2020-09-15

F-Measure Optimisation and Label Regularisation for Energy-Based Neural Dialogue State Tracking Models

Anh Duong Trinh
Technological University Dublin, anhduong.trinh@tudublin.ie

Robert J. Ross
Technological University Dublin, robert.ross@tudublin.ie

John D. Kelleher
Technological University Dublin, john.d.kelleher@tudublin.ie

Recommended Citation

Trinh A.D., Ross R.J., Kelleher J.D. (2020.) F-Measure Optimisation and Label Regularisation for Energy-Based Neural Dialogue State Tracking Models. *The 29th International Conference on Artificial Neural Networks, ICANN 2020*, Bratislava, Slovakia, September 15–18, 2020. doi:10.1007/978-3-030-61616-8_64

Funder: Science Foundation Ireland, European Regional Development Fund (ERDF)

F-Measure Optimisation and Label Regularisation for Energy-based Neural Dialogue State Tracking Models

Anh Duong Trinh[1,2 [0000−0003−4778−6357]], Robert J. Ross[1,2 [0000−0001−7088−273X]], and John D. Kelleher[1,3 [0000−0001−6462−3248]]

[1] ADAPT Centre
[2] School of Computer Science
[3] Information, Communications & Entertainment Institute
Technological University Dublin, Ireland
{anhduong.trinh,robert.ross,john.d.kelleher}@tudublin.ie

Abstract. In recent years many multi-label classification methods have exploited label dependencies to improve performance of classification tasks in various domains, hence casting the tasks to structured prediction problems. We argue that multi-label predictions do not always satisfy domain constraint restrictions. For example when the dialogue state tracking task in task-oriented dialogue domains is solved with multi-label classification approaches, slot-value constraint rules should be enforced following real conversation scenarios.
To address these issues we propose an energy-based neural model to solve the dialogue state tracking task as a structured prediction problem. Furthermore we propose two improvements over previous methods with respect to dialogue slot-value constraint rules: (i) redefining the estimation conditions for the energy network; (ii) regularising label predictions following the dialogue slot-value constraint rules. In our results we find that our extended energy-based neural dialogue state tracker yields better overall performance in term of prediction accuracy, and also behaves more naturally with respect to the conversational rules.

Keywords: Neural dialogue state tracking · Energy-based learning · F-measure optimisation · Label regularisation · Multi-label classification · Dialogue processing.

1 Introduction

Task-oriented dialogue systems have a wide range of applications in the modern technology world. The performance of dialogue systems depends directly on the performance of their dialogue state tracking (DST) components, that are responsible for maintaining meaningful dialogue representations including user intents and dialogue context. Although the dialogue state tracker plays an essential role, it is far from perfect due to various factors [20].

Task-oriented dialogue systems are typically restricted in specific closed domains, and cast dialogue states as sets of slot-value pairs. Within this setting the dialogue state tracking task can be interpreted as a multi-task classification problem, where tracking the value of each slot is by itself a classification task. This interpretation is applied for various public dialogue domains [3, 19].

To date various deep learning approaches have been proposed to tackle the dialogue state tracking problem as a combination of individual tasks [10, 17] or in a multi-task learning-based fashion [13]. Among multi-task learning-based approaches it is also common to treat the dialogue state tracking task as a multi-label classification problem [21]. While classic multi-label classification methods assume independence between class labels, more recent approaches tend to explore the role of label dependencies in the task, that casts the task itself as a structured prediction problem. From a practical point of view, structured prediction models have shown significant improvements in natural language processing. Particularly in the dialogue processing field, recent structured dialogue state trackers [14, 15] demonstrate that accounting for label associations can boost the performance when used to supplement a classic deep learning approach. In these models the label dependencies are captured via an energy function that is implemented with a deep learning architecture in the so-called energy-based learning methodology [8].

Despite the fact that a structured prediction methodology has already improved multi-label classification models' performance, we argue that there is still room for improvement. In particular, multi-label classifiers do not naturally enforce strict restrictions on dialogue states such that each slot has only one activated value at any time during the conversation. Such restrictions can be thought of as additional constraints on the structured prediction task. In order to investigate this phenomenon we propose a modelling approach to improve energy-based neural dialogue state trackers focusing on: (i) revising and redefining the energy-based estimation of the dialogue states; and (ii) applying slot-value constraint rules via label regularisation. Furthermore we conduct a detailed error analysis on the impact of the mentioned points on the structured prediction performance.

We proceed by introducing the domains in which we work in a more detail before going on to detail the specifics of our contributed model and subsequent analysis.

2 Task-oriented Dialogue Domains

The effectiveness of structured prediction is based on the assumption of dependencies between label classes. We base our work on the analysis of a series of well-known dialogue datasets that have moderate size and are limited to a single closed domain. The first two datasets we chose come from the second and the third dialogue state tracking challenge (DSTC2 & 3) competitions. The third dataset was created more recently with Wizard-of-Oz crowd-sourced data collection framework, and is named WOZ 2.0[4]. The details of these datasets are as follow:

- DSTC2 [3] is a restaurant information dataset that consists of spoken conversations. It includes 1612 dialogues for training, 506 dialogues for validation, and 1117 dialogues for testing. DSTC2 dialogue states consist of three subtasks: *Joint goals*, *Search methods*, and *Requested slots*; and among these the latter two have been solved with various machine learning and deep learning approaches. The most difficult task is *Joint goals*, which requires tracking the value of four informable slots:

[4] From here we simplify the name of this dataset to WOZ as in common practice.

food, *price range*, *area*, and *name*. However the slot *name* is omitted from many works due to the lack of its appearance in the whole dataset.
- DSTC3 [4] is a spoken dialogue dataset in the tourism information domain. It contains 2286 dialogues in a complete set. The dialogue states are defined in the same way as the DSTC2 challenge. In this work we solve the *Joint goals* task of only four informable slots, *food*, *price range*, *area*, and *type*, as we omit other slots due to their extremely low appearance frequency in the data.
- WOZ [19] is a chat-based restaurant information dataset that shares the same ontology as DSTC2. The WOZ dataset includes 1200 dialogues in total, split into the training, validation, and test sets with a ratio 3:1:2. The WOZ data is collected in a Wizard-of-Oz chat environment, therefore it is cleaner than DSTC2. Subsequently the tracking results can be much higher. The WOZ dialogue state tracking task requires capturing the *Joint goals* of three slots (*food*, *price range*, and *area*), as well as the *Requested Slots*.

In all domains we focus on the most challenging task, *Joint goals*, and study the impact of label dependencies between informable slots in the tracking process.

To verify that interlabel dependencies are present, we investigated label dependencies of DSTC2 & 3 data with Pearson's chi-square test and related measurements [16]. The statistical test analysis shows that there exist dependencies between dialogue slots. In this work we conduct similar statistical tests for the WOZ dataset. We observe the associations between WOZ data slots with the Cramer's V coefficient as follow: *food – price range* 0.316; *food – area* 0.302; *price range – area* 0.180. The analysis indicates that there exist dependencies between the WOZ slots.

3 Energy-based Dialogue State Tracker

Energy-based learning [8] focuses on exploiting the dependencies between different variables in the system. The classic concept of learning energy values is that the energy function should be trained to assign lower values to correct variable configurations, and higher energy for undesired variable configurations.

The architecture of an energy-based model is usually split into two components: a feature function $F(X)$, and an energy function $E(F(X), Y)$, where X and Y are input and output variables respectively. In the implementation of our energy-based dialogue state tracker we develop the feature function with a hierarchical recurrent neural network to transform dialogue data into meaningful high-dimensional representations, and the energy function with a deep learning network to estimate the goodness of fit between variables. We detail these below before discussing the learning and inference strategies.

3.1 Multi-task Recurrent Neural Feature Network

Our feature function network $F(X)$ is designed with a hierarchical recurrent neural network architecture to extract dialogue data in a fixed-size vector representations (Fig. 1). The architecture consists of three main elements:

- A bidirectional LSTM layer [7] for user input;
- An encoder for machine acts for the DSTC2 & 3 data, or a bidirectional LSTM layer for machine transcripts for the WOZ data. We parse the DSTC machine acts with the technique proposed by Henderson et al. [6];
- A LSTM layer to handle dialogue turns, that takes the output of the above two elements as the input and produces the vector representations of dialogues on a turn-based basis.

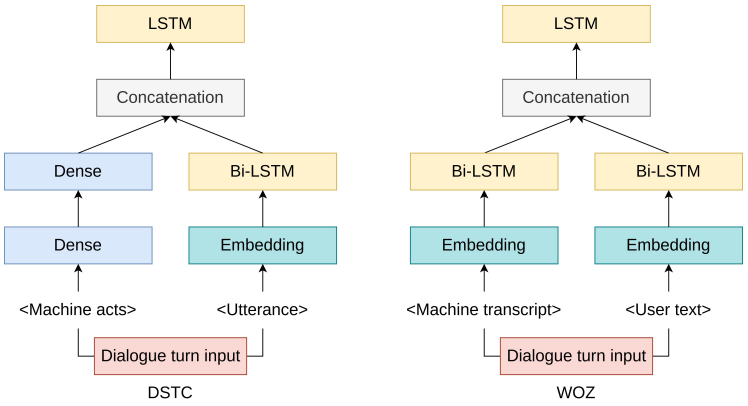

Fig. 1. Multi-task Recurrent Neural Feature Network for DSTC and WOZ datasets.

Following the practice of pretraining the feature network to achieve higher results [1], we pretrain our feature network with a multi-task learning method proposed by Trinh et al. [13]. The representations extracted with this feature network are in turn fed into the subsequent energy network.

3.2 Deep Neural Energy Network

Our energy network is developed based on the concept of the Structured Prediction Energy Network (SPEN) [1], that consists of two energy terms: local energy and global energy. It is formulated as follow:

- Energy function

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

- Local energy represents the relationships between input and label variables

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

– Global energy calculates the associations between label variables

$$E_{global}(Y) = W_2^\top f(W_1^\top Y) \tag{3}$$

Here we use the tuple, $\theta = \{W, W_1, W_2\}$, to capture the energy trainable parameters, $f(\cdot)$ is a non-linearity function for the global energy, and L is the number of classes in the target.

This has been applied previously in other structured dialogue state trackers [14, 15].

4 Energy-based Learning Strategy

The training of an energy-based model involves learning functions such that the energy function is trained to assign minimal energy value for desired variable configurations, while giving higher energy to undesired configurations. For this purpose, we implement a variant of the energy-based learning methodology based on the Deep Value Networks [2], that use a F-measurement to evaluate the compatibility between variables. However, while this variant has been successfully applied to structured dialogue state tracking [14], it is not as we noted earlier well designed for the case of outputs where certain types of constraints on those outputs much be adhered to.

4.1 Ground Truth Energy

Initially Gygli et al. [2] propose to define the ground truth energy $E^*_{F_1}$ through the use of the dice coefficient F_1 measurement as the estimation for the fitness of the predicted labels and ground truth labels. This measurement was invented to evaluate discreet classification output, and now is modified to fit the dialogue state predictions as continuous variables, that also makes it differentiable for the training process. The ground truth dice coefficient F_1 is defined as:

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

where Y is the predicted labels, Y^* is the ground truth labels, $Y \cap Y^* = \sum_i \min(y_i, y_i^*)$ and $Y \cup Y^* = \sum_i \max(y_i, y_i^*)$ are extended for continuous output variables.

We argue that the sums $\sum_i \min(y_i, y_i^*)$ and $\sum_i \max(y_i, y_i^*)$ are in fact the lower and upper boundaries of these vectors, therefore they indicate the extreme values in all cases. In a multi-label classification task the differentiable F_1 metric can be defined in a more relaxed manner [18]:

$$E^*_{F_1}(Y, Y^*) = \frac{2\sum_i y_i y_i^*}{\sum_i y_i + \sum_i y_i^*} \tag{5}$$

When comparing the two F_1 scores in Equations 4 and 5, it is not difficult to mathematically prove that

$$\begin{aligned} \sum_i \min(y_i, y_i^*) &= \sum_i y_i y_i^* \\ \sum_i \min(y_i, y_i^*) + \sum_i \max(y_i, y_i^*) &= \sum_i y_i + \sum_i y_i^* \end{aligned} \tag{6}$$

given the fact that any ground truth label y_i^* can hold only the value 0 or 1.

However, we argue that Equation 4 makes the differential process discontinuous based on the nature of the operations $\min$ and $\max$. Therefore we propose to use Equation 5 as the formula for the ground truth energy for our energy-based learning experiments. We retain the cross entropy loss function for the experiments:

$$L(E, E^*_{F_1}) = -E^*_{F_1} \log E - (1 - E^*_{F_1}) \log(1 - E) \tag{7}$$

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

4.2 Label Regularisation

In order to apply the dialogue restriction, that requires assigning only one value to a slot at any time of the conversation, onto the dialogue state prediction, we propose a label regularisation term that would penalise the predictions that activate greater or fewer values than the number of activated values in the ground truth labels. We formulate this regularisation term as follow:

$$R(Y, Y^*) = \left(\frac{\sum_i y_i - \sum_i y_i^*}{\sum_i y_i^*} \right)^2 \tag{8}$$

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

Our use of the term regularisation is based on its more general meaning and is fundamentally different from the L_2 or L_1 regularisation that instead penalise excessive parameter values.

Ultimately the objective function including the label regularisation term for training the energy network in our proposal is formulated as follow:

$$\begin{aligned} \mathcal{L} &= L(E, E^*_{F_1}) + \alpha R(Y, Y^*) \\ &= \left(- E^*_{F_1} \log E - (1 - E^*_{F_1}) \log(1 - E) \right) + \alpha \left(\frac{\sum_i y_i - \sum_i y_i^*}{\sum_i y_i^*} \right)^2 \end{aligned} \tag{9}$$

where α is the regularisation coefficient.

The whole learning process is visualised in Fig. 2.

Overall, the redefinition of the objective function, that includes redesigning the ground truth energy and introducing the label regularisation, is a novel contribution to work on structured prediction dialogue state tracking.

5 Experiment Setup

As outlined above to achieve the best results the feature network should be pretrained. Therefore we stage our experiments in two phases. Firstly, we train the feature network in a multi-task learning system where the target variables are assumed to be independent. We train the multi-tasks systems for each dataset five times with different weight initialisation and select the best one to extract dialogue features. Secondly, we train the

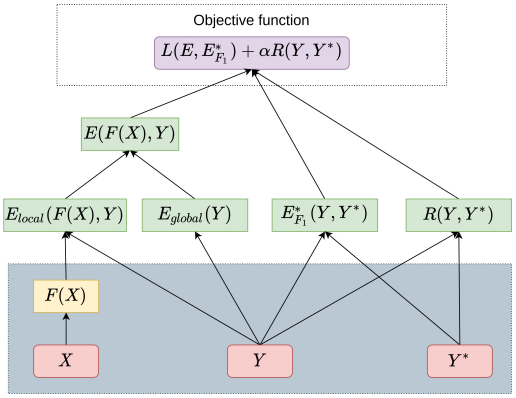

Fig. 2. The learning process of our energy-based dialogue state tracker.

energy network while freezing the feature network. The energy network is also trained five times, but the outcomes of the network are ensembled into an end prediction for evaluation. The performance of both the multi-task feature network and the energy network will be reported separately to show the improvement based on label dependencies that we aim to leverage.

As the DSTC2 and WOZ data are split into training, validation, and test subsets, we use them directly according to the purpose of these sets. Meanwhile, the DSTC3 data is provided in a single set, therefore we split it into five folds and trained our system with a cross validation technique.

6 Results & Discussion

We report the overall performance of both our multi-task feature system and energy-based system against the DSTC2 & 3 and WOZ data, and benchmark them against the state-of-the-art models in Table 1. The state-of-the-art models are selected if either they achieve the highest accuracy result or they are related to our work.

We find that the energy-based learning approach boosts the results on top of the multi-task learning methodology, up to 12% accuracy, across the datasets. That improvement strongly indicates that the impact of label dependencies in dialogue domains is significant.

Although our energy-based dialogue state trackers yield competitive results for the *Joint goals* task in all three datasets, they do not yet outperform the state-of-the-art performance seen for example in the globally-conditioned encoder (GCE) for WOZ [11], the hybrid system [17] for DSTC2, and the multi-domain system for DSTC3 [9]. However, none of the state-of-the-art systems model the label dependencies in an explicit manner when performing dialogue state tracking. Given that, we believe that their performance could be improved if structured prediction is applied, in particular with the energy-based learning methodology.

Table 1. Performances of the state-of-the-art and our dialogue state tracking systems on the DSTC 2 & 3 and WOZ data. The results for *Joint Goals* are reported with the Accuracy metric. The baseline models for DSTC2 & 3 were proposed during the competitions [3, 4], and the baseline model for WOZ is the very first work on this dataset [10].

Model	DSTC2	DSTC3	WOZ
Globally-conditioned encoder (GCE) [11]	-	-	0.885
Hybrid system [17]	0.796	-	-
Multi-domain system [9]	0.774	0.671	-
Word-based system [6]	0.768	-	-
Global-locally self-attentive tracker (GLAD) [21]	0.745	-	0.881
Unsupervised RNN-based system [5]	-	0.646	-
Our work			
Energy-based system	0.774	0.651	0.875
Multi-task feature system	0.709	0.531	0.841
Baseline [3, 4, 10]	0.719	0.575	0.844

As our contributions are centred on the redefinition of the F_1 metric as well as regularising the label constraint rules for the energy-based model, we conduct an analysis of the effectiveness of these phenomena in the dialogue state tracking results. We also conduct an analysis on errors to emphasise the role of label regularisation in structured multi-label classification.

6.1 Improvement based on F_1 Metric

To evaluate the improvement based on the redefinition of F_1 we benchmark our energy-based model during the development phase with the work done by Trinh et al. [14], which was developed based on the Deep Value Networks algorithm [2] (Table 2). Here we report our work in the development phase, that does not include the label regularisation, hence the performance is different from Table 1.

Table 2. Performances of the energy-based dialogue state trackers with different F_1 metrics on the DSTC 2 & 3 data. The results for *Joint Goals* are reported with Accuracy metric.

Energy-based DST Model	DSTC2	DSTC3
Multi-label classification F_1 (this work)	0.769	0.642
Dice coefficient F_1 (Equation 4) [14]	0.760	0.622

From this we can observe that by redefining the F_1 measurement the energy-based model achieves a slight improvement, that is approximately 1% accuracy for DSTC2 and 2% accuracy for DSTC3. We cannot compare the performance against the WOZ dataset, as it was not reported with the dice coefficient F_1 metric.

6.2 Effectiveness of Label Regularisation

The impact of label regularisation is reported in two types of analysis: the overall performance of our energy-based model before and after including the regularisation term, and the proportion of correct predictions over the total number of dialogue turns that follow the slot-value constraint rules with different thresholds.

With label regularisation we find that the overall performance of our energy-based trackers is not improved significantly (Table 3). The performance accuracy improvement across three datasets is less than 1%, that is small in comparison with the result differences seen between the multi-task feature system and the energy-based system, or when comparing with the achievement based on the F_1 measure redefinition.

Table 3. Performances of the energy-based dialogue state trackers with and without label regularisation on the DSTC 2 & 3 and WOZ data. The results for *Joint Goals* are reported with Accuracy metric.

Energy-based DST Model	DSTC2	DSTC3	WOZ
With label regularisation	0.774	0.651	0.875
Without label regularisation	0.769	0.642	0.866

We conduct another analysis to evaluate the behaviours of our energy-based systems when tracking dialogue states with our regularisation (Table 4). In this analysis we set different threshold values, and consider a value activated if the predicted belief score of this value exceeds the threshold. Among the correct predictions we count the number of those predictions satisfying the slot-value constraint rules in task-oriented dialogue domains such that each slot has only one activated value at any moment. Finally we report the proportion of recorded numbers against the total number of dialogue turns in the datasets.

Table 4. Analysis of the label regularisation on the energy-based dialogue state tracking on the DSTC 2 & 3 and WOZ data. The results are reported with the proportion (%) of the correct predictions over the total number of dialogue turns, that follow the slot-value constraint rules.

Threshold	DSTC2		DSTC3		WOZ	
	+Reg	–Reg	+Reg	–Reg	+Reg	–Reg
0.5	76.1	75.6	65.0	63.9	87.2	86.1
0.7	73.7	72.8	64.6	62.4	85.7	83.8
0.9	63.4	59.6	62.8	59.3	80.9	78.7

We find that with different threshold values the energy-based systems with label regularisation consistently outperform those without the regularisation term. This finding indicates that the impact of label regularisation in the dialogue state tracking process

is systematic. Not only can it improve the overall performance, but it also guides the system's prediction behaviour towards the requirement of specific domains.

6.3 Error Analysis

To our knowledge only one example of a comparative error analysis of dialogue state trackers was given by Smith [12]. In that analysis the author reports the error distributions of tracking models over three error types of possible deviations from the true joint goal for every turn in the dialogue. We find that these error types match our label regularisation analysis as such:

- Missing attributes (MA) is the error where the tracker fails to recognise a value for a slot despite it being present in data. In our scenario we interpret this as the label regulariser assigning the number of activated values less than the number of slots.
- Extraneous attributes (EA) is the error where the tracker classifies unnecessary values for a slot when they are not mentioned in the data. In our task it is similar to the situation when the number of activated values is bigger than the number of slots.
- False attributes (FA) is the error that occurs when the tracker assigns a false value to a slot. In this case the number of activated values satisfies the slot-value constraint rules that we apply, but the tracked dialogue state is wrong due to the false value.

The analysis results of error distributions are reported in Table 5, where we compare the behaviours of our energy-based dialogue state tracker with and without the label regularisation. We set the activated threshold 0.5 for the error analysis.

Table 5. Error distributions of the energy-based dialogue state trackers on the DSTC 2 & 3 and WOZ data. The results are reported with the proportion (%) of the number of errors with the respective type over the total number of incorrectly predicted turns.

Dataset	Label	#Turns	Error distributions		
			MA	FA	EA
DSTC2	+Reg	2235	436 (19.5%)	1283 (57.4%)	516 (23.1%)
	–Reg	2285	660 (28.9%)	919 (40.2%)	706 (30.9%)
DSTC3	+Reg	6532	1724 (26.4%)	3319 (50.8%)	1489 (22.8%)
	–Reg	6700	2144 (32.0%)	2533 (37.8%)	2023 (30.2%)
WOZ	+Reg	627	96 (15.3%)	399 (63.6%)	132 (21.1%)
	–Reg	672	226 (33.6%)	236 (35.1%)	210 (31.3%)

We observe that where the label regularisation is present, the error distributions move toward the FA error type, that indicates that the majority of errors still satisfy the slot-value constraint rules of the dialogue domains. Meanwhile the energy-based model without the label regularisation term produces more evenly distributed errors with a special case of the WOZ dataset. This finding outlines the effectiveness of the label regularisation term in the training process of our energy-based tracker.

In the comparative error analysis of Smith [12] the error distributions align with the difference in difficulty observed in tracking different slots. For example in the DSTC2 data the error distributions are relative to the order $\{food >> area >> price\ range\}$, that follows the setting of the ontology where the slot *food* has the biggest set of values, while the slot *price range* has the smallest one. However, we do not find this phenomenon in our error analysis. It can be explained that as we treat the dialogue state tracking task as a multi-label classification problem, we flatten the label set and make all the values of any slots equally important.

7 Conclusion

In this paper we demonstrated that an energy-based structured prediction methodology can be improved with additional constraint integration. We have examined this in the context of dialogue systems. By proposing to mathematically optimise the quality measurement, and regularise label classes, we demonstrated that our energy-based model's behaviours achieve a high level of satisfactory in a number of dialogue domains. We also provided a systematic analysis on the tracker's performance regarding the overall improvement, and in particular the error distributions. The error analysis is essential to understand the mechanism of dialogue state tracking process, subsequently it helps to improve future models.

We note that there are elements of the energy-based learning methodology that we can continue to develop. For the learning process of our energy-based model we see that including the label regularisation is not the only possible solution, instead, for example we can also regularise the constraint rules directly in the energy function formulation. On the other hand, the performance of our energy-based tracker can be boosted by an inference strategy where we apply the inference process multiple times to generate multiple alternative predictions and then apply a reranking process to select the best overall prediction as output.

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.

References

1. Belanger, D., McCallum, A.: Structured Prediction Energy Networks. In: Proceedings of the 33rd International Conference on Machine Learning. vol. 48 (2016)
2. Gygli, M., Norouzi, M., Angelova, A.: Deep Value Networks Learn to Evaluate and Iteratively Refine Structured Outputs. In: Proceedings of the 34th International Conference on Machine Learning (2017)
3. Henderson, M., Thomson, B., Williams, J.D.: The Second Dialog State Tracking Challenge. In: Proceedings of the SIGDIAL 2014 Conference. pp. 263–272 (2014)

4. Henderson, M., Thomson, B., Williams, J.D.: The Third Dialog State Tracking Challenge. In: Proceedings of 2014 IEEE Workshop on Spoken Language Technology. pp. 324–329 (2014)
5. Henderson, M., Thomson, B., Young, S.: Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. In: Proceedings of 2014 IEEE Workshop on Spoken Language Technology. pp. 360–365 (2014)
6. Henderson, M., Thomson, B., Young, S.: Word-Based Dialog State Tracking with Recurrent Neural Networks. In: Proceedings of the SIGDIAL 2014 Conference. pp. 292–299 (2014)
7. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
8. LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M.A., Huang, F.J.: A Tutorial on Energy-Based Learning. Predicting Structured Data (2006)
9. Mrksic, N., O'Seaghdha, D., Thomson, B., Gasic, M., Su, P.H., Vandyke, D., Wen, T.H., Young, S.: Multi-domain Dialog State Tracking using Recurrent Neural Networks. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. pp. 794–799 (2015)
10. Mrksic, N., O'Seaghdha, D., Wen, T.H., Thomson, B., Young, S.: Neural Belief Tracker: Data-Driven Dialogue State Tracking. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/P17-1163
11. Nouri, E., Hosseini-Asl, E.: Toward Scalable Neural Dialogue State Tracking Model. In: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), 2nd Conversational AI workshop (2018)
12. Smith, R.W.: Comparative Error Analysis of Dialog State Tracking. In: Proceedings of the SIGDIAL 2014 Conference. pp. 300–309 (2014)
13. Trinh, A.D., Ross, R.J., Kelleher, J.D.: A Multi-Task Approach to Incremental Dialogue State Tracking. In: Proceedings of The 22nd workshop on the Semantics and Pragmatics of Dialogue, SEMDIAL. pp. 132–145 (2018)
14. Trinh, A.D., Ross, R.J., Kelleher, J.D.: Capturing Dialogue State Variable Dependencies with an Energy-based Neural Dialogue State Tracker. In: Proceedings of the SIGDial 2019 Conference. pp. 75–84 (2019)
15. Trinh, A.D., Ross, R.J., Kelleher, J.D.: Energy-Based Modelling for Dialogue State Tracking. In: Proceedings of the 1st Workshop on NLP for Conversational AI. pp. 77–86 (2019)
16. Trinh, A.D., Ross, R.J., Kelleher, J.D.: Investigating Variable Dependencies in Dialogue States. In: Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue. pp. 195–197 (2019)
17. Vodolan, M., Kadlec, R., Kleindienst, J.: Hybrid Dialog State Tracker with ASR Features. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL. vol. 2, pp. 205–210 (2017)
18. Wang, B., Li, C., Pavlu, V., Aslam, J.: Regularizing Model Complexity and Label Structure for Multi-Label Text Classification. In: Proceedings of KDD'17. Halifax, Nova Scotia - Canada (2017). https://doi.org/10.1145/nnnnnnn.nnnnnnn
19. Wen, T.H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-barahona, L.M., Su, P.h., Ultes, S., Young, S.: 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, EACL. pp. 438–449 (2017)
20. Williams, J.D., Raux, A., Henderson, M.: The Dialog State Tracking Challenge Series: A Review. Dialogue & Discourse **7**(3), 4–33 (2016). https://doi.org/10.5087/dad.2016.301
21. Zhong, V., Xiong, C., Socher, R.: Global-Locally Self-Attentive Dialogue State Tracker. In: Proceedings of the 56th Annual Meeting ofthe Association for Computational Linguistics. pp. 1458–1467 (2018)