Energy-based Neural Modelling for Large-Scale Multiple Domain Dialogue State Tracking

Anh Duong Trinh [†], **Robert J. Ross** [†], **John D. Kelleher** [‡]
ADAPT Centre
[†] School of Computer Science
[‡] Information, Communications & Entertainment Institute
Technological University Dublin, Ireland
{anhduong.trinh, robert.ross, john.d.kelleher}@tudublin.ie

Abstract

Scaling up dialogue state tracking to multiple domains is challenging due to the growth in the number of variables being tracked. Furthermore, dialog state tracking models do not yet explicitly make use of relationships between dialogue variables, such as slots across domains. We propose using energy-based structure prediction methods for large-scale dialogue state tracking task in two multiple domain dialogue datasets. Our results indicate that: (i) modelling variable dependencies yields better results; and (ii) the structured prediction output aligns with the dialogue slot-value constraint principles. This leads to promising directions to improve state-of-the-art models by incorporating variable dependencies into their prediction process.

1 Introduction

Task-oriented dialogue systems have been developed to assist users in many fields (Brixey et al., 2017; Zhao et al., 2019). In recent years it is a rising trend to scale-up task-oriented dialogue systems from single domain to multiple domains to improve the generalisability of models and support transfer of knowledge across domains. This leads to a new challenge in handling dialogues in the multi-domain context, that in turns increases the work load of the dialogue manager, and in particular the dialogue state tracking component. On the other hand, a number of works have demonstrated the benefit of processing multiple domains, for example it has been shown that such models yield better performances across domains in comparison with single domain trackers constructed and trained with the same approach (Mrksic et al., 2015).

Dialogue state tracking in task-oriented dialogue systems frequently uses a multi-slot representation for the dialogue state, thus casting the task as a multi-task classification problem. In these scenarios, an increase in the number of domains is equivalent to an increase in the number of slots, this in turn enlarges the models and makes the task more challenging. While traditionally one can develop a number of models to track dialogue states in each domain separately, recent advanced techniques tend to train dialogue state trackers in the multi-domain environment. Such multi-domain trackers produce state-of-the-art results (Kim et al., 2020; Heck et al., 2020).

To date state-of-the-art dialogue state trackers have treated the task as a set of individual domain-dependent classification problems (Heck et al., 2020; Wu et al., 2019; Zhou and Small, 2019). However, we argue that such approaches leave room for improvement; particularly with the consideration of the nature of human-machine interactions (Landragin, 2013). Specially, we argue that the multi-task classification methodology usually does not take into account the relationships between dialogue slot variables, despite the fact that these factors can play an essential part in the dialogue state prediction (Trinh et al., 2019a). Therefore, we propose to explicitly incorporate dialogue variable associations into the prediction process in a multi-domain dialogue environment, thus casting the dialogue state tracking task a structured prediction problem.

In this paper we demonstrate the manner, in which dialogue variable dependencies make an impact on the dialogue state tracking process in a multiple domain context. We choose two newly published multiple domain datasets, MultiWOZ 2.0 (Budzianowski et al., 2018) and MultiWOZ 2.1 (Eric et al., 2019), to conduct our study. These datasets contain a large number of dialogues across several different domains, thus they are practical for our study. Our investigation is detailed in three stages:

Proceedings of 4th Workshop on Structured Prediction for NLP, pages 33–42

- **Data analysis** – It is important to clearly determine whether variable dependencies exist in dialogue data, and to what extent they present in dialogue states. These questions can be solved by performing statistical tests on dialogue data (Trinh et al., 2019c).

- **Model development** – Since we treat the dialogue state tracking task as a structured prediction problem, we develop an energy-based tracking model for the task, where the energy-based learning methodology has been found effective in handling variable dependencies (Trinh et al., 2019b).

- **Evaluation & Analysis** – We evaluate the performance of our energy-based model and benchmark it against state-of-the-art trackers. Furthermore, we conduct an analysis study on the effectiveness of dialogue variable dependencies on the dialogue state tracking process in comparison with a multi-task deep learning method.

To the best of our knowledge, there have been structured prediction models developed for dialogue state tracking in single domains, but no work has been performed for multiple domains. On the other hand, several multi-domain dialogue state trackers study the topic of variable dependencies to some extent, but do not provide a detailed analysis on this phenomenon. Therefore, the contributions of our work are two-fold: (i) a large-scale structured prediction model for multi-domain dialogue state tracking; and (ii) a systematic analysis of variable dependencies across dialogue slots and domains.

The work presented in this paper is an empirical research of our previous work on capturing variable dependencies in dialogue states within single dialogue domains (Trinh et al., 2019a,b). We demonstrate that the energy-based method has good generalisability when applied to track dialogue states in multiple domain settings.

2 Variable Associations in Multi-Domain Dialogue

There are a number of works that to some extent have studied the variable associations in dialogue data in both single and multiple domain contexts. Single-domain dialogue variable dependencies were explicitly studied in the work by Trinh et al. (2019c,a). The associations between slots in single domain dialogue data are demonstrated to be beneficial factors for dialogue state tracking, and structured prediction approaches such as energy-based learning are effective in studying this phenomenon. On the other hand, although there has been no explicit study on variable dependencies in multiple domain dialogue data, we can indirectly infer the benefit of modelling such dependencies. Mrksic et al. (2015) show that shared models across dialogue domains yield better results than their domain-specific counterparts. Similarly in the TRADE model, Wu et al. (2019) highlighted the correlations between domains by training the base model on all of the domains except one, then fine-tuning on the remaining domain.

Since we focus on multiple domain dialogue state tracking, we conduct our study on MultiWOZ 2.0 (Budzianowski et al., 2018) and MultiWOZ 2.1 (Eric et al., 2019), two novel chat-based multi-domain dialogue datasets. We perform statistical tests on the dialogue data, and present the data analysis results in Figure 1. The statistical tests are Pearson's chi-squared test, which is useful for detecting pairwise dependencies between variables, and the chi-square test-based Cramer's V measurement, that measures the dependency strength once confirmed (Trinh et al., 2019c). In Figure 1 we present the heatmap of measured Cramer's V between all slot pairs in MultiWOZ 2.1 dataset, since this dataset contains manually fixed labels based on MultiWOZ 2.0 data.

The analysis explicitly confirms the variable dependencies in the multiple domain dialogue data, where pairwise statistical significance coefficient $\mathbf{p < 0.05}$ for all slot pairs. These dependencies exist on both slot and domain levels. Our analysis results also align to some extent with the cosine similarity of slot embedding presented in the TRADE model (Wu et al., 2019).

3 Energy-based Learning Dialogue State Tracking

Energy-based learning (LeCun et al., 2006) is an approach to structured prediction that can be used to account for variable dependencies in a supervised learning process. The core concept of the approach is to represent the associations of all variables in the system with a scalar value called *energy*, and to train an *energy function* that assigns low energy values to valid combinations of variables. There

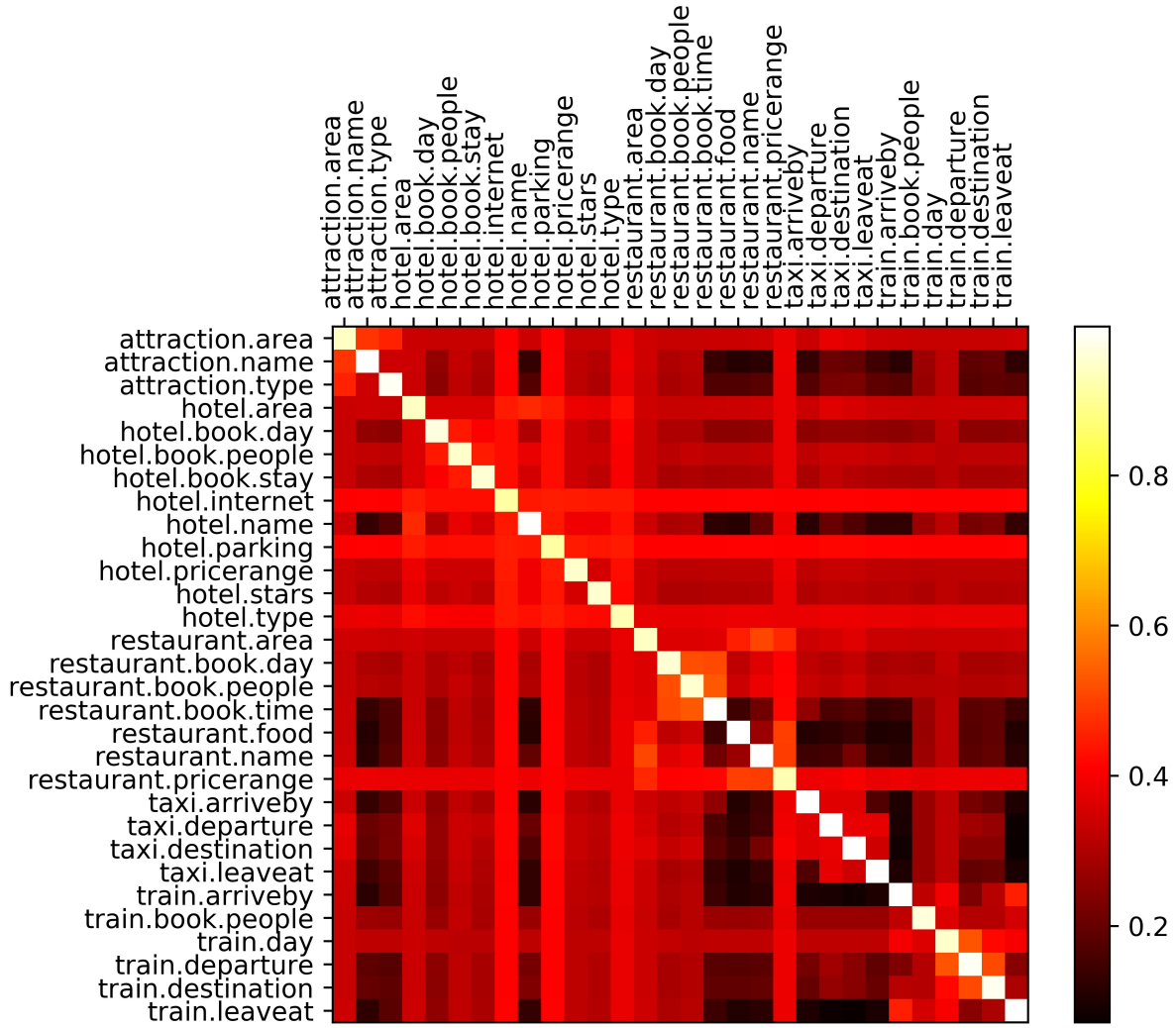

Figure 1: Cramer's V assessment of variable dependencies in MultiWOZ 2.1 data

are two key functions in the energy-based dialogue state tracker that we have developed:

- **Feature function** $F(X)$ – As a first step, we transform raw data into a distributed representation; this can be done with advanced techniques such as combinations of embedding and recurrent neural networks (Kelleher, 2019). The feature function can be either pretrained separately as an auxiliary task or jointly trained with the energy function.

- **Energy function** $E(F(X), Y)$ – The energy function is designed to capture variable dependencies and present them via a scalar value called *energy*. In our work, we develop the energy function with a deep learning architecture called Structured Prediction Energy Networks (SPEN) (Belanger and McCallum, 2016) to capture the dependencies between input and output variables, as well as among output variables.

The working mechanism of an energy-based model is different from a standard feedforward deep learning model:

- **Learning process** – During the learning process the energy function is typically trained to assign lower energy values to correct variable configurations, i.e. the desired output can be predicted with the minimal energy value with respect to our input. In our work we adopt a variant of the learning strategy detailed for the Deep Value Networks (DVN) architecture (Gygli et al., 2017) for this task.

- **Inference process** – Since in the energy-based learning methodology the energy function is trained to be an estimator for the good-

ness of fit between variables in the system, the output variables cannot be predicted in a straight forward manner. Therefore, we perform multiple inference loops guided by the gradient of the energy surface to find a set of labels for a given input using the trained energy function.

3.1 Hierarchical Recurrent Neural Feature Network

Task-oriented dialogues consist of multiple turns, where each turn contains machine and user actions. In the MultiWOZ datasets these actions are presented in a sentence format instead of dialogue act semantic representations. To accommodate the structure of multiple domain dialogue data, we make use of a multi-task LSTM-based dialogue state encoder (Trinh et al., 2018). In the description below we denote dialogue input data X, and the multi-task LSTM network $F(X)$. The architecture of our feature network is visualised in Figure 2.

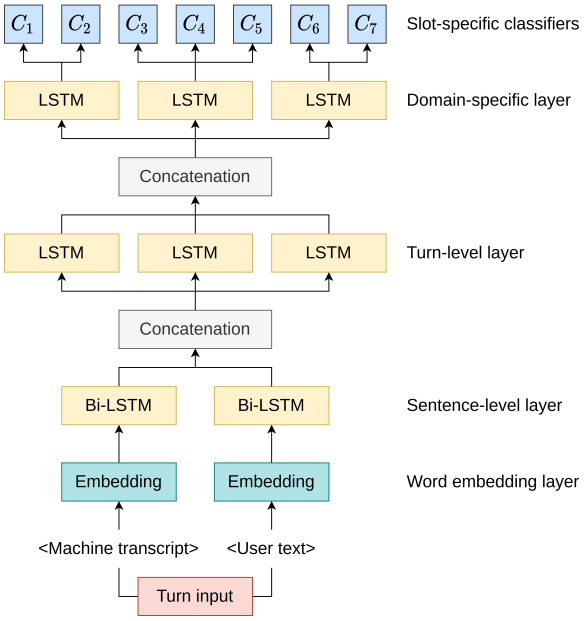

Figure 2: Multi-task Recurrent Neural Feature Network for MultiWOZ datasets. All recurrent units are LSTM (Hochreiter and Schmidhuber, 1997).

Our LSTM-based feature network consists of 5 layers:

- Word embedding layer – The word embedding layer is trained from scratch due to the small vocabulary present in the data.
- Sentence-level LSTM layer – To transform the sentence into vector representations, we make use of bidirectional LSTM structure (Bi-LSTM) (Schuster and Paliwal, 1997). In this layer machine and user transcripts are processed with separate Bi-LSTM cells, then their output vectors are concatenated before being fed into the next layer.
- Turn-level LSTM layer – A number of unidirectional LSTM cells are used to roll out the dialogue by turns. As highlighted in an earlier multi-task LSTM-based model (Trinh et al., 2018), using a number of LSTM cells can extract more useful information. The output of all the LSTM cells is concatenated into joint vectors, and treated as dialogue turn representations.
- Domain-specific LSTM layer – For each domain in the data we assign one LSTM cell to specialise the information downstream from the overall dialogue to the domain level.
- Slot-specific classifiers – The output layer consists of a number of slot-specific classifiers. Each classifier produces the prediction of the slot it corresponds to with a *softmax* activation function.

We pretrain this feature network $F(X)$ following the method as highlighted in a number of works on energy-based learning (Belanger and McCallum, 2016; Trinh et al., 2019b). It should be noted that the dialogue features can be extracted as the output of either the turn-level layer or the domain-specific layer. From our experiments, we have observed that the domain-specific LSTM layer produces more meaningful representations, thus it is more beneficial to pass on the energy function.

3.2 Deep Learning Energy Network

Since we focus on studying the variable dependencies between slots, our energy function must include the term for this phenomenon explicitly. We base the design of our energy network on the concept of Structured Prediction Energy Networks (SPEN) (Belanger and McCallum, 2016). The SPEN network is developed as a deep learning architecture to define an energy function that includes two individual energy terms, *local energy* and *global energy*:

$$E(F(X), Y) = E_{local}(F(X), Y) + E_{global}(Y) \quad (1)$$

Local energy is computed between input and output (label) variables, and is intended to capture the agreement between feature representations and labels:

$$E_{local}(F(X), Y) = \sum_{i=1}^{L} y_i W_i^\top X \tag{2}$$

where W is the set of trainable parameters, $Y = \{y_i\}^L$ is a label vector, and L is the number of label classes.

Global energy meanwhile is the energy term that captures the relationship between labels independently of the input features:

$$E_{global}(Y) = W_{g2}^\top f(W_{g1}^\top Y) \tag{3}$$

where weights W_{g1} and W_{g2} are trainable parameters, and $f(\cdot)$ is a non-linear function.

3.3 Learning Process

The purpose of the learning process is to train the energy function to measure the goodness of fit between variables correctly. It is important to design a suitable objective function to ensure that the energy function is well trained (Trinh et al., 2020).

For multi-label classification tasks, F_1 measurement is a common evaluation metric. In our structured dialogue state tracking task we make use of the F_1 metric for continuous variables, and interpret it as the ground truth energy:

$$E^*_{F_1}(Y, Y^*) = \frac{2\sum_i y_i y_i^*}{\sum_i y_i + \sum_i y_i^*} \tag{4}$$

where Y is the predicted labels, and Y^* is the ground truth labels.

Since the ground truth energy is calculated with our F_1 measurement, its value can only fall into the range $[0, 1]$. Therefore, it is appropriate to use a cross entropy function as the loss function between predicted and ground truth energies:

$$L(E, E^*_{F_1}) = -E^*_{F_1} \log E - (1 - E^*_{F_1}) \log(1 - E) \tag{5}$$

where $E = E(F(X), Y)$ is the predicted energy, and $E^*_{F_1} = E^*_{F_1}(Y, Y^*)$ is the ground truth energy.

There exist slot-value constraint rules in the task-oriented dialogue state tracking task such that at any time in the conversation each slot can be classified with not more than one value. However, multi-label classification methods do not include a mechanism to control the output prediction following these rules. Therefore we introduce a regularisation term to encourage our energy-based tracker to shape the output into the desired format:

$$R(Y, Y^*) = \left(\frac{\sum_i y_i - \sum_i y_i^*}{\sum_i y_i^*} \right)^2 \tag{6}$$

where Y is the predicted output, and Y^* is the ground truth labels.

Our final objective function including the label regularisation term for the learning process of the energy network is formulated as follow:

$$\mathcal{L} = L(E, E^*_{F_1}) + \alpha R(Y, Y^*) \tag{7}$$

where α is a regularisation coefficient.

This learning process is visualised in Fig. 3.

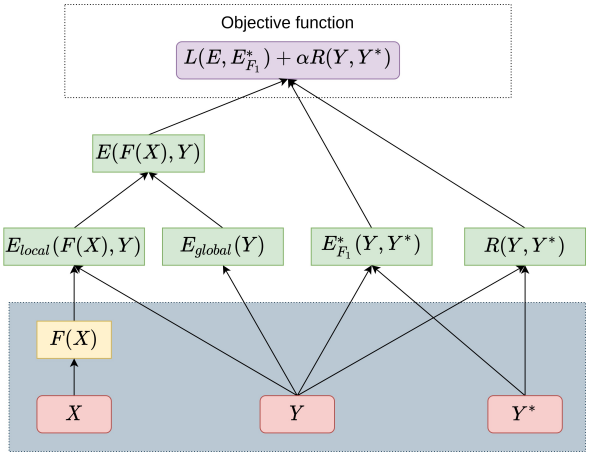

Figure 3: The learning process of our energy-based dialogue state tracker. The grey area denotes a frozen network where the parameters have been pretrained.

3.4 Inference Process

The energy function, as described above, can be interpreted as an estimator of the goodness of fit of the variables in the system. However, at prediction time we do not have the output variables that are an essential part of the energy formulation. Instead, to determine these values we perform a loopy inference process guided by the gradient of the energy surface.

We start with a random hypothesis and use gradient ascent to update the output hypothesis:

$$\begin{aligned} Y^{(0)} &= \{random(y_i)\}^L \\ Y^{(t+1)} &= \mathcal{P}_Y\left(Y^{(t)} + \eta \nabla_Y E(F(X), Y^{(t)})\right) \end{aligned} \tag{8}$$

where $\mathcal{P}_y$ is the projection operation to shape the predicted output to the output variable space $Y =$

$\{y_i\}^M \in \{[0,1]\}^M$, and η is the learning rate for gradient ascent.

Here, it should be noted that the energy function is an estimator for our F_1 measurement of the predicted output; thus, we aim to maximise the F_1 score to achieve the desired prediction:

$$E(F(X), Y^{(t)}) \sim E^*_{F_1}(Y^{(t)}, Y^*) \qquad (9)$$

4 Experiments

As indicated earlier, we have selected the multiple domain dialogue datasets, MultiWOZ 2.0 (Budzianowski et al., 2018) and MultiWOZ 2.1 (Eric et al., 2019), to conduct our study of variable dependencies. Since the MultiWOZ 2.0 dataset was known to contain a lot of labelling errors, the latter version MultiWOZ 2.1 was manually annotated to correct them. Each dataset contains more than 10000 dialogues across 7 domains, split into three subsets: train, development and test for training, validation and test purposes respectively.

However, following the common practice of other previous works we excluded two domains that rarely appear in the datasets. We followed the data processing and scoring scripts from the TRADE model (Wu et al., 2019) for our dialogue state tracking task.

Our experiments were conducted in two stages: first, we trained a multi-task learning network to extract dialogue features; then, we experimented on the energy-based learning level to explore inter-label dependencies.

Our model's hyperparameters are presented in Table 1.

We train both feature network and energy-based models with Adam optimiser (Kingma and Ba, 2015) for 300 epochs. To avoid the overfitting problem, we apply the early stopping technique and find that our models converge shortly after 200 epochs. We trained the feature network 3 times for each dataset, and selected the best model to extract features. The energy-based network was trained 5 times and the predictions were ensembled into the ultimate dialogue states for evaluation.

Hyper parameter	Value
Energy-based Network	
Word embedding size	300
LSTM number of turn-level cells	5
LSTM number of units	128
LSTM drop out	0.2
LSTM output activation	tanh
Energy non-linearity function $f(\cdot)$	tanh
Inference process	
Number of iterations	50
Inference learning rate	0.001
Learning process	
Objective function	Equation 7
Regularisation coefficient	0.01
Optimiser	Adam
Learning rate	0.001
Maximal global gradient norm	5.0

Table 1: Basic hyper parameters used in experiments constructing the energy-based dialogue state tracker.

5 Results & Discussion

We evaluate the performance of both our multi-task feature system and the energy-based tracker with an *Accuracy* metric as is common in dialogue state tracking. The results are reported in Table 2 alongside results of a number of state-of-the-art systems to our knowledge.

Overall, our energy-based dialogue state tracker yields competitive results in comparison to models that account for variable relationships using techniques such as attention mechanism (Kumar et al., 2020; Zhong et al., 2018) and transfer learning (Wu et al., 2019). When accounting for the variable dependencies with the energy-based method, we improve the belief state tracking results by large margins, i.e., 13.9% for MultiWOZ 2.0 and 18.1% for MultiWOZ 2.1. We believe that there are at least two reasons for this large improvement:

- High quality features are extracted from dialogue data due to the architecture of a hierarchical multi-task LSTM network. As we extract input features from domain-specific LSTM cells, the features contain both dialogue information up to current turns as well as domain information.

- The associations between variables, in particular label dependencies, are accounted for explicitly; hence more information is available for the classification of each slot than would be available in a straightforward multi-task classification process.

While the energy-based system does not achieve the state-of-the-art performance, it should be noted that state-of-the-art systems currently employ a

Model	MultiWOZ 2.0	MultiWOZ 2.1
TripPy (Heck et al., 2020)	-	**0.553**
Schema-guided (Chen et al., 2020)	0.512	0.552
DST-Picklist (Zhang et al., 2019)	-	0.533
SOM-DST (Kim et al., 2020)	**0.517**	0.530
MA-DST (Kumar et al., 2020)	-	0.519
DSTQA (Zhou and Small, 2019)	0.514	0.512
COMER (Ren et al., 2019)	0.488	-
TRADE (Wu et al., 2019)	0.486	0.456
HyST (Goel et al., 2019)	0.442	-
Neural reading (Gao et al., 2019)	0.411	-
GCE (Nouri and Hosseini-Asl, 2018)	0.363	-
GLAD (Zhong et al., 2018)	0.356	-
Our work		
Energy-based system	0.488	0.547
Multi-task feature system	0.349	0.366

Table 2: Performances of state-of-the-art and presented dialogue state tracking systems on MultiWOZ 2.0 & 2.1 data. The results for belief states are reported with the Accuracy metric.

very wide variety of modelling techniques while the currently presented work focuses on the addition of a mechanism to guide final labelling. For example, TripPy (Heck et al., 2020), which achieves the highest accuracy in MultiWOZ 2.1 data, is based on span-prediction and a number of memory mechanisms. Meanwhile, SOM-DST (Kim et al., 2020) improves the dialogue state tracking efficiency with a selectively overwriting memory mechanism. Both of these however do not explicitly look at the variable dependencies as potentially useful factors of dialogue states. The practical use of the energy-based learning method may lie in its use to fine tune results to take into account variable dependencies. Given the fact that the energy-based model is developed separately from the feature network, we can apply it to state-of-the-art models to investigate the effectiveness of variable dependencies in different situations.

One final observation with respect to the results is differences in performance across MultiWOZ 2.0 and 2.1 datasets. Even though the labels in MultiWOZ 2.1 dataset are corrected with manual labour, meaning the data is less noisy than the MultiWOZ 2.0 data, not all systems yield better results in MultiWOZ 2.1 than in MultiWOZ 2.0, e.g., models such as TRADE (Wu et al., 2019) and DSTQA (Zhou and Small, 2019) perform better with the original noisy data. In contrast, we observe that other state-of-the-art systems including our energy-based tracker perform better with cleaner data (MultiWOZ 2.1); this is of course a common phenomenon in supervised learning.

5.1 Variable Dependencies Analysis

In term of accuracy score our energy-based tracker outperforms the multi-task feature system by a large margins. However, the accuracy metric does not in itself verify the system's ability to capture variable dependencies. In order to evaluate the effectiveness of the energy-based learning method in capturing variable dependencies, we conduct an analysis on the performance of our trackers on the MultiWOZ 2.1 test set. Specifically, we analyse pairwise variable dependencies with Pearson's chi-squared test and measure their strength with Cramer's V coefficient as detailed earlier in Section 2. We present the results of variable association analysis between a number of slots in Table 3 with respect to test labels, labels produced by the Energy-based Tracker and labels produced by our Multi-Task Learning tracker. Here, we only show the dependencies between a subset of the slots purely for space reasons. If we were to show more or all of them, the table wouldn't fit in the template. We have, however, done the analysis of the dependencies for other slots and the results indicate that the other slots have similar tendencies, and more importantly that the data we present is representative of this more general pattern.

		hotel price range	restaurant price range	taxi departure	taxi destination	train departure	train destination
Test label							
attraction	area	0.200	0.236	0.272	0.276	0.107	0.089
hotel	area	0.225	0.315	0.218	0.218	0.094	0.078
restaurant	area	0.214	0.411	0.254	0.286	0.093	0.107
Energy-based tracker							
attraction	area	0.182	0.173	0.193	0.194	0.095	0.096
hotel	area	0.236	0.336	0.199	0.199	0.075	0.078
restaurant	area	0.256	0.419	0.254	0.321	0.120	0.109
Multi-task feature system							
attraction	area	0.291	0.194	0.153	0.151	0.086	0.084
hotel	area	0.147	0.232	0.149	0.160	0.055	0.056
restaurant	area	0.287	0.213	0.137	0.137	0.124	0.126

Table 3: Data analysis on variable dependencies in the performance of multi-task and energy-based trackers in MultiWOZ 2.1 data. The variable dependencies are reported with Cramer's V coefficient. In the table, the first block is variable dependencies in labels of the test set, while the second block is variable dependencies detected by our energy-based model, and the last block is the performance of the multi-task feature system.

The analysis results demonstrate that the energy-based tracker more consistently mirrors the association strengths seen in the test labels then does our baseline Multi-Task Learning approach. It is evidenced by smaller margins in Cramer's V coefficients between the *Energy-based tracker* and the *Test label* results than seen between the *Multi-task system* results and the *Test label* results[1]. There are, however, very few exceptions to this trend, namely the *attraction.area – restaurant.price range* and *attraction.area – train.destination* pairs where the multi-task based system has produced associations closer to the test label case than does the energy-based model.

Overall, we argue that the ability to capture variable dependencies between slots across dialogue domains explains the reason why the energy-based method outperforms the multi-task learning approach.

5.2 Slot-Value Constraint Analysis

Dialogue states of many task-oriented dialogue systems must satisfy a slot-value constraint principle that each slot must not have more than one value in the belief state of any turn. Specifically, the value of each informable slot can be either *none* if it is not mentioned by users, or a specific value, for example *Chinese* for the slot *food* in domain *restaurant* if information is provided by the user. While the underlying multi-task feature system follows this rule strictly due to the use of the output *softmax* activation function in slot-specific classifiers, the energy-based tracking model is not guaranteed to maintain this strict constraint.

To overcome this challenge, we proposed a label regularisation term (Equation 6) in the objective function detailed in Section 3.3. To evaluate the effectiveness of this mechanism, we conduct an additional analysis to determine the behaviours of our energy-based system based on this regularisation. This analysis is conducted in two stages:

- First, we train and evaluate our energy-based method on the dialogue data without the label regularisation term. Thus, the loss function (Equation 5) becomes our learning objective in this baseline case.

- Second, we set different threshold values, and calculate the proportion of correct predictions over the total number of dialogue turns that follow slot-value constraint rules with different thresholds. A value is considered activated if the predicted belief score of this value exceeds the threshold. This stage is conducted for our energy-based method both with and without the regularisation term.

[1]It should be noted that stronger associations do not necessarily indicate better tracking performance – our goal is to capture valid associations not to arbitrarily increase the number of associations seen in label outputs.

The slot-value constraint analysis is presented in Table 4.

Threshold	MultiWOZ 2.0		MultiWOZ 2.1	
	+Reg	–Reg	+Reg	–Reg
0.5	45.7	36.8	52.4	48.3
0.7	29.7	26.3	39.4	35.1
0.9	16.8	15.5	18.3	18.1

Table 4: Analysis of the impact of label regularisation on the energy-based dialogue state tracking on the MultiWOZ 2.0 & 2.1 data. The results are reported with the proportion (%) of correct predictions over the total number of dialogue turns that follow the slot-value constraint rules. +Reg/–Reg denotes the presence/absence of the label regularisation in the learning process.

The analysis result demonstrates that our energy-based systems with the label regularisation consistently outperforms those that do not include this term in the learning process with different belief score thresholds. Here, the label regularisation helps guide the system's prediction behaviour towards the requirement of the task-oriented domains. We can conclude that the impact of label regularisation on dialogue state tracking is systematic.

6 Conclusion

In this paper we demonstrated the effectiveness of applying the energy-based learning method to a large-scale dialogue state tracking task in multiple domains. We showed that the energy-based method is capable of capturing the dependencies between dialogue variables such as slots across domains, thus it improves the performance over a multi-task deep learning system significantly. Our analyses also showed that the structured prediction method can produce dialogue states that follow dialogue slot-value constraint rules in contrast with a multi-label classification method.

Although the results achieved with the energy-based method are competitive with published dialogue state tracking systems, they are not yet state of the art. There are several directions to investigate the further impact of an energy-based methodology on the dialogue state tracking task. One promising direction is the application of our energy-based method on top of an existing state-of-the-art systems to further improve that system's performance. Another direction is to refine the energy-based structure and investigate various strategies for the learning and inference processes to improve the ability to integrate captured dependencies into the structured prediction at a higher level. Furthermore our long term goal is to apply the structured learning approach in tracking different aspects of the conversations such as personality and preference as well as user intents.

Acknowledgements

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106 at the ADAPT SFI Research Centre at Technological University Dublin. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant No. 13/RC/2106.

References

David Belanger and Andrew McCallum. 2016. Structured Prediction Energy Networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48.

Jacqueline Brixey, Rens Hoegen, Wei Lan, Joshua Rusow, Karan Singla, Xusen Yin, Ron Artstein, and Anton Leuski. 2017. SHIHbot : A Facebook chatbot for Sexual Health Information on HIV / AIDS. In *Proceedings of the SIGDIAL 2017 Conference*, pages 370–373.

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 2018 Conference on Empirical Methods in Natural Language Processing*.

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. In *Association for the Advancement of Artificial Intelligence*.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Anuj Kumar Goyal, Peter Ku, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-tur. 2019. Dialog State Tracking: A Neural Reading Comprehension Approach. In *Proceedings of the SIGDial 2019 Conference*, pages 264–273.

Rahul Goel, Shachi Paul, and Dilek Hakkani-Tur. 2019. HyST: A Hybrid Approach for Flexible and Accurate Dialogue State Tracking. In *Proceedings of the INTERSPEECH 2019 Conference*.

Michael Gygli, Mohammad Norouzi, and Anelia Angelova. 2017. Deep Value Networks Learn to Evaluate and Iteratively Refine Structured Outputs. In *Proceedings of the 34th International Conference on Machine Learning*.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašić. 2020. TripPy: A Triple Copy Strategy for Value Independent Neural Dialog State Tracking. In *Proceedings of the SIGDial 2020 Conference*.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

John D Kelleher. 2019. *Deep Learning*. The MIT Press.

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2020. Efficient Dialogue State Tracking by Selectively Overwriting Memory. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics (ACL)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Adarsh Kumar, Peter Ku, Anuj Goyal, Angeliki Metallinou, and Dilek Hakkani-Tur. 2020. MA-DST: Multi-Attention-Based Scalable Dialog State Tracking. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020)*.

Frédéric Landragin. 2013. *Man-Machine Dialogue: Design and Challenges*. ISTE Ltd and John Wiley & Sons, Inc.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc' Aurelio Ranzato, and Fu Jie Huang. 2006. A Tutorial on Energy-Based Learning. *Predicting Structured Data*.

Nikola Mrksic, Diarmuid O'Seaghdha, Blaise Thomson, Milica Gasic, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2015. Multi-domain Dialog State Tracking using Recurrent Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 794–799.

Elnaz Nouri and Ehsan Hosseini-Asl. 2018. Toward Scalable Neural Dialogue State Tracking Model. In *Proceedings of the 2nd Conversational AI workshop, NeurIPS 2018*.

Liliang Ren, Jianmo Ni, and Julian McAuley. 2019. Scalable and Accurate Dialogue State Tracking via Hierarchical Sequence Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1876–1885.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Anh Duong Trinh, Robert J. Ross, and John D. Kelleher. 2018. A Multi-Task Approach to Incremental Dialogue State Tracking. In *Proceedings of The 22nd workshop on the Semantics and Pragmatics of Dialogue, SEMDIAL*, pages 132–145.

Anh Duong Trinh, Robert J. Ross, and John D. Kelleher. 2019a. Capturing Dialogue State Variable Dependencies with an Energy-based Neural Dialogue State Tracker. In *Proceedings of the SIGDial 2019 Conference*, pages 75–84.

Anh Duong Trinh, Robert J. Ross, and John D. Kelleher. 2019b. Energy-Based Modelling for Dialogue State Tracking. In *Proceedings of the 1st Workshop on NLP for Conversational AI*, pages 77–86.

Anh Duong Trinh, Robert J. Ross, and John D. Kelleher. 2019c. Investigating Variable Dependencies in Dialogue States. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue*, pages 195–197.

Anh Duong Trinh, Robert J. Ross, and John D. Kelleher. 2020. F-Measure Optimisation and Label Regularisation for Energy-Based Neural Dialogue State Tracking Models. In *Artificial Neural Networks and Machine Learning ICANN 2020*.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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.

Guoguang Zhao, Jianyu Zhao, Yang Li, Christoph Alt, Robert Schwarzenberg, Leonhard Hennig, Stefan Schaffer, Sven Schmeier, Changjian Hu, and Feiyu Xu. 2019. MOLI: Smart Conversation Agent for Mobile Customer Service. *Information (Switzerland)*, 10(2).

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-Locally Self-Attentive Dialogue State Tracker. In *Proceedings of the 56th Annual Meeting ofthe Association for Computational Linguistics*, pages 1458–1467.

Li Zhou and Kevin Small. 2019. Multi-domain Dialogue State Tracking as Dynamic Knowledge Graph Enhanced Question Answering.