 Joint Accuracy Performance

The experimental results on WOZ 2.0 corpus are presented in Table 1. The joint accuracy of SUMBT is compared with those of the baseline models that are described in Section 3.2 as well as previously proposed models. The models incorporating the contextual semantic encoder BERT beat all previous models. Furthermore, the three baseline models, `BERT+RNN`, `BERT+RNN+Ontology`, and the slot-dependent SUMBT, showed no significant performance differences. On the other hand, the slot-independent SUMBT which learned the shared information from all across domains and slots significantly outperformed those baselines, resulting in 91.0% joint accuracy. This implies the importance of utilizing common knowledge through a single model.

Table 2 shows the experimental results of the slot-independent SUMBT model on MultiWOZ corpus. Note that MultiWOZ has more domains and slots to be learned than WOZ 2.0 corpus. The SUMBT greatly surpassed the performances of previous approaches by yielding 42.4% joint accuracy. The proposed model achieved state-of-the-art performance in both WOZ 2.0 and MultiWOZ datasets.

4.2 Attention Weights Analysis

Figure 2 shows an example of attention weights as a dialog progresses. We can find that the model attends to the part of utterances which are semantically related to the given slots, even though the slot-value labels are not expressed in the lexically same way. For example, in case of ‘price range’ slot-type at the first turn, the slot-value label is ‘moderate’ but the attention weights are relatively

Model	Joint Accuracy
NBT-DNN (Mrkšić et al., 2017)	0.844
BT-CNN (Ramadan et al., 2018)	0.855
GLAD (Zhong et al., 2018)	0.881
GCE (Nouri and Hosseini-Asl, 2018)	0.885
StateNetPSI (Ren et al., 2018)	0.889
BERT+RNN (baseline 1)	0.892 ($\pm$0.011)
BERT+RNN+Ontology (baseline 2)	0.893 ($\pm$0.013)
Slot-dependent SUMBT (baseline 3)	0.891 ($\pm$0.010)
Slot-independent SUMBT (proposed)	**0.910** ($\pm$0.010)

Table 1: Joint goal accuracy on the evaluation dataset of WOZ 2.0 corpus.

Model	Joint Accuracy
Benchmark baseline [6]	0.2583
GLAD (Zhong et al., 2018)	0.3557
GCE (Nouri and Hosseini-Asl, 2018)	0.3558
SUMBT	**0.4240** ($\pm$0.0187)

Table 2: Joint goal accuracy on the evaluation dataset of MultiWOZ corpus.

[5] The pretrained model is published in https://github.com/huggingface/pytorch-pretrained-BERT

[6] The benchmark baseline is the model proposed in Ramadan et al. (2018) and the performance is described in http://dialogue.mi.eng.cam.ac.uk/index.php/corpus/

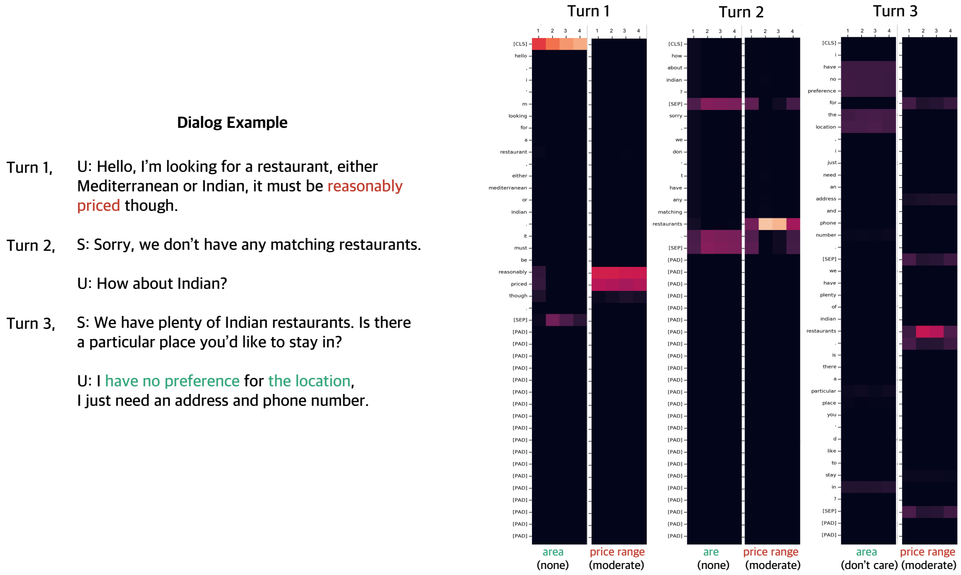

Figure 2: Attention visualizations of the first three turns in a dialog (WOZ 2.0). At each turn, the first and second columns are the attention weights when the given slots are 'area' and 'price range', respectively. The slot-value labels are denoted in the parentheses.

high on the phrase 'reasonably priced'. When appropriate slot-values corresponding to the given slot-type are absent (i.e., the label is 'none'), the model attends to [CLS] or [SEP] tokens.

5 Conclusion

In this paper, we propose a new approach to universal and scalable belief tracker, called SUMBT which attends to words in utterances that are relevant to a given domain-slot-type. Besides, the contextual semantic encoders and the non-parametric discriminator enable a single SUMBT to deal with multiple domains and slot-types without increasing model size. The proposed model achieved the state-of-the-art joint accuracy performance in WOZ 2.0 and MultiWOZ corpora. Furthermore, we experimentally showed that sharing knowledge by learning from multiple domain data helps to improve performance. As future work, we plan to explore whether SUMBT can continually learn new knowledge when domain ontology is updated.

Acknowledgements

We would like to thank Jinyoung Yeo and anonymous reviewers for their constructive feedback and helpful discussions. We are also grateful to SK T-Brain Meta AI team for GPU cluster supports to conduct massive experiments.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint*, arXiv:1607.06450.

Paweł Budzianowski, Tsung Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, 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*, pages 5016–5026. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1870–1879. Association for Computational Linguistics.

Yun-Nung Chen, Dilek Hakkani-Tür, and Xiaodong He. 2016. Zero-shot learning of intent embeddings for expansion by convolutional deep structured semantic models. In *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6045–6049. IEEE.

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.

Rahul Goel, Shachi Paul, Tagyoung Chung, Jeremie Lecomte, Arindam Mandal, and Dilek Hakkani-Tur. 2018. Flexible and scalable state tracking framework for goal-oriented dialogue systems. *arXiv preprint*, arXiv:1811.12891.

Matthew Henderson, Blaise Thomson, and Steve Young. 2014. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the SIGDIAL 2014 Conference*, pages 292–299. Association for Computational Linguistics.

Young-Bum Kim, Dongchan Kim, Anjishnu Kumar, and Ruhi Sarikaya. 2018. Efficient large-scale neural domain classification with personalized attention. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 2214–2224. Association for Computational Linguistics.

Nikola Mrkšić and Ivan Vulić. 2018. Fully statistical neural belief tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 108–113. 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*, pages 1777–1788. 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 (Short Papers)*, pages 432–437. Association for Computational Linguistics.

Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. 2018. Towards universal dialogue state tracking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2786. Association for Computational Linguistics.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations (ICLR)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M. Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449. Association for Computational Linguistics.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive dialogue state tracker. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 1458–1467. Association for Computational Linguistics.