2023

Structured Dialogue State Management for Task-Oriented Dialogue Systems

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

School of Computer Science

Technological University Dublin

Structured Dialogue State Management for Task-Oriented Dialogue Systems

Submitted by:

Anh Duong Trinh

Supervisors:

Prof. John D. Kelleher

Dr. Robert J. Ross

Thesis Submitted for the degree of

Doctor of Philosophy

January 2023

Abstract

Human-machine conversational agents have developed at a rapid pace in recent years, bolstered through the application of advanced technologies such as deep learning. Today, dialogue systems are useful in assisting users in various activities, especially task-oriented dialogue systems in specific dialogue domains. However, they continue to be limited in many ways. Arguably the biggest challenge lies in the complexity of natural language and interpersonal communication, and the lack of human context and knowledge available to these systems. This leads to the question of whether dialogue systems, and in particular task-oriented dialogue systems, can be enhanced to leverage various language properties. This work focuses on the semantic structural properties of language in task-oriented dialogue systems. These structural properties are manifest by variable dependencies in dialogue domains; and the study of and accounting for these variables and their interdependencies is the main objective of this research.

Contemporary task-oriented dialogue systems are typically developed with a multiple component architecture, where each component is responsible for a specific process in the conversational interaction. It is commonly accepted that the ability to understand user input in a conversational context, a responsibility generally assigned to the dialogue state tracking component, contributes a huge part to the overall performance of dialogue systems. The output of the dialogue state tracking component, so-called dialogue states, are a representation of the aspects of a dialogue relevant to the completion of a task up to that point, and should also

capture the task structural properties of natural language. Here, in a dialogue context dialogue state variables are expressed through dialogue slots and slot values, hence the dialogue state variable dependencies are expressed as the dependencies between dialogue slots and their values. Incorporating slot dependencies in the dialogue state tracking process is herein hypothesised to enhance the accuracy of postulated dialogue states, and subsequently potentially improve the performance of task-oriented dialogue systems.

Given this overall goal and approach to the improvement of dialogue systems, the work in this dissertation can be broken down into two related contributions: (i) a study of structural properties in dialogue states; and (ii) the investigation of novel modelling approaches to capture slot dependencies in dialogue domains.

The analysis of language's structural properties was conducted with a corpus-based study to investigate whether variable dependencies, i.e., slot dependencies when using dialogue system terminology, exist in dialogue domains, and if yes, to what extent do these dependencies affect the dialogue state tracking process. A number of public dialogue corpora were chosen for analysis with a collection of statistical methods being applied to their analysis.

Deep learning architectures have been shown in various works to be an effective method to model conversations and different types of machine learning challenges. In this research, in order to account for slot dependencies, a number of deep learning-based models were experimented with for the dialogue state tracking task. In particular, a multi-task learning system was developed to study the

leveraging of common features and shared knowledge in the training of dialogue state tracking subtasks such as tracking different slots, hence investigating the associations between these slots. Beyond that, a structured prediction method, based on energy-based learning, was also applied to account for explicit dialogue slot dependencies.

The study results show promising directions for solving the dialogue state tracking challenge for task-oriented dialogue systems. By accounting for slot dependencies in dialogue domains, dialogue states were produced more accurately when benchmarked against comparative modelling methods that do not take advantage of the same principle. Furthermore, the structured prediction method is applicable to various state-of-the-art modelling approaches for further study.

In the long term, the study of dialogue state slot dependencies can potentially be expanded to a wider range of conversational aspects such as personality, preferences, and modalities, as well as user intents.

Declaration

I certify that this thesis which I now submit for examination for the award of PhD, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work.

This thesis was prepared according to the regulations for graduate study by research of the Technological University Dublin and has not been submitted in whole or in part for another award in any other third level institution. The work reported on in this thesis conforms to the principles and requirements of the TU Dublin's guidelines for ethics in research.

TU Dublin has permission to keep, lend or copy this thesis in whole or in part, on condition that any such use of the material of the thesis be duly acknowledged.

Signature: *Anh Duong Trinh*

Date: 12.01.2023

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.

I would like to take this opportunity to thank my supervisors, Professor John D. Kelleher and Dr. Robert J. Ross, for their guidance and patience throughout my PhD programme. Their efforts contribute to a great part of my success today.

I would like to also thank a number of my colleagues who have helped me at various stages of my study: Giancarlo Salton, Eoin Rogers, Annika Lindh, and Elizabeth Hunter. Their assistance was invaluable to me.

Last, but not least, I would like to thank my partner David Ferguson, my good friend Thi Nguyet Que Nguyen, and my family for their emotional support.

Table of Contents

List of Abbreviations **xi**

List of Figures **xiii**

List of Tables **xv**

1 Introduction **1**

1.1 Research Objectives 9

1.2 Investigation Methods and Scope 13

1.3 Research Contributions 16

1.3.1 Publications 19

1.4 Structure of Dissertation 21

2 Literature Review **24**

2.1 Dialogue State Representations 25

2.1.1 Dialogue Acts 26

2.1.2 Dialogue Attributes and Values 30

2.1.3 Dialogue States 31

2.1.4 Dialogue Frames 33
2.2 Dialogue State Tracking Methods 35
2.2.1 Rule-based Systems 37
2.2.2 Generative Systems 38
2.2.3 Discriminative Systems 41
2.2.4 Hybrid Systems 44
2.2.5 Domain Specific Dialogue Corpora 50
2.2.6 State-Of-The-Art Models 55
2.3 Structured Prediction Applications in Dialogue Research 60
2.4 Summary 62

3 Inter-Slot Dependencies in Dialogue States 65

3.1 Dialogue Datasets 66
3.2 Analysis Methods 69
3.2.1 Pearson's Chi-Square Test 69
3.2.2 Measuring Slot Dependencies 71
3.3 Dialogue Slot Dependencies Analysis 73
3.3.1 Single Dialogue Domain 73
3.3.2 Multiple Dialogue Domain 75
3.4 Discussion 77
3.5 Summary 78

4 Harnessing Domain Structure with Multi-Task Learning 80

4.1 Overview of Multi-Task Learning 81
4.2 Multi-Task Dialogue State Tracker 84
4.3 Results & Analysis 88
4.3.1 Analysis of Slot-Based Performance 91
4.4 Summary 92

5 Studying Slot Dependencies with Energy-Based Learning 94

5.1 Overview of Energy-Based Learning 96
5.2 Energy-Based Dialogue State Tracking Model 99
5.2.1 Hierarchical Recurrent Neural Feature Network 99
5.2.2 Deep Neural Energy Network 101
5.3 Energy-Based Modelling Strategies 104
5.3.1 Learning Process 105
5.3.2 Inference Process 109
5.3.3 End-to-End Learning 111
5.4 Results & Analysis 116
5.4.1 Energy-based Modelling Performance 116
5.4.2 Analysis of Slot-Based Performance 120
5.4.3 Analysis of Slot Dependencies in Predicted States 122
5.5 Summary 123

6 Slot Value Regularisation for Energy-Based State Tracking 125

6.1 Overview of Dialogue State Constraints 126

6.2 Modified EBL Architecture 129
6.2.1 Multi-Task Recurrent Neural Feature Network 129
6.3 A Modified Learning Process 131
6.3.1 Ground Truth Energy 132
6.3.2 Slot Value Regularisation 133
6.4 Results & Analysis 135
6.4.1 Modified Energy-Based Modelling Performance 136
6.4.2 Effectiveness of Redefined F_1 Metric 138
6.4.3 Effectiveness of Slot Value Regularisation 139
6.4.4 Analysis of Slot Value Constraint Rules 140
6.4.5 Analysis of Error Distributions 141
6.5 Summary 144

7 Generalisability to Multiple Dialogue Domains 146

7.1 Overview of Multi-Domain Dialogue 147
7.2 Large-Scale EBL Dialogue State Tracking 149
7.2.1 Large-Scale Recurrent Neural Feature Network 149
7.2.2 Large-Scale Deep Neural Energy Network 152
7.3 Energy-Based Learning Processes 152
7.4 Results & Analysis 154
7.4.1 Analysis of Slot Dependencies 157
7.4.2 Analysis of Slot Value Constraint Rules 159
7.4.3 Analysis of Error Distributions 160

7.5 Summary 162

8 Conclusion **164**

8.1 Summary of Contributions 165

8.2 Directions for Future Work 168

References **174**

Appendices **214**

A Experiment Training Details **215**

B List of Publications **217**

C List of Employability and Discipline Specific Skills Training **220**

List of Abbreviations

BERT Bidirectional Encoder Representations from Transformers

CNN Convolutional Neural Network

DeepSPIN Deep Structured Prediction in NLP

DRT Discourse Representation Theory

DST Dialogue State Tracking

DSTC Dialogue State Tracking Challenge

EBL Energy-Based Learning

ERC European Research Council

GPT Generative Pre-trained Transformer

HIS Hidden Information State

LSTM Long Short-Term Memory

MultiWOZ Multi-Domain Wizard-Of-OZ

NLP Natural Language Processing

POT Probabilistic Ontology Trees

RNN Recurrent Neural Network

SGD Schema-Guided Dialogue

WOZ Wizard-Of-Oz

List of Figures

1.1 The common architecture of Spoken Dialogue Systems. 4

1.2 An example of dialogue states in task-oriented dialogue systems. 7

2.1 ISO 24617-2 dialogue act properties (Bunt et al., 2017). 27

3.1 Cramer's V assessment of slot type dependencies in MultiWOZ 2.1 data 76

4.1 Machine act autoencoder. 85

4.2 Sequential dialogue input. 85

4.3 The multi-task baseline model (a) and the proposed multi-task dialogue state tracker (b). 87

5.1 The hierarchical recurrent neural architecture to transform dialogue input into fixed-size vector representations. For the sake of simplicity of presentation, the unidirectional LSTM and concatenation layers are jointly presented as LSTM cells rolling up dialogue turns. m denotes encoded machine acts, u denotes vector representations for user utterances, and h denotes hidden states representing dialogue turn information. 100

5.2 The deep neural structures of local and global energy functions. 103

5.3 The learning process of the energy-based dialogue state tracker. 107

5.4 The inference process of the proposed energy-based dialogue state tracker. 110

5.5 The working mechanism of the end-to-end energy-based dialogue state tracker. 112

6.1 Multi-task Recurrent Neural Feature Network for DSTC and WOZ datasets. 130

6.2 The learning process including the value regularisation of the proposed energy-based dialogue state tracker. 135

7.1 Large-Scale Recurrent Neural Feature Network for MultiWOZ data. . . 150

List of Tables

1.1 An overview of investigation methods for dialogue state tracking. The list of abbreviations used in this table: D – discriminative, G – generative, H – hybrid, DL – deep learning, DE – data efficient 15

2.1 The 10 dimensions of dialogue acts in the ISO 24617-2 scheme based on the work by Bunt et al. (2020, 2017). 29

2.2 An example of DSTC2 dialogue state slot and value combinations. . . . 31

2.3 An example of DSTC2 dialogue states following user utterances. 33

2.4 An overview of different types of dialogue state tracking methods. . . . 36

2.5 Summary of the state-of-the-art dialogue state tracker entries for the DSTC series. 54

2.6 Summary of the state-of-the-art dialogue state trackers for the DSTC series. * denotes the approach submitted during the competition time. 56

2.7 Summary of the state-of-the-art dialogue state trackers for the WOZ and MultiWOZ datasets. 57

2.8 Summary of the state-of-the-art dialogue state trackers for the SGD dataset retrieved from the competition report by Rastogi et al. (2020a). 58

2.9 Summary of the state-of-the-art dialogue frame trackers. 59

3.1 The analysis of informable slot appearance (%) in DSTC 2 & 3, calculated over the number of dialogues and turns in the whole dataset. . . . 68

3.2 Overview information of the chosen dialogue corpora for this research. . 69

3.3 Interpretation of Cramer's V coefficient (Field, 2017). 73

3.4 Statistical assessment of slot dependencies in the DSTC2 data. 74

3.5 Statistical assessment of slot dependencies in the WOZ and DSTC3 data in the Cramer's V coefficient. 74

4.1 Performance of the proposed multi-task models and related state-of-the-art systems on DSTC2 testset evaluated with the accuracy metric. 89

4.2 Detailed performance of the proposed multi-task models on DSTC2 informable slots. 91

5.1 Comparison of the end-to-end and separate process algorithms for energy-based learning. 114

5.2 Performances of state-of-the-art and the energy-based dialogue state tracking systems on DSTC2 & 3 data. 117

5.3 Performances of the energy-based dialogue state tracking systems per slot and for *Joint goals* of those present in the task. 120

5.4 Proportional reduction in errors (%) of the energy-based system for each slot and the *Joint goals*. 121

5.5 Analysis of slot dependencies on the DSTC2 & 3 test data. The results are reported in the Cramer's V coefficient. 123

6.1 An example of dialogue states with and without slot value constraint rules. 127

6.2 Performances of the state-of-the-art and the proposed dialogue state tracking systems on the DSTC 2 & 3 and WOZ data. 137

6.3 Performances of the energy-based dialogue state trackers with different F_1 metrics on the DSTC2 & 3 data. 138

6.4 Performances of the energy-based dialogue state trackers with and without value regularisation on the DSTC2 & 3 and WOZ data. 139

6.5 Analysis of the value regularisation on the energy-based dialogue state tracking on the DSTC 2 & 3 and WOZ data. 140

6.6 An example of the three error types of dialogue state tracking. *MA* denotes missing attributes, *EA* denotes extraneous attributes, and *FA* denotes false attributes. 142

6.7 Error distributions of the energy-based dialogue state tracker on the DSTC2 & 3 and WOZ data. 143

7.1 Performances of state-of-the-art and the energy-based dialogue state tracking systems on MultiWOZ 2.0 & 2.1 data. 154

7.2 Analysis of slot dependencies in the MultiWOZ 2.1 testset, and the predicted dialogue states of the energy-based model and the multi-task feature network. 158

7.3 Analysis of the impact of value regularisation on the energy-based dialogue state tracking on the MultiWOZ 2.0 & 2.1 data. 160

7.4 Error distributions of the energy-based dialogue state tracker on the MultiWOZ 2.0 & 2.1 data. 161

A.1 Hyper parameters used in experiments constructing the energy-based dialogue state tracker. 216

Chapter 1

Introduction

A dialogue system is a computational system which communicates with users through conversational activities. Dialogue systems can interact with users in many ways. Commonly used information retrieval centred dialogue applications, i.e. conversational help assistants, are developed to assist users in specific tasks such as customer service (Hewitt & Beaver, 2020; G. Zhao et al., 2019), educational activities (Okonkwo & Ade-Ibijola, 2021; J. A. Kumar, 2021), and health care (Cristofori et al., 2021; K.-H. Liang et al., 2021). Therefore, they are classified as task-oriented dialogue systems. In contrast, general purpose dialogue systems may instead aim to maximise the user engagement with a so-called chitchat ability (Hardy et al., 2021; Ishii et al., 2021; Sun et al., 2021; Gabriel et al., 2020).

Both dialogue system types have their own advantages and disadvantages. For example, task-oriented dialogue systems are very specific to the domain they are built for, and as such they can achieve high performance in solving the domain

specific task, but perform badly for out-of-domain questions. On the other hand, general purpose dialogue systems can entertain users for a long period on a wide range of topics, while not satisfying users on specific queries. Today, task-oriented dialogue systems are deployed for various activities thanks to their high performance, for example smart campus building hosts (Sieińska et al., 2020) and ticket booking assistants (Byrne et al., 2021; Al-Ajmi & Al-Twairesh, 2021).

Dialogue system development involves a wide range of research areas. From the cognitive and linguistic perspectives, human-machine conversations inherit all the properties of natural language. Hence, dialogue systems are expected to obtain the ability to handle conversations with users in a natural manner, i.e. as natural as human-human conversations (Landragin, 2013). Achieving this level of conversational ability is a big challenge even for advanced technology due to the complex structure of language. From the technical perspective, building quality dialogue systems, in particular task-oriented dialogue systems, still faces many challenges such as learning in a low-resource environment and adapting to users' behaviours (Z. Zhang et al., 2020; Ward & Devault, 2016; Ward & DeVault, 2015).

In general, there are two ways to develop task-oriented dialogue systems: modular (Truong et al., 2017; F. Chen, 2020; K. Liang et al., 2020) and end-to-end (S. Lee et al., 2019; B. Liu et al., 2018; J. D. Williams et al., 2017). In the modular case, a task-oriented dialogue system consists of various components with different functionalities (Harrison et al., 2020; Bowden et al., 2018). Meanwhile, an end-to-end task-oriented dialogue system is a single model that receives natural

language input and responds with natural language output (B. Liu & Lane, 2018; J. D. Williams & Liden, 2017). The downside of modular systems is that their development requires large-scale dialogue data for each of the components in the architecture. Meanwhile, although modular dialogue systems are more complicated to construct, the performance of task-oriented dialogue systems with modular architectures is more interpretable and stable than in the case of their end-to-end counterparts. Hence, modular dialogue systems are more commonly developed and applied.

The modular architecture can be applied to many different dialogue system types, including spoken dialogue systems. A prototypical spoken dialogue system architecture is presented in Figure 1.1. Here the working mechanism of the spoken dialogue system starts when an automatic speech recogniser component receives users' utterances, then analyses and transforms them into speech hypotheses. After that, a spoken language understanding component uses these hypotheses to produces semantic representations and passes these onto a dialogue manager. The dialogue manager is the core component of any modular dialogue system, and consists of two subcomponents: a dialogue state tracker and a dialogue policy planner. Here, the dialogue state tracker maintains dialogue representations from received input, while the dialogue planning unit generates an appropriate response. This generated response needs transforming back to the natural language form – which is taken care of by a natural language generation component. At the end of the path, a text-to-speech unit generates verbal speech and transmits it to users. This

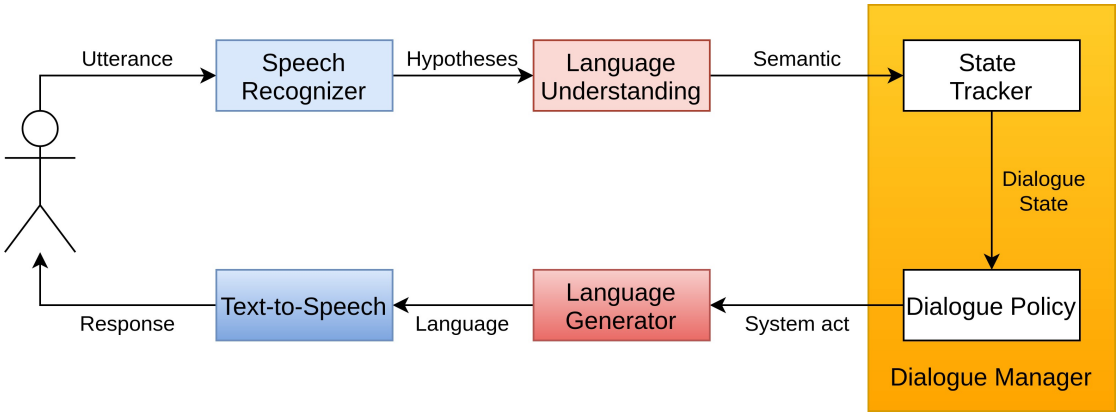

Figure 1.1: The common architecture of Spoken Dialogue Systems.

process repeats until users achieve their goals (Ross, 2009).

Dialogue systems are also developed with different modalities such as, but not limited to, text, speech, and haptic (or touch) channels. The system in Figure 1.1 is designed to handle human speech with two specific components, a speech recogniser and a text-to-speech unit. In order to accommodate other modalities, the architecture can be modified to include more components, or indeed omit a few of them. For example, a text-based dialogue system does not need audio processing components such as the speech recogniser and the text-to-speech unit (Nakano & Komatani, 2020; Kelleher et al., 2005), while a video chatbot must include a camera to capture users' gestures and facial expressions (DeVault et al., 2014).

In the most popular dialogue system architectures to date, the process of handling dialogue mainly lies within the dialogue manager. The dialogue manager is the core part of any dialogue system. This component is commonly split into two subcomponents with different functionalities: a dialogue state tracker, and a dialogue planning unit. The dialogue state tracker is responsible for maintaining

the dialogue context, that includes dialogue history, user intents, and other knowledge. The dialogue information representations maintained by this tracker are often known as dialogue states, the idea of which was inspired by the information state update approach to dialogue management (Traum, 2000; Traum & Larsson, 2001, 2003). The dialogue planning unit, on the other hand, generates an appropriate response policy to these dialogue states. Generally these subcomponents can be developed together in an end-to-end manner (J. D. Williams et al., 2017; X. Li et al., 2017; Serban et al., 2016) or independently (Budzianowski et al., 2017; P.-H. Su et al., 2017; T. Zhao & Eskenazi, 2016).

Dialogue State Tracking (DST) is an essential but very challenging task for the development of dialogue systems, in particular task-oriented dialogue systems. The quality of a dialogue state tracker has a huge impact on the overall system performance. More accurate dialogue states help improve the appropriateness of the response, while bad dialogue state predictions can lead the conversation moving in a wrong direction, and hence making users dissatisfied (J. D. Williams, 2012). While DST can be used in implementing many different types of dialogue manager, in practice we are mostly concerned with the very common slot-filling dialogue management paradigm.

The development of the dialogue state tracker component itself faces many challenges. For instance, from the technical point of view, the dialogue state tracker suffers from the imperfection of current technology. In the spoken dialogue system architecture in Figure 1.1 the speech recognising and the language understanding

components can produce errors in the process, thus creating more noise for the dialogue state tracker to handle since it takes the output of these components as the input for dialogue states. On the other hand, from the linguistic point of view, the complexity of human conversational activities is very challenging to model computationally.

The content of the dialogue state tracker, the so-called dialogue states, are the dialogue information representations maintained and updated during conversations that reflect user intents such as what information users provide to the system, and what questions they ask. In general, dialogue states are defined in various formats depending on the purpose of the dialogue systems. For example, a popular form of dialogue state in task-oriented dialogue systems is a combination of slot and value pairs that are predefined in an ontology; these are for example used in the Dialogue State Tracking Challenge series, and illustrated in Figure 1.2 (Henderson, Thomson, & Williams, 2013, 2014a,b). Dialogue states can also be represented in a semantic format together with dialogue acts (Young et al., 2010) or using more complex structures such as multiple frames (Asri et al., 2017).

Dialogue state tracking should address various challenges in processing human-machine conversations, for example the challenges in cognitive and linguistic aspects (Landragin, 2013). Here, the challenges are considered with respect to the nature of human-machine conversational interactions, and not the issue of system development. In general, a dialogue system is characterised as an artificial cognitive system that should possess common human knowledge and behave based

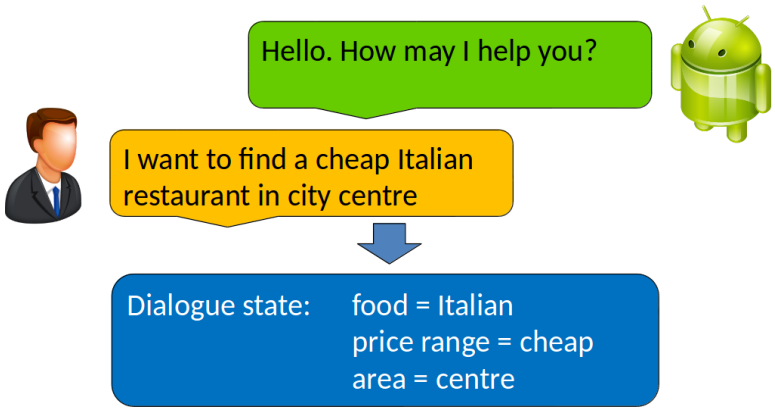

Figure 1.2: An example of dialogue states in task-oriented dialogue systems.

on both this knowledge and built-in data. The modelling of an artificial cognitive system often follows the path of computational cognitive science, that tends to resemble humans in processing the information flow. Landragin (2013) detailed the human cognitive system as a complex set of different mechanisms including, but not limited to, perceptive mechanisms, attentive mechanisms, memory, the ability to represent information and reason, and the ability to learn. Although the development of a useful task-based dialogue system may not require the system to implement all of these cognitive mechanisms at the automatic and individualised level found in humans, it is useful for dialogue system research to be capable of these mechanisms in order that the system's interaction with humans is as natural as possible. For instance, the dialogue information flow is represented by dialogue states that are required to include all context and dialogue history up to the current point of conversations.

The cognitive challenges are well linked to the linguistic aspects since dialogue systems must handle language in general. The modern trend of multimodalities

in dialogue adds more research disciplines on top of the primary concern of natural language processing. The language understanding unit, when included in the dialogue system architecture, can perform language analysis to a certain extent, for example extracting linguistic meaning from different surface forms of language. However, it neglects the conversational analysis such as discourse and pragmatic structure. These important pieces of conversational information are in practice implicitly captured by the dialogue state tracking component. Here, the difference between the human cognitive system and the artificial cognitive system in relation to natural language processing can be highlighted with the structural properties of language. Humans automatically process natural language as a whole piece of information, and in so doing take account of relationships and dependencies between information pieces. Meanwhile dialogue systems split the conversational content into pieces for analysis purposes, and in doing so often process information pieces independently.

In the dialogue management of task-oriented dialogue systems, it is important to highlight that the structure of meaning in dialogue states is an ontological concept, rather than a linguistic aspect. Here, dialogue states are produced based on the idea of different dialogue structures such as tasks, slots, frames, and others. Some of these structures can be considered dialogue variables. These are in fact conceptual structure related to dialogue systems, not linguistic structures directly based on natural language. Therefore, the study of the structural properties of language in dialogue state management can be understood as the study of conceptual

relationships and dependencies between variables in the dialogue state structures.

Given the difference between the human and artificial cognitive systems, dialogue state tracking models often do not take into account the structural nature of language in the process of producing dialogue states. Hence, the scope of this research is at the highest level to study natural language's structural properties in the management of dialogue states.

1.1 Research Objectives

Human-machine conversations make use of the structural properties of natural language, therefore dialogue systems should account for these. In general cases, structural properties are characterised by the relations within a given structure (Korbmacher & Schiemer, 2018). This characteristic is also seen in the case of natural language, whereas there exist many forms of structures such as syntactic and semantic structures (Garvin, 1976). Consequently, human-machine conversations have their own forms of structures presented in dialogue states. Thus, the study of the structural properties, i.e., the relations within a given structure, in dialogue states and the leveraging of these structural properties is the objective of this research.

From the perspective of dialogue management, the structural properties of dialogue states are characterised by the relationships among dialogue slots and their outputs, i.e. slot values. In this manner, the structure in the dialogue state presents itself as a conceptual structure rather than a purely linguistic structure.

Nevertheless, dialogue systems, in particular task-oriented dialogue systems, often treat different slots within the dialogue state structures as separate pieces to process. A highlighted example is the work by Henderson, Thomson, & Young (2014b), where a separate neural network was developed for each dialogue slot. Although this processing manner is straightforward and very effective in specific domains, it is argued in this dissertation that this approach neglects the structural properties of language described above, and hence has clear room to be improved upon.

Specifically, as dialogue state tracking is the main dialogue process and dialogue states are considered the full task-relevant information representations of dialogues, the structural properties must be reflected in those states to some extent. There is a lack of systematic and detailed studies that can explore and exploit the fact that the dialogue states are structured and not simply collections of independent slots. Hence, this research aims to investigate the structural properties of dialogue states, that are represented in the format of inter-slot dependencies within the dialogue domains. It is hypothesised that by accounting for the inter-slot dependencies within dialogue states that are structured, the state of the art in dialogue state tracking can be improved.

Based on the argument above, the main research question to be addressed in this dissertation is:

How can the performance of dialogue state tracking models be improved by incorporating knowledge of the structural properties of the

dialogue states?

In order to address this research question, the following issues are addressed as supporting questions:

- **RQ-1**: *To what extent are structural properties prevalent in task-oriented dialogue?*

  As mentioned above, dialogue states are considered the full task-relevant representations of human-machine conversations. It is thus essential to conduct a systemic corpus-based study to investigate the structural properties of the language, i.e., the slot dependencies, presented in dialogue states. The result should provide a strong background for further study of the incorporation of the structural properties into an automatic prediction process. The research addressing this sub-question is primarily presented in Chapter 3.

- **RQ-2**: *Since multiple tasks are often accomplished in parallel in dialogue, in what way can a machine learning approach to dialogue state tracking be used to take advantage of this fact?*

  Following the study of the structural properties of dialogue data, an essential next step is to investigate whether the relationships between dialogue slots, presented as dialogue state tracking subtasks, are useful for producing dialogue states. Dialogue state tracking is often split into a set of multiple tasks that have close relationships with each other. The information achieved in one task can be hypothetically helpful in solving another task. Among machine learning methods, multi-task learning is a common technique for

various language processing tasks, that accounts for the relations between tasks. Combining these facts, the multi-task learning approach might be useful for a study of the relations among dialogue state subtasks. The research addressing this sub-question is primarily presented in Chapter 4.

- **RQ-3**: *If there are correlations between variables typical of a structured prediction task, i.e., slot dependencies, how can a structured prediction approach to machine learning be applied to leverage these?*

  The slot dependencies are hypothesised to have an impact on the prediction process, therefore leveraging these relationships with a suitable machine learning-based approach should be a promising research direction. Structured prediction methods are good at capturing the structured dependencies among variables, and have a handful of applications in various natural language processing fields. Applying a structured prediction approach to the dialogue state tracking task brings the novelty to the project. The research addressing this sub-question is primarily presented in Chapter 5.

- **RQ-4**: *How can it be determined whether a structured prediction approach makes a difference to the process of dialogue state tracking?*

  A data-centred approach such as a structured prediction method should improve the overall performance of the system. However, this improvement can be achieved by different machine learning approaches, which are not specific to the structured task. Therefore, a systemic evaluation of how the

structured prediction method leverages the structural properties of the data in the prediction process is a critical aspect of this research. The research addressing this sub-question is primarily presented in Chapter 6.

- **RQ-5**: *To what extent can a structured learning process be generalised across domains?*

  Dialogue domains have evolved from single domain to multiple domains in recent years to fit the nature of conversations and to improve user satisfaction. In multi-domain dialogues, slots are dependent not only within a particular domain, but also across domains. This levels up the challenge for dialogue state management and leads to the question of how to leverage the structural properties of multi-domain dialogue data with a structured prediction approach. The research addressing this sub-question is primarily presented in Chapter 7.

1.2 Investigation Methods and Scope

Dialogue representations in task-oriented dialogue systems are generally represented with the logical and semantic forms that include dialogue acts followed by a set of act items (Schatzmann, 2008). However, task-oriented dialogue domains often include a predefined ontology of different attributes and values that are incorporated into the definition of dialogue understanding (Ross & Bateman, 2009). In this case, dialogue representations, or so-called dialogue states, are defined by

a set of attributes, or so-called slots, and their values are then tracked for every turn. The use of dialogue acts is not actually vital for task-oriented dialogue systems. Instead dialogue state tracking shifts to the task of defining correct values for dialogue slots.

Despite so much effort in research for the dialogue state tracking task, modelling solutions are far from perfected. The methodology trend has evolved into big and complex systems that model discriminative and hybrid dialogue state tracking techniques. The evolution has also expanded to deep learning-based modelling and data-efficient approaches (Shalyminov, 2020). These methods currently yield state-of-the-art performances in various domains and datasets (Table 1.1). Following this trend, modelling deep learning architectures for the dialogue state tracking task is an appropriate approach.

However, the main research objective of this dissertation is to explore and exploit the structural properties of dialogue states, wherein there is a lack of a systematic and detailed study. For this reason, structured prediction approaches (LeCun et al., 2006) are seen as promising for the task. Combining this with the research trend outlined above, the application of deep learning-based structured prediction methods to dialogue state tracking is a promising direction. Another aim of this dissertation is to design and develop deep learning-based architectures that are effective and not difficult to replicate, so that the research community can benefit further from the work.

It is worth clarifying the terminology used throughout my dissertation and the

Table 1.1: An overview of investigation methods for dialogue state tracking. The list of abbreviations used in this table: D – discriminative, G – generative, H – hybrid, DL – deep learning, DE – data efficient

Dataset	Investigation method	Properties		
		Type	DL	DE
DSTC1	Structured discriminative model (S. Lee, 2013)	D	✓	✓
DSTC2	Sequence-to-sequence model (Feng et al., 2021)	G	✓	✓
DSTC3	Multi-domain neural belief model (Mrksic et al., 2015)	D	✓	✓
DSTC4	Hybrid tracker with hand-crafted rules (Dernoncourt et al., 2016)	H	✓	✓
DSTC5	Multi-channel CNN model (H. Shi et al., 2016b)	D	✓	✓
WOZ	Amendable generation model (Tian et al., 2021)	H	✓	✓
MultiWOZ 2.0	Knowledge-aware graph-enhanced GPT2 model (W. Lin et al., 2021)	H	✓	✓
MultiWOZ 2.1	TripPy + SaCLog model (Dai et al., 2021)	H	✓	✓
SGD	Machine reading comprehension & WD classification model (Ma et al., 2020)	H	✓	✓
Frames	Frame tracking model (Schulz et al., 2017)	D	✓	✓

difference between the modelling approaches used in experiments. The focus of my research is the slot-based dialogue system type among task-oriented dialogue systems. From here, task-oriented dialogue systems and slot-based systems have the same interpretation in this dissertation. The research objective is to investigate slot dependencies in dialogue states, wherein inter-slot dependencies are defined by the correlation between different slots being assigned particular values. In some places throughout the dissertation, inter-slot dependencies can be shortened to slot dependencies, and these terms have the same meaning.

Since dialogue state tracking for task-oriented dialogue systems is to assign

correct values for dialogue slots for each dialogue turn, the common approach is to treat each of the slots as a subtask and to develop multinomial classification models for said slots. The dialogue states are then assembled by joining the output of these models. In my research, on the one hand the multi-task learning approach is the assembling of multiple multinomial classification models and training them in a multi-task learning manner. On the other hand, in order to explicitly investigate the structural properties of dialogue states the values of dialogue slots are joined and predicted together, hence multiple values are assigned to multiple slots at the same time. Therefore, the structured prediction approach is similar to multi-label classification modelling when performing predictions. These approaches will be detailed in further chapters of this dissertation.

In the scope of this dissertation, it is worth noting that the investigation of structured prediction methods to dialogue state tracking mainly targets the study of the structural properties of dialogue states, but is not intended at this point to compete with state-of-the-art dialogue state tracking approaches. Nevertheless, I hope that this study may help to improve state-of-the-art results in the long term through the incorporation of these structural methods with other state-of-the-art developments.

1.3 Research Contributions

This dissertation delivers a number of contributions to the human-machine dialogue research field. Ultimately it presents a detailed and systematic study of the

structural properties of dialogue states, that are represented by the dependencies between dialogue slots and values. This research also demonstrates the benefits of incorporating these properties into the dialogue state tracking process with deep learning-based modelling approaches.

A corpus-based study was conducted to determine whether slot dependencies exist in dialogue states, and if yes, to what extent these dependencies vary among slots. The dialogue slot dependencies were examined with statistical tests and their related metrics. This study provides a solid ground for the fact that conversational language preserves linguistic properties, in particular the structural nature, of human language. This finding is very beneficial for further study in the community. This study is published in Trinh et al. (2019a,c)

A number of unique and novel deep learning-based models were also developed for the dialogue state tracking task where the slot dependencies in dialogue states needed accounting for. In detail:

- A multi-task learning model with a novel architecture was developed to track dialogue states while making use of the shared developed functions between dialogue state subtasks. This approach demonstrates that a multi-task learning model that consists of both shared parameters between all subtasks and separate parameters for individual subtasks benefits from both shared and separate training channels. The model with this complex architecture performs better than individual independent models of subtasks and multi-task learning models that do not require subtasks to share trained parameters.

This multi-task learning model is detailed in Trinh et al. (2018).

- Structured prediction is a novel method applied to the dialogue state tracking task. A structured prediction method, namely energy-based learning, was applied to explicitly account for dialogue slot dependencies. This approach demonstrates that taking into account the structural properties of dialogue states improve the tracking process significantly. This application is the first of its kind in the dialogue state tracking research field. This energy-based dialogue state tracker was presented in Trinh et al. (2019a,b).

It is important to ensure that the predicted dialogue states follow the rules of dialogue domains. Hence, the evaluation of these phenomena plays an essential part in the application of energy-based learning in dialogue processing. In this research a systematic evaluation method was also developed to determine whether an energy-based state tracking method produces quality dialogue states. The evaluation measures are also applicable to other works in the community. This evaluation study was detailed in Trinh et al. (2020b).

Another contribution of this dissertation is to prove the generalisability of the structured prediction methodology, in particular energy-based learning, to the problem of dialogue state tracking. It is well known that changing from single dialogue domain to multiple dialogue domains complicates the dialogue state tracking task and can make many methods infeasible (Balaraman et al., 2021; H. Chen et al., 2017). However, this research demonstrates the flexibility of the energy-based learning methods when working with different dialogue domains regardless of the

number of dialogue slots and their values. This work was published in Trinh et al. (2020a).

1.3.1 Publications

The research contributions are represented below with a list of my publications from throughout my PhD programme, that include a number of long papers, extended abstracts, and position papers presented at various venues.

The list of long papers includes:

- Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2018). 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). Cited as (Trinh et al., 2018).

- Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2019a). Capturing Dialogue State Variable Dependencies with an Energy-based Neural Dialogue State Tracker. In Proceedings of the SIGDIAL 2019 conference (pp. 75–84). Cited as (Trinh et al., 2019a).

- Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2019b). Energy-Based Modelling for Dialogue State Tracking. In Proceedings of the 1st workshop on NLP for conversational AI (pp. 77–86). Cited as (Trinh et al., 2019b).

- Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2020a). Energy-based Neural Modelling for Large-Scale Multiple Domain Dialogue State Tracking. In

Proceedings of the 4th workshop on structured prediction for NLP (SPNLP) (pp. 33–42). Cited as (Trinh et al., 2020a).

- Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2020b). F-Measure Optimisation and Label Regularisation for Energy-Based Neural Dialogue State Tracking Models. In Proceedings of the 29th international conference on artificial neural networks (ICANN) (pp. 798–810). Cited as (Trinh et al., 2020b).

The list of extended abstracts is as follow:

- Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2017). Incremental Joint Modelling for Dialogue State Tracking. In Proceedings of the 21st workshop on the semantics and pragmatics of dialogue (SemDial) (pp. 176–177). Cited as (Trinh et al., 2017).

- Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2019c). Investigating Variable Dependencies in Dialogue States. In Proceedings of the 23rd workshop on the semantics and pragmatics of dialogue (SemDial) (pp. 195–197). Cited as (Trinh et al., 2019c).

The list of position papers is presented below:

- Trinh, A. D. (2017). Dialogue Management Modelling. In Proceedings of the 13th workshop on spoken dialogue systems for PhDs, postdocs & new researchers (YRRSDS) (pp. 23–24). Cited as (Trinh, 2017).

- Trinh, A. D. (2019). Dialogue State Tracking. In Proceedings of the 15th workshop on spoken dialogue systems for PhDs, postdocs & new researchers (YRRSDS) (pp. 18–19). Cited as (Trinh, 2019).

1.4 Structure of Dissertation

The research in this dissertation is structured as follows:

- **Chapter 2** presents an extensive literature review into the research field. This chapter is split into three main areas. The first part is the review of the representations of human-machine conversations, in particular the dialogue representations used in various task-oriented systems. In the second part, I describe and critique state-of-the-art dialogue state tracking methods. Lastly, I review the application of structured prediction methods in dialogue processing.

- **Chapter 3** outlines the main evidence for the assumption that structural properties are present in dialogue states. To do so, an investigation of the inter-slot dependencies among dialogue slot types in a number of task-oriented dialogue corpora is conducted. This finding provides a strong motivation for my further work on integrating slot dependencies into the dialogue state tracking process.

- **Chapter 4** presents the first experimental attempt to study the relationships between slots in dialogue data. In this chapter, a multi-task learning

approach is designed to engage slot dependencies as learning features into the dialogue state tracking process. Here the contribution of these slot dependencies in comparison with other deep learning methods that ignore these features is highlighted.

- **Chapter 5** presents the second experimental attempt that explicitly studies the structural properties of dialogue states. In this chapter, an energy-based learning method is proposed for the dialogue state tracking task. This approach explicitly accounts for dialogue slot dependencies in the learning process, thus making the dialogue state tracking task a structured prediction problem.

- **Chapter 6** studies the dialogue state principles in the task-oriented dialogue domains. In order to do so, an approach to enhance the performance of the energy-based dialogue state tracker from the previous chapter is proposed. This approach consists of two elements: an improvement in the mathematical formulation for the energy-based learning method, and the enforcement of dialogue state constraint rules.

- **Chapter 7** presents a generalisation of the proposed structured prediction method when applied to tracking dialogue states for multiple dialogue domains. Moving from a single domain to multiple domains increases the complexity of the dialogue state tracking task, hence challenging the efficiency of any proposed methodology. Nevertheless, the structural properties of natural

language are universal, hence the proposed energy-based learning technique should be effectively used in any scale.

- **Chapter 8** summarises the contributions of this dissertation and proposes a number of future research directions.

Chapter 2

Literature Review

Natural Language Processing (NLP), and in particular dialogue processing, has a long history of study. In this chapter, the literature review focuses on three areas of NLP that underpin the main research question as introduced in the previous chapter. First, it addresses the topic of dialogue state representations in Section 2.1, which I see is a core underpinning of dialogue understanding. Second, the state of the art in dialogue state tracking methods are reviewed in Section 2.2 in order to outline the trends and movements in this very important technology domain. Finally, Section 2.3 provides an overview of structured prediction applications in dialogue research, because, as has been outlined in the previous chapter, my argument is that understanding the role relationship between concepts under discussion is key to improving the language understanding pipeline in dialogue.

2.1 Dialogue State Representations

Conversation representations are crucial to dialogue systems. The analysis and representation of information within discourse has a long tradition, with foundational work in this space including the focus stack model (Grosz & Sidner, 1986), Discourse Representation Theory (DRT) (Kamp & Reyle, 1993; Kamp et al., 2010) and distributional semantics (Clark et al., 2016; Fagarasan et al., 2015; Coecke et al., 2010). These works and many of their derivatives are focused on creating and maintaining data structures that enable reference resolution within a (linguistic) discourse (Kelleher, 2003). Consequently, this tradition of work is outside the scope of this dissertation which is focused on dialogue state tracking as distinct from reference resolution.

Overall, natural language understanding can be represented in the logical form, that in turn can be used to represent conversational language. However, the specifications of dialogues require their representations to contain rich information with respect to the language exchange. In particular, modern approaches to dialogue representations for task-oriented dialogues are based on more specific architectures called **dialogue state** or **belief state** architecture (Jurafsky & Martin, 2020).

Commonly, belief states in task-oriented dialogues are structured with dialogue acts followed by a (possibly empty) sequence of dialogue act items, for example

the CUED dialogue state format (Schatzmann, 2008)[1]:

$$acttype(attr1 = value1, attr2 = value2, ...)$$

where *acctype* is the type of dialogue act defined in the system taxonomy, and the pairs $(attr, value)$ give the details of dialogue states by providing more information related to the user intents.

Inheritedly, in this structure format dialogue states are constructed with two main components: dialogue acts, that represent the interactive function of the turn in the dialogue flow; and dialogue items, that represent attributes of user intents in the domain context (Jurafsky & Martin, 2020). All the components play roles in representing dialogue state. While the act types indicate intention of users in a semantic context, the attributes and values provide precise information. Dialogue acts are usually universal, while the attributes and values vary domain by domain. In task-oriented dialogue systems these attributes are usually predefined in an ontology. Classifying dialogue acts and predicting dialogue attributes-values of dialogue state representations correctly are equally important.

2.1.1 Dialogue Acts

Task-oriented dialogue systems make use of dialogue acts to represent the underlying meaning of interactions with users. It is understood that different dialogue

[1]The complete CUED dialogue act list was reproduced and presented in Appendix A of Thomson (2009)'s PhD dissertation.

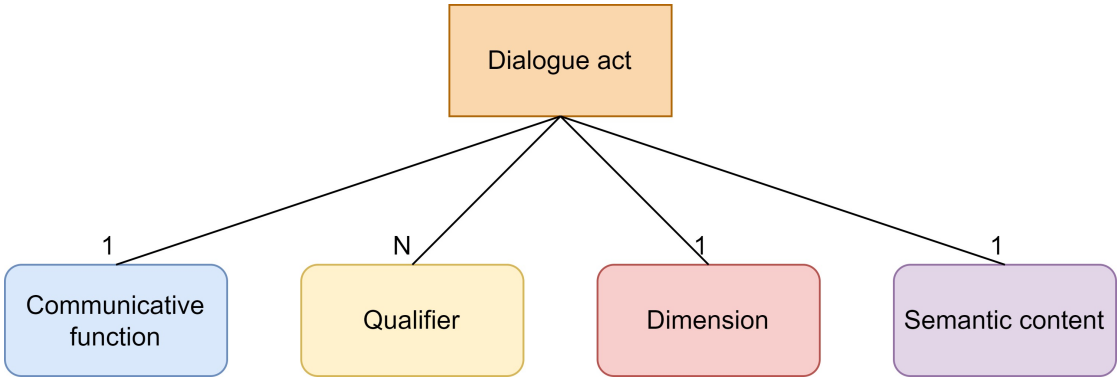

Figure 2.1: ISO 24617-2 dialogue act properties (Bunt et al., 2017).

systems require different types of acts, hence various dialogue act taxonomies have been developed to fit the purpose. It is also understood that these acts are designed for specific tasks in the domains of those systems. Examples of dialogue act taxonomies include DIT (Bunt, 1994), DIT++ (Bunt, 2009), DAMSL (Allen & Core, 1997), MRDA (Shriberg et al., 2004), HCRC Map Task (Anderson et al., 1991), Vermobil (Alexandersson et al., 1998), and SWBD-DAMSL (Jurafsky et al., 1997).

However, since dialogue acts are designed to fit different tasks, the question of having a universal system is challenging. In early work, Traum (2000) raised a number of questions on dialogue act taxonomies around the issues of defining taxonomies, new uses, and standardisation efforts for a discussion. More recently, a semantic scheme of dialogue acts, ISO24617-2 (2020), has been proposed by Bunt et al. (2017, 2020) to standardise the requirements and functionalities of dialogue acts. The ISO scheme presents 4 main aspects of dialogue acts. Figure 2.1 demonstrates the main aspects of dialogue acts and the entity relations between them as set out in that proposal.

- **Dimension**, also described as a category of the semantic content of dialogue, is an annotation scheme to assign communicative functions to dialogue segments. The ISO 24617-2 scheme supports a multidimensional annotation, and defines 10 dialogue act dimensions, where a dialogue act belongs to a single dimension. The list of these dimensions and their description are presented in Table 2.1.

- **Communicative function** expresses the nature of conversational interactions between parties, for example the agreement or disagreement between speakers. The ISO 24617-2 scheme defines more than 60 communicative functions, that are split into 2 groups of general purpose functions and dimension specific functions. However, the number of communicative functions is not fixed. Additional functions can be added if they follow the taxonomy of communicative functions.

- **Qualifier** of a dialogue act is a combination of attribute-value pairs that were mentioned above under the term dialogue item. These items are added to show that a dialogue act is performed conditionally with the conditions are set by the act qualifier. The task of identifying these qualifiers for dialogue acts is also important, and will be detailed in the next section (Section 2.1.2).

- **Semantic content** is the additional semantic information such as semantic roles, time and space information, and the annotation of rhetorical relations.

The ISO 24617-2 scheme assumes that a dialogue act has 1 communicative

Table 2.1: The 10 dimensions of dialogue acts in the ISO 24617-2 scheme based on the work by Bunt et al. (2020, 2017).

Dimension	Description
1. Task	Dialogue acts of this dimension move the task flow forward, that in turn motivates the conversation.
2. Auto-Feedback	Dialogue acts provide the feedback of previous utterances processed by the current speaker.
3. Allo-Feedback	Dialogue acts give the feedback of previous utterances processed by the current addressee.
4. Turn management	These acts are in charge of managing the turns of dialogue, that includes obtaining, keeping, releasing, or assigning the right to speak.
5. Time management	These acts manage the use of time in the interaction between user and system.
6. Contact management	These acts manage the structuring of the dialogue and the progression of topics.
7. Discourse structuring	Dialogue acts of this dimension handle the structuring of dialogue, that includes but not limited to topic management, opening and closing dialogues and subdialogues.
8. Own communication management	Dialogue acts represent the speaker's editing actions of their contribution.
9. Partner communication management	Dialogue acts represent the speaker's editing actions of another speaker's contribution.
10. Social obligations management	Dialogue acts are responsible for social functions such as greeting, introduction, apologising, thanking, and farewell.

function, 1 dimension, possibly multiple qualifiers, and 1 semantic content (as seen in Figure 2.1).

There also exist different types of relations between dialogue acts:

- **Dependence relations** express semantic relations between dialogue acts. The ISO 24617-2 standard defines 2 types of dependence relations in dialogue: functional dependence relations, and feedback dependence relations. The functional dependence relations occur with responsive dialogue acts such as

question-answer, while the feedback dependence relations stick with feedback dialogue acts such as explanation and interpretation.

- **Rhetorical relations** indicates how two dialogue acts and their semantic contents are related to each other.

As mentioned above, the taxonomy of dialogue acts can evolve through time reflecting increases in modelled conversation complexity. For example, *Task Management*, a dimension proposed in DAMSL, has been identified for potential future inclusion in the ISO scheme (Bunt et al., 2020). Hence, such standardisation should be periodically reviewed and updated.

2.1.2 Dialogue Attributes and Values

Dialogue attribute-value combinations play the role of qualifiers of dialogue acts to express the conditions with which the dialogue acts are performed during conversation. The task of identifying these attributes and values in a current dialogue turn, also called the slot-filling task, is a special case of a supervised semantic parsing task in the broader natural language processing domain. This task requires that in each turn of a dialogue, a set of domain-specific slot and value pairs is correctly classified, such that these indicate the user intent. A popular example of this task was captured in the first three competitions of the Dialogue State Tracking Challenge (DSTC) series (J. Williams et al., 2013; Henderson, Thomson, & Williams, 2014a,b).

Task-oriented dialogue systems often predefine the set of dialogue slots and

Table 2.2: An example of DSTC2 dialogue state slot and value combinations.

Utterance:	*I am looking for a fancy French restaurant in city centre.*	
Slot & Value:	food	french
	price range	expensive
	area	centre

their allowed values in advance in an ontology; hence the task of slot-filling is fully supervised learning. For example, the ontology of the second dialogue state tracking task defines 4 slots (*food*, *price range*, *area*, and *name*) in a restaurant information domain, that in turn have a set of predefined values each. An example of the slot filling task in the DSTC2 domain is presented in Table 2.2. Depending on the domain and data, the complexity of tracking different slots is different, for instance tracking *food* values may be harder than *price range* or *area*.

There are various methods to tackle the task of slot-filling for task-oriented dialogue systems, ranging from rule-based to machine learning-based approaches. The review of these methods is detailed later in Section 2.2.

2.1.3 Dialogue States

Dialogue states are used to denote the full representations of a particular point in a dialogue. Originally, the meaning of dialogue states was deeply related to linguistic aspects of language representations. However, the loosely modern use of the term *dialogue state* now indicates mainly user intents during the conversation, especially in the case of task-oriented dialogue systems.

Tracking dialogue states in a dialogue system includes the subtasks of classify-

ing the user's most recent dialogue act and identifying the current state of values for each slot. Here, the dialogue states must not only include current slot value filling results, but also summarise all the user's constraints up to the current turn. In many dialogue systems, these two tasks are handled by different components, such that the dialogue act classification task inherits the result of the language understanding unit, and the slot filling task is dealt with by the dialogue manager. Hence, the dialogue act classification task is often trivial in task-oriented dialogue systems. The main task of dialogue state tracking itself is thus a slot-filling task in most cases.

For example, the DSTC2 restaurant information domain defines dialogue state as the combination of three components for each dialogue turn, but yet these do not include the dialogue act classification task (Henderson, Thomson, & Williams, 2014a):

1. The goal constraint for each informable slot such as *food* and *price range*.

2. A set of Requested slots that are queried by users, the results of which should be informed by the system in the next turn.

3. A search method that indicates how users communicate with the system during the current turn.

An example of dialogue states in the DSTC2 domain is presented in Table 2.3. In this example, the main focus is the constraints of dialogue slots and values, as well as the Requested slots based on the user's query. The dialogue states are joint

Table 2.3: An example of DSTC2 dialogue states following user utterances.

System:	Hi. How can I help you?
User:	I want Chinese food near city centre. *inform(food=Chinese, area=centre)*
System:	There is Beijing restaurant for your query.
User:	What is the address and phone number? *inform(food=Chinese, area=centre); request(address), request(phone)*

representations of different components in the dialogue domain.

A good dialogue system must have a mechanism to track dialogue states accurately following the sequence of a dialogue, and adjust dialogue states based on the new observations as time goes by. While this is simple to explain, achieving robust results is not trivial.

2.1.4 Dialogue Frames

Dialogue frames can be interpreted as an extended version of slot-based dialogue states (Asri et al., 2017). Similar to dialogue states, the dialogue data is expressed in terms of dialogue acts, dialogue slots, and slot values. The difference between dialogue frames and dialogue states lies in the definition of newly defined semantic dialogue frames. A semantic dialogue frame consists of 4 components:

1. **User requests** – slots whose values are requested by the user for this frame. This is similar to the subtask of requestable slots in the DSTC2 dialogue state specification.

2. **User binary questions** – user questions that include the constraints of slot types and slot values.

3. **Constraints** – slots with particular values set by either the user or the system. This component is similar to DSTC2 Joint goals. The difference is that the constraints in frames can be also set by the system, while in DSTC2 only the user can impose the constraints.

4. **User comparison requests** – slots whose values the user wants to know and compare in multiple frames. Based on the presence of this component, the dialogue representations at any point can be multiple frames.

Since a dialogue can be represented in the multiple frame format, the frame tracking task requires the tracking of multiple frames simultaneously. In this sense, the dialogue frame tracking task can be considered an extension of the dialogue state tracking task (Henderson, 2015). In the dialogue state tracking task, the dialogue history is compressed into a single semantic frame, and the system updates dialogue states within this frame throughout the conversation. On the other hand, in the frame environment, the dialogue history is stored in multiple frames, that allows the system to refer to any previous state without erasing the current frame. For example, if the user provides a new constraint, a new semantic frame is created within the system and becomes active as the true current dialogue representation. Hence, the dialogue frame tracking task is significantly more difficult than the dialogue state tracking task, as it requires identifying and updating the active frame as well as maintaining all the frames in the dialogue history for each dialogue turn.

It is worth noting that in addition to the above there are many other types of

dialogue states such as **Information State** and **State Machine Style** dialogue representation. Here, the information state approach involves a flexible approach to representation of context that aims to recognise a set of dialogue acts along with their dialogue attributes and values (Traum & Larsson, 2001, 2003). Meanwhile, the state machine style is used to model the behaviours of systems rather than of users (Jurafsky & Martin, 2020). However, since the focus of my research is on the slot-filling type of dialogue states, the two mentioned types of dialogue states are out of the scope of this dissertation. The following section presents a review of dialogue state tracking methods on slot-based dialogue states.

2.2 Dialogue State Tracking Methods

Generally Dialogue State Tracking (DST) in task-oriented dialogue systems is relatively straightforward and less complicated than in general purpose conversational systems as the domain is more constrained (Deriu et al., 2021). In task-oriented dialogue systems, slots and their possible values are predefined in a domain-specific ontology, and dialogue states are constructed from these slot value pairs (J. D. Williams et al., 2016). For example, a valid dialogue state in the restaurant information domain might be {*food=chinese, price range=cheap, area=centre*}.

Over the past 30 years, dialogue state tracking techniques have moved from hand-crafted rules (Zue et al., 2000; Larsson & Traum, 2000) towards deep learning-based methods (Balaraman et al., 2021). Today's state-of-the-art dialogue systems benefit from deep learning-based dialogue state trackers (Feng et al., 2021; Tian

Table 2.4: An overview of different types of dialogue state tracking methods.

	Dialogue State Tracking Methods			
	Rule-based	Generative	Discriminative	Hybrid
Parameter fine-tuning	Manual	Automatic	Automatic	Semi-automatic
Data requirement	No	Yes	Yes	Yes
Unseen situation generalisation	No	Yes	No	Yes
Performance	Limited	Good	Good	Good

et al., 2021) or those implemented with a mixture of deep learning and other techniques (Dai et al., 2021; S. Li et al., 2021; T. Yu et al., 2021). Generally there are many ways to categorise research techniques for dialogue state tracking, for example neural versus non-neural methods. J. D. Williams et al. (2016) and Henderson (2015) split dialogue state tracking techniques into three groups: hand-crafted rules, generative methods, and discriminative methods. This approach to categorisation captures the overall summary of a great number of proposed models. However, based on the recent trend in research, hybrid methods that combine two or more different techniques should be added into the list as a fourth group. An overview of these technique groups is presented in Table 2.4.

The following sections will present a brief overview of each of the four technique groups mentioned above, followed by an overview of public dialogue corpora for the dialogue state tracking task and the state of the art.

2.2.1 Rule-based Systems

Early dialogue systems such as the MIT JUPITER weather information system (Zue et al., 2000) and the TRINDI dialogue move engine toolkit (Larsson & Traum, 2000) used hand-crafted rules for dialogue state tracking. The basic concept of hand-crafted rules for dialogue state tracking is that they are designed to map the previous dialogue state s and the current language understanding hypothesis u to a new dialogue state s':

$$s' = F(s, u) \tag{2.1}$$

where $F(\cdot)$ is the set of rules manually defined for the dialogue state tracking task.

Applying the rules directly as in Equation 2.1 can only track a single dialogue state at a time. Therefore a modification is needed to track multiple dialogue states in parallel. This modification is designed to compute the belief score $b(s')$ of the new dialogue state s' rather than just to purely map the previous dialogue state s to the new state s'. With this rule, Equation 2.1 is reformulated as follows:

$$b(s') = F(s, u) \tag{2.2}$$

The rule-based models with this modification (Fix & Frezza-Buet, 2015; Sun et al., 2014a; Z. Wang & Lemon, 2013) have been shown to overcome some language understanding errors by using a language understanding N-best list with confidence scores.

Regardless of tracking single or multiple dialogue states, the common point

of rule-based approaches is that the models require the language understanding component to provide the semantic representations of user utterances. However, in some cases the language understanding component might not be reliable due to its error rates. Therefore to ensure the performance of dialogue state tracking models, it is also important to improve the language understanding component itself (Kadlec et al., 2014; Sun et al., 2014b; Zhu et al., 2014). For example, Kadlec et al. (2014) used the confidence scores to correct language understanding hypotheses, while Sun et al. (2014b) and Zhu et al. (2014) developed their own semantic parsers and trained them on speech recognition hypotheses.

Although using hand-crafted rules has the advantage that the systems do not require any data to train, these systems have a crucial limitation that formula parameters are not derived directly from real dialogue data and require careful manual tuning (J. D. Williams et al., 2016). Ultimately, this limitation motivates the use of data-driven methods such as machine learning and deep learning.

2.2.2 Generative Systems

Generative dialogue state tracking models typically process dialogue with a Bayesian mechanism. The probability of the current dialogue state *s'* is computed based on the previous dialogue state *s*, the system action *a*, and the new observation *o* of user action by applying Bayesian inference. For example, a simplified Bayesian

equation (Young et al., 2010) is formulated as in Equation 2.3.

$$b(s') = \eta P(o'|s', a) \sum_{s} P(s'|a, s) b(s) \tag{2.3}$$

where $b(s)$ is the previous distribution over dialogue states, $b(s')$ is the updated distribution, $P(s'|a, s)$ is the probability of dialogue state changing to s' given the current state s and the system action a, $P(o'|s', a)$ is the probability of the new observation given the new dialogue state s' and the system action a, and η is a normalising constant.

Equation 2.3 shows that if the new belief score is calculated directly based on the current dialogue state, many other factors of the ongoing conversations would be ignored, for example dialogue history and context. Therefore there exist various techniques to modify Bayesian networks for generative dialogue state tracking systems with specific settings such as including a term accumulating for dialogue history (J. D. Williams & Young, 2007; S. Lee & Stent, 2016), conditional probability terms expressing context (DeVault & Stone, 2007; Perez & Radford, 2016), or goal change handling (B.-J. Lee et al., 2014). In detail, S. Lee & Stent (2016) used a task frame parser to handle the input and a context fetching model to deal with dialogue history before updating the dialogue state. Both focusing on the dialogue context, B.-J. Lee et al. (2014) included a goal change handling model and a system-user action pair weighting model in their DST system to compute hidden information state, while Perez & Radford (2016) developed a probabilistic matching model to extract mentions, search information in an ontology and rank

the candidates.

In the early days of dialogue research, there were many generative dialogue state tracking models which enumerated all possible dialogue states (Roy et al., 2000; B. Zhang et al., 2001; Meng et al., 2003; J. D. Williams et al., 2005), but soon they faced a challenge due to an enormous number of dialogue states. As a solution, various techniques were proposed to approximate Bayesian computation. A few related works proposed the grouping of dialogue states into partitions to present belief distributions called Hidden Information State (HIS) (Young et al., 2007, 2010; J. D. Williams, 2010; Gasic & Young, 2011). In the work proposed by K. Kim et al. (2008), a frame-based belief state representation was used to reduce the complexity of belief update. Meanwhile Mehta et al. (2010) represented the space of user intentions, i.e. dialogue states, in the form of Probabilistic Ontology Trees (POT) and performed computation of dialogue states only for the m-best most probable cases. Similarly, dynamic probabilistic ontology trees were used to track dialogue states and to capture dialogue history (Raux & Ma, 2011; Ma et al., 2012). Another approximation technique was proposed by Thomson & Young (2010), where the method was based on the loopy belief propagation algorithm (Ihler et al., 2005).

Overall, generative methods yield better results than hand-crafted rules, and have the ability to generate unseen situations in the training process, which is a big advantage for open-domain spoken dialogue systems.

2.2.3 Discriminative Systems

Discriminative approaches were applied to the dialogue state tracing task as early as the work by Bohus and Rudnicky (Bohus & Rudnicky, 2006), where the authors developed a machine learning-based multi-class logistic regression model. Discriminative DST models compute belief scores for dialogue states based on observing dialogue properties, that can be formulated, for example, as in Equation 2.4:

$$b(s') = P(s'|f_{a,o,h}) \tag{2.4}$$

where $f_{a,o,h}$ are features extracted from dialogue data such as the system response a, speech recognition or language understanding output o, and dialogue history h.

The key issue of discriminative methods is to build an effective mechanism to extract dialogue features f that give the dialogue state tracker the advantage of deciding belief state based on a large number of features. For specific domains, discriminative dialogue state tracking models often outnumber other methods and yield better results (S. Lee, 2013; Metallinou et al., 2013). Proposed discriminative methods for DST include conditional random fields (H. Ren et al., 2013), Markovian maximum entropy models (K. Yu et al., 2015), and neural networks (Henderson, Thomson, & Young, 2014b; H. Shi et al., 2016a), among others.

In general, dialogues can be cast as sequential data, therefore there exists various approaches accommodating sequential dialogue data in the dialogue state tracking task (Feng et al., 2021; J. Zhao et al., 2021; Jagfeld & Vu, 2017; S. Kim &

Banchs, 2014). Considering this sequential labelling technique, many researchers apply linear-chain conditional random fields to the dialogue state tracking task with different feature extraction techniques (S. Kim & Banchs, 2014; S. Lee & Eskenazi, 2013; Ma & Fosler-Lussier, 2014). Another approach is Markovian maximum entropy model for dialogue state tracking, where the previous turn prediction can be used as features for the current turn estimation (H. Ren et al., 2014a,b).

It is worth noting that there have been also several attempts to solve dialogue state tracking tasks as non-sequence problems such as mapping dialogue states and hypothesis-specific features (Metallinou et al., 2013), structured discriminative method (S. Lee, 2013), deep neural networks (Henderson, Thomson, & Young, 2013), and web-style ranking (J. D. Williams, 2014).

Deep learning-based models are commonly used for the dialogue state tracking task as they are helpful in processing sequential dialogue data. Many models are based around Recurrent Neural Network (RNN) (Yoshino et al., 2016; Mrksic et al., 2015; Henderson, Thomson, & Young, 2014b) and Convolutional Neural Network (CNN) (Mrksic et al., 2017; H. Shi et al., 2016b,a) architectures. When the attention mechanism (Bahdanau et al., 2015) was proposed for various natural language processing tasks and achieved great performances, there were attempts to engage this mechanism to track dialogue states (Hori et al., 2016; Jang et al., 2016). Since then, there has been a great amount of research with the further development of the attention mechanism. One of the directions, for example, is developing models with self defined global and local attention terms over dialogue

features that are accounted for in predicting dialogue states (Zhong et al., 2018; Nouri & Hosseini-Asl, 2018). Balaraman & Magnini (2019) proposed an attention-based dialogue state tracker, that consists of a global encoder and a number of slot-attentive decoders, for the purpose of scalable deployment in real-world applications. The non-autoregressive dialogue state tracker (Le et al., 2020), and its further development, the improvised non-autoregressive model (B. Li et al., 2021), were also based on the attention mechanism and multiple encoder-decoder architecture.

A further trend is to develop transformer-based dialogue state trackers (Balaraman et al., 2021). Arguably the best known transformer architecture is Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). This architecture has been implemented in a great number of dialogue works, for example the work proposed by Zeng & Nie (2021), Lai et al. (2020), and Ruan et al. (2020). Here, the main approach is that the transformer architecture is used to encode the dialogue context and current turn input, then a number of task-specific classifiers are used to produce dialogue state predictions (Gulyaev et al., 2020; Zheng et al., 2020).

From another perspective, there is also extensive research on models using recurrent neural networks to process speech recognition hypotheses in an incremental fashion (Zilka & Jurcicek, 2015a,b; Platek et al., 2016; Jagfeld & Vu, 2017). These incremental trackers operate directly on the speech recognition output, therefore they can avoid errors produced by the language understanding unit. In detail,

incremental dialogue state trackers proposed by Platek et al. (2016) and Zilka & Jurcicek (2015a,b) are capable of operating on the top live speech recognition hypothesis on a word-by-word basis and producing live dialogue states in real time. Meanwhile, Jagfeld & Vu (2017) encoded hypotheses of word confusion networks to predict dialogue states in a fixed time frame smaller than single word units. However, word confusion networks contain a lot of speech-to-text errors that in turn limit the performance of this incremental system.

Discriminative methods outperform both generative methods and hand-crafted rule-based systems, but still have their own disadvantages in that they are effective only when there are enough training data. When there are not enough labelled data, discriminative models often include some generative techniques to deal with unseen training situations (Feng et al., 2021; C.-S. Wu et al., 2020, 2019).

2.2.4 Hybrid Systems

Hybrid systems are dialogue state trackers that use more than one specific method from those mentioned above. It is clear that one method standing alone can achieve good results, but with some limitations. For instance, discriminative methods yield better results than other methods in specific domains but suffer from data insufficiency. On the other hand generative methods have the ability to consider many possible outcomes irrespective of whether they are seen or unseen in a dataset. In many cases combining those techniques can show outstanding improvements on the same tasks.

Since there are three general method groups, as seen above, with various sub-methods specifically, there are a huge number of ways to combine them. For example, there are many models proposed with the combination of discriminative and generative methods. In an early example of this, Henderson, Thomson, & Young (2014a) developed an RNN-based discriminative model with an online unsupervised adaptation technique to generate unseen slot value pairs. S. Lee & Eskenazi (2013) proposed a maximum entropy discriminative model combined with generative method and unsupervised prior adaptation. A different combination of a generative model with discriminative re-scoring mechanism was introduced by D. Kim et al. (2013).

Another hybrid combination is to apply manual rules to provide additional inference during dialogue state tracking, as dialogue corpora are normally relatively small in comparison with texts or documents in natural language processing. In fact applying hand-written rules has proven to boost the performance of discriminative and generative models. In Zilka et al. (2013)'s work, hand-crafted rules are implemented in the discriminative maximum entropy model to compute transition probabilities between states. The MSIIP systems proposed by M. Li & Wu (2016) and Y. Su et al. (2016) use a discriminative classifier to generate similar semantic structures to dialogue states from the utterance of each turn, then to apply rule-based strategies to predict dialogue states. K. Yu et al. (2015) proposed a hybrid framework based on constrained Markov Bayesian polynomials to formulate a universal rule-based system for the dialogue state tracking task, which

allows data-driven rule generation. Similarly, Mrkšić & Vulić (2018) applied two variants of statistical Markovian update techniques on top of a neural tracking architecture to create a robust framework for building resource-light dialogue state tracking models. If the hand-crafted rules are differentiable, the hybrid tracker can be trained in an end-to-end fashion (Vodolan et al., 2017, 2015).

Following recent trends in attention-based modelling, many hybrid systems have been developed by combining a deep attentive neural architecture with various techniques for robust results (C.-S. Wu et al., 2020, 2019; Xu & Hu, 2018). Since attention-based models are discriminative systems, the authors incorporated generative techniques on top to handle unseen situations in training data. In detail, C.-S. Wu et al. (2019) employed a soft-gated pointer-generator copying mechanism to add a distribution over the vocabulary and a distribution over the dialogue history into a single output distribution. Later, based on this architecture, C.-S. Wu et al. (2020) proposed a self-supervised approach to dialogue state tracking with two auxiliary tasks: preserving latent consistency and modelling conversational behaviour. On the other hand, Xu & Hu (2018) developed an end-to-end sequence-to-sequence architecture based on Vinyals et al. (2015)'s pointer networks to extract unknown slot values in dialogue states.

Transformers have demonstrated big advantages in learning language representations in various NLP tasks including dialogue state tracking. Hybrid trackers can make use of transformers to encode dialogues, then apply various rule-based and generative techniques to predict dialogue states. Similar to transformer-based

discriminative systems featured in the previous section, a great number of hybrid models have been developed based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). S. Kim et al. (2020) proposed a BERT-based model with a selectively overwritting strategy for more efficiency in dialogue state tracking, while Gao et al. (2019) applied a reading comprehension approach to a BERT-based architecture. In another BERT-based model, J.-G. Zhang et al. (2020) proposed a dual strategy that consists of a slot-gate classification module and a set of tracking rules for different types of slots. Another technique is to apply span detection modelling on top of a BERT-based encoder such as the work by X. Shi et al. (2020), while Chao & Lane (2019) went further with both a span prediction module and a rule-based update. Coming at it from a different perspective, Heck et al. (2020) applied the triple copy strategy on top of a BERT-based context encoder. This model became a baseline for a series of further developments where authors incorporated more techniques such as curriculum learning (Dai et al., 2021), conversational semantic parsing (T. Yu et al., 2021), and controllable counterfactual generation (S. Li et al., 2021). It is also worth noting the attention mechanism is useful not only at the encoding stage, but also in the decoding process such as with the use of a pointer generator (Feng et al., 2021).

Other transformer architectures such as the Generative Pre-trained Transformer (GPT) (Radford et al., 2018, 2019), the standard bidirectional encoder-decoder transformer BART (Lewis et al., 2020), and the text-to-text transfer transformer T5 (Raffel et al., 2020), are also used to develop hybrid dialogue state

trackers that yield competitive results. For example, Hosseini-Asl et al. (2020) developed a GPT-2-based tracker with causal language modelling as a generative decoder. Meanwhile, Tian et al. (2021) developed a model based on the GPT-2 architecture and a 2-stage generation mechanism. Moreover, the BART and T5 architectures were used in the work by Z. Lin et al. (2020), in which the authors proposed Levenshtein belief spans for efficient dialogue state tracking with a minimal generation length.

Another recent trend for dialogue state tracking methods is to make use of additional information in the training process. W. Lin et al. (2021) developed a transformer-based dialogue state tracker enhanced with a knowledge graph. Similarly, L. Zhou & Small (2019) incorporated a dynamic knowledge graph into a question answering style dialogue state tracker. Furthermore, extra information can be provided as the guiding schema for dialogue state tracking as in the case of Schema-Guided Dialogue (SGD) (Rastogi et al., 2020a). As transformers had been widely used, many models had the guiding schema embedded as extra features for the encoders (Balaraman & Magnini, 2020; Ruan et al., 2020; Gulyaev et al., 2020; Zheng et al., 2020; X. Shi et al., 2020). At the same time, various non-transformer approaches also yield competitive results for schema-guided dialogue such as machine reading comprehension model (Ma et al., 2020), question answering with data augmentation approach (Mou et al., 2020), and the tracker based on attention graph (L. Chen et al., 2020).

Besides those above, there are several systems with original approaches which

can be considered hybrid. Many dialogue state trackers were enhanced with graph-based architectures including the state graph (Zeng & Nie, 2020), the hierarchical task graph (Shen & Wang, 2020), and the combination of the state graph and historical state copy mechanism (P. Wu et al., 2020). Goel et al. (2019) proposed an approach of flexible dialogue state tracking for each slot type. A special case study of hybrid systems is the combination of the reading comprehension method (Gao et al., 2019) and the joint dialogue state tracking technique (Goel et al., 2019), that yields better performance than both the original approaches. Another original approach is to explore dialogue representation based on the author-topic model and combine it with support vector machine classification to track dialogue states (Dufour et al., 2016). Finally, it is also worth mentioning the robust hybrid dialogue state tracker (Dernoncourt et al., 2016), that was developed using elaborate string matching, coreference resolution tailored for dialogues and a few other improvements to operate on large ontologies.

Hybrid methods may also include systems that contain only the output of different models; an example of this is the SJTU system proposed by Sun et al. (2014b). The SJTU system is the combination of a deep neural network, a Markovian maximum entropy model, and a rule-based model. Each of these models produces the best result for each component of dialogue states. Then, the ultimate dialogue states are produced by combining the best output of the mentioned models. This is ultimately an ensemble approach, and the literature is full of many such approaches.

Hybrid methods generally solve the problem that discriminative methods suffer from. Applying generative improvement techniques or hand-crafted rules gives the model the ability to generalise distributions over unseen slot value pairs. However, hybrid methods still have their own limitation, for instance, the need to fine tune parameters to match different components in the model. Hybrid systems also require in-depth analysis into the effects and benefits of ensembling different types of techniques.

2.2.5 Domain Specific Dialogue Corpora

With the research direction changing towards advanced technologies such as deep learning, data-driven approaches have gained more popularity in recent years. However, to the best of my knowledge there exist only a few public domain-specific dialogue corpora that are annotated for dialogue state research.

The Dialogue State Tracking Challenge (DSTC) series is a common testbed for explicit research in dialogue state slot-filling tasks. The tasks in DSTC are proposed to range from simple to complex. Early editions of the DSTC required systems to track static user intents in human-machine dialogues (J. Williams et al., 2013), while late editions changed to tracking dynamic user intents (Henderson, Thomson, & Williams, 2014a,b). In the later editions trackers were supposed to operate on human-human conversations (S. Kim, D'Haro, Banchs, Williams, & Henderson, 2016; S. Kim, D'Haro, Banchs, Williams, Henderson, & Yoshino, 2016). To date, the DSTC2 dataset is the most popular dataset from this series

used in research.

The main task of DSTC, *Joint goals*, is to predict a value for each informable slot in the ontology. Informable slot and value pairs are in fact the main intents that users provide. A side task in relation to slots in the domain is *Requested slots*, when users ask for some specific information from the system. For example, in the DSTC2 restaurant information domain, users can tell the system what type of food and what area in town they want (*Joint goals*), while they can also ask for the address and phone number of the restaurant they are interested in (*Requested slots*).

DSTC data were gathered from spoken conversations that contain a lot of noise due to the imperfection of the automatic speech recognition unit used in the collection system. Meanwhile, chat-based dialogues such as those collected in the Wizard-Of-Oz (WOZ) dataset[2] can avoid those errors (Wen et al., 2017; Mrksic et al., 2017). Similar to the DSTC2 data, this WOZ corpus covers the restaurant information domain.

The similarity of the DSTC2 and WOZ datasets also lies in the fact that they cover only a single domain with a reasonably small number of dialogues. More recently, a multi-domain dialogue corpus, Multi-Domain Wizard-Of-OZ (MultiWOZ), was introduced to remove the limitation of single-domain corpora (Budzianowski et al., 2018). This MultiWOZ dataset is a fully-labelled collection of human-human chat-based conversations, that contains around 10000 dialogues across seven different domains.

[2]From here Wen et al. (2017) WOZ 2.0 dataset will be referred to simply as the WOZ dataset.

After the fifth edition of DSTC, the dialogue state tracking task was discontinued in the competition series until the introduction of Schema-Guided Dialogue (SGD) dataset in DSTC8 (Rastogi et al., 2020b). The SGD dataset is a multi-domain dataset that exceeds the MultiWOZ dataset in all parameters such as the number of domains, dialogues, slots, and values. The data here were created with synthetic implementations and generated semantic dialogue representations. Then these dialogues representations were paraphrased into natural language utterances by a crowd-sourcing service. Therefore, the nature of conversations in this dataset is different from other dialogue corpora mentioned above. Another difference is that in schema-guided dialogues a schema listing the supported slots and intents is provided along with their natural language descriptions for each service, that is based on the data simulation.

As introduced earlier in the chapter, another expansion of the dialogue state tracking task is dialogue frame tracking (Asri et al., 2017). The core idea of dialogue frame tracking is to incorporate memory of all dialogue states in the conversation into frames, so that the system can refer to any previous state at any moment of conversation. In detail, the dialogue frame tracking task has two phases: (i) at each turn of the dialogue the tracker is required to detect if a new frame should be generated; (ii) then it decides which frame in the pool is the best candidate for the dialogue states of this turn. Dialogue frame tracking is thus different from the common dialogue state tracking concept where the systems should focus only on tracking the current dialogue state and ignore previous dialogue

states. The dialogue frame tracking task is available for research via the Frames corpus (Asri et al., 2017).

As seen in the last section, in recent years a large number of approaches have been proposed to solve the dialogue state tracking task. There is unfortunately no way all dialogue state trackers can be directly compared as they operate on different tasks and datasets. However, it can be useful to group tracking systems by the corpora they are based on, and reporting the best approaches. For example, in the DSTC series, the performances of these models are reported on the main task *Joint goals* and side task *Requested slots*. The common evaluation metric used to benchmark dialogue state tracking systems is *Accuracy*, as it was the feature metric proposed for early DSTC competitions. The later challenges, DSTC4 & 5, instead used the F_1 score (Kelleher et al., 2015) as the feature evaluation metric. I compare the top two submitted approaches for each DSTC competition and present them in Table 2.5.

Overall, the top performing models for each DSTC competition are either discriminative or hybrid systems. Although the accuracy metric provides a view on how good dialogue state tracking systems are, it does not contain detailed analysis of the performance. For DSTC2, Smith (2014) presented a comparative error analysis for model entries, that included a set of error types of the results. The finding of the work was that there was no single best approach for the DSTC2 tasks. In detail, the top two trackers in this competition did not yield the best results across all the subtasks: the web-style ranking model (J. D. Williams, 2014) came first on

Table 2.5: Summary of the state-of-the-art dialogue state tracker entries for the DSTC series.

Competition	Model	Performance	
		Joint goals	Requested slots
DSTC1	Discriminative maximum entropy model (S. Lee & Eskenazi, 2013)	0.438	-
	Combined tracking model (D. Kim et al., 2013)	0.345	-
DSTC2	Web-style ranking model (J. D. Williams, 2014)	0.784	0.957
	Word-based RNN model (Henderson, Thomson, & Young, 2014b)	0.768	0.978
DSTC3	Unsupervised RNN model (Henderson, Thomson, & Young, 2014a)	0.646	0.943
	Knowledge-based model (Kadlec et al., 2014)	0.630	0.923
DSTC4	Hybrid tracker with hand-crafted rules (Dernoncourt et al., 2016)	0.579	-
	Hybrid probabilistic framework (M. Li & Wu, 2016)	0.388	
DSTC5	Multi-channel CNN model (H. Shi et al., 2016b)	0.452	-
	RNN model with attention mechanism (Hori et al., 2016)	0.395	-

the *Joint goals* task, while the word-based recurrent neural tracker (Henderson, Thomson, & Young, 2014b) outperformed it on the *Requested slots* task. Henderson, Thomson, & Williams (2014a) also reported the case of ensembling and stacking of all entries to achieve outperforming results over single entries.

During the DSTC4 & 5 challenges, due to the issue of very limited data, participants were allowed to use additional out-of-domain data to boost their models' performance. The top performing trackers overcame the limitation with more ad-

vanced techniques such as a hybrid method (Dernoncourt et al., 2016; M. Li & Wu, 2016) and an attention mechanism (Hori et al., 2016).

For the interested readers, a more detailed list of task-oriented dialogue corpora with dialogue state annotations is presented in the work by Z. Zhang et al. (2020).

In the next section, I will detail the state-of-the-art approaches to dialogue state tracking, broken down by dialogue corpora introduced in this section.

2.2.6 State-Of-The-Art Models

Due to the popularity of the DSTC series, there have been many attempts to improve dialogue state tracking mechanisms and reported results on these competition corpora even many years after the competitions have ended. Therefore we report the performance of state-of-the-art systems on the main task, *Joint goals*, of DSTC in Table 2.6. Here, the DSTC4 & 5 datasets are omitted from this work because of the dataset privacy policy and relative lack of published works.

Later works on the DSTC tasks show that the dominant techniques for dialogue state tracking in task-oriented dialogue systems are still discriminative (Feng et al., 2021; Mrksic et al., 2015; S. Lee, 2013) and hybrid (Vodolan et al., 2017; K. Yu et al., 2015). In particular, Mrksic et al. (2015) showed that the dialogue state tracking models could benefit from training in multiple domains to improve their performance on a single domain.

While the DSTC series is a case study of dialogue state tracking, where a common testbed and evaluation metrics are provided to compare different approaches,

Table 2.6: Summary of the state-of-the-art dialogue state trackers for the DSTC series. * denotes the approach submitted during the competition time.

Dataset	Model	Joint goals
DSTC1	Structured discriminative model (S. Lee, 2013)	0.454
	*Discriminative maximum entropy model (S. Lee & Eskenazi, 2013)	0.438
	Constrained Markov Bayesian Polynomial (K. Yu et al., 2015)	0.402
DSTC2	Sequence-to-sequence model (Feng et al., 2021)	0.850
	Hybrid model with ASR features (Vodolan et al., 2017)	0.796
	StateNet model (L. Ren et al., 2018)	0.755
DSTC3	Multi-domain neural belief model (Mrksic et al., 2015)	0.671
	*Unsupervised RNN model (Henderson, Thomson, & Young, 2014a)	0.646
	Constrained Markov Bayesian Polynomial (K. Yu et al., 2015)	0.634

the series does not have a monopoly on dialogue state tracking datasets. As mentioned earlier, Wen et al. (2017)'s Wizard-of-Oz (WOZ) dataset is a newer dataset that is similar to the DSTC2 corpus as it covers the restaurant search domain. Budzianowski et al. (2018)'s multi-domain Wizard-of-Oz (MultiWOZ) dataset is meanwhile a fully-labelled collection of human-human written conversations that spans over multiple domains and topics. I report the recent works on both the WOZ and MultiWOZ datasets in Table 2.7. Similar to DSTC2, the dialogue state tracking tasks of WOZ require the trackers to produce predictions over *Joint goals* and *Requested slots*. The evaluation is also conducted with the *accuracy* metric.

It is observed that neural network-based methods are effective for both single

Table 2.7: Summary of the state-of-the-art dialogue state trackers for the WOZ and MultiWOZ datasets.

Dataset	Model	Joint goals
WOZ	Amendable generation model (Tian et al., 2021)	0.9137
	Sequence-to-sequence model (Feng et al., 2021)	0.912
	Effective sequence-to-sequence model (J. Zhao et al., 2021)	0.910
MultiWOZ 2.0	Knowledge-aware graph-enhanced GPT2 model (W. Lin et al., 2021)	0.5486
	Transformer model (Zeng & Nie, 2021)	0.5464
	DST-Picklist model (J.-G. Zhang et al., 2020)	0.5439
MultiWOZ 2.1	TripPy + SaCLog model (Dai et al., 2021)	0.6061
	TripPy + CoCoAug model (S. Li et al., 2021)	0.6053
	TripPy + SCoRe model (T. Yu et al., 2021)	0.6048

and multiple dialogue domains and produce the state of the art. In particular, the sequence-to-sequence approaches work well on dialogue data as a special case of natural language generation, as shown in the case study of the WOZ data (Tian et al., 2021; Feng et al., 2021; J. Zhao et al., 2021), On the other hand, transformers have shown a positive impact in the case study of the MultiWOZ 2.0 & 2.1 data. In detail, the BERT-based model (Zeng & Nie, 2021) and the GPT2-based model (W. Lin et al., 2021) are on top of the leader board for the MultiWOZ 2.0 data. Meanwhile, the three hybrid models developed from Heck et al. (2020)'s TripPy system yield the best results on the MultiWOZ 2.1 data (Dai et al., 2021; S. Li et al., 2021; T. Yu et al., 2021).

Table 2.8: Summary of the state-of-the-art dialogue state trackers for the SGD dataset retrieved from the competition report by Rastogi et al. (2020a).

Dataset	Model	Joint goals
SGD	Machine reading comprehension & WD classification model (Ma et al., 2020)	0.8653
	*Team 14 entry (*no published paper*)	0.7726
	Zero-shot BERT-based model (Ruan et al., 2020)	0.7375

The Schema-Guided Dialogue corpus is relatively new to the public. The state-of-the-art approaches for this corpus were reported mainly during the competition time (Rastogi et al., 2020a). There are 25 teams participating in this competition, of which the top three entries are reported in Table 2.8.

The state of the art proposed by Ma et al. (2020) is a hybrid dialogue state tracker that consists of two models for different types of dialogue slots. In detail, span-based and numerical slots are tracked by a machine reading comprehension model, while boolean and text-based slots are predicted by a classification model with wide and deep features. On the other hand, Ruan et al. (2020) proposed to fine-tune a BERT-based model to perform zero-shot dialogue state tracking. This model also consists of a number of modules for different purposes: intent prediction, slot prediction, slot transfer prediction, and user state summarisation. Despite gaining the second highest result, team 14 did not publish their paper at the workshop, therefore their approach remains an unanswered question.

For the Frames corpus, I report the currently available dialogue frame trackers that I am aware of; these are a complex frame tracking model (Schulz et al., 2017) and a recurrent neural network-based baseline tracker (Asri et al., 2017)

Table 2.9: Summary of the state-of-the-art dialogue frame trackers.

Dataset	Model	Performance	
		Slot-based	Act-based
Frames	Frame tracking model (Schulz et al., 2017)	0.764	0.957
	Recurrent neural network-based tracker (Asri et al., 2017)	0.613	0.668

(see Table 2.9). Performance is evaluated with the *Accuracy* metric for *Slot-based* and *Act-based* frame tracking. Slot-based frame prediction is conducted solely on slot value pairs, while act-based frame tracking accounts for the probability of frame references including dialogue acts[3].

The frame tracking model proposed by Schulz et al. (2017) is in fact a hybrid system, that includes a set of encoding rules to boost the performance over the simple multi-task learning-based baseline by Asri et al. (2017). To date there are not many published works on the dialogue frame tracking task.

Overall, the dialogue research community has moved towards developing data-driven dialogue state trackers (J. D. Williams et al., 2016; Henderson, 2015), implementing large scale systems (Dai et al., 2021; S. Li et al., 2021), and training end-to-end trackers (Tian et al., 2021; Feng et al., 2021). However, many methods are still under the influence of the model-centric trend such as the transformer-based models (W. Lin et al., 2021; Zeng & Nie, 2021). These systems often overlook the structural properties of natural language, and in this case the structural properties of dialogue state representations.

[3]An utterance in dialogue contains one or multiple dialogue acts. For example "I want Chinese food" can be represented as *inform(food=Chinese)*. Here the dialogue act is *inform*, the slot is *food*, and the value of this slot is *Chinese*.

2.3 Structured Prediction Applications in Dialogue Research

Deep learning has been a revolutionary technology for Natural Language Processing (NLP) for the last decade with breakthrough results in various tasks (Kelleher, 2019). However, many neural NLP models overlook the structural complexity of human language, which in turn leads to issues in performance. For example, many machine translation systems do not recognise name entities in text and mistakenly translate them in the outcome (Martins et al., 2022). However, as introduced in Chapter 1, many NLP tasks can be characterised as structured prediction problems, wherein the interdependent structures of outputs are taken into account for prediction. Subsequently, structured prediction methods have been widely used in various NLP tasks (Dev et al., 2021). For example, the research project DeepSPIN[4], funded by the European Research Council (ERC), focuses on applying structured prediction methods to three highly challenging NLP applications: machine translation, quality estimation, and dependency parsing (Martins et al., 2022).

Although dialogue processing is a subset of NLP and many dialogue tasks can also be characterised as structured prediction problems, there is a lack of structured prediction investigation in dialogue research overall. The lack of research on structured prediction for dialogue is evidenced from the workshop series on

[4]DeepSPIN: Deep Structured Prediction in NLP. Project website: `https://deep-spin.github.io/`. ERC project website: `https://cordis.europa.eu/project/id/758969`.

Structured Prediction for NLP, which at the time of writing is in the 6th edition and consists of overall 47 published papers, yet only 1 paper presented work on dialogue tasks, which was in fact my own work (Trinh et al., 2020a).

Nevertheless, there have been attempts in the community to implement approaches that account for interdependencies between slots in dialogue tasks. Although it is not necessarily the explicit dependencies among slot outputs that these works consider, they are notable in their own rights. Shu et al. (2019) and Mehri et al. (2019) approached the structural aspect for the system development when they developed structured networks for dialogue systems. These structured systems account for the interdependencies among latent variables by the connections between neural layers of different components in the architecture. On the other hand, there are approaches accounting for the structural properties of language in dialogues. For example, Tseng et al. (2019) proposed to use tree-structured semantics to enhance dialogue language generation, while Kurfali & Ostling (2019) and Z. Liu et al. (2021) made use of the discourse relations and coreference to improve the understanding of conversations.

It is worth noting that there are indeed approaches that are very close to structured prediction methods. Tanaka et al. (2021) proposed to use the label propagation algorithm to classify thoughtful actions of the dialogue system given an ambiguous user request. In the label propagation algorithm, unlabelled data points are assigned values based on the influence of neighbour labelled data points, hence this algorithm accounts for the value dependencies among data points to

some extent. From another perspective, J. Zhou et al. (2021) developed a dialogue state tracking system that predicts dialogue states with a multi-level fusion mechanism. This mechanism allows the system to study the transition probability between dialogue states, i.e. potential outputs, to perform the end prediction.

Meanwhile, in the Structured Prediction for NLP workshops, there have been a number of works on the name entity recognition task, which can be considered very close to dialogue state tracking (Ma et al., 2022; ter Horst & Cimiano, 2020; A. Gupta & Durrett, 2019; Stratos, 2017). Based on the similarity between tasks, these structured prediction approaches are potentially suitable for dialogue state tracking.

2.4 Summary

This chapter has reviewed three major points of the literature around the dialogue state tracking task.

- Dialogue state representations are crucial to the development of dialogue systems. In the common format, dialogues are represented with dialogue acts, that underlie the interactions with users, and dialogue attributes, that provide further clarifications into user intents. This format of dialogue representations is also called dialogue states, and the task of predicting dialogue states is called dialogue state tracking. In many task-oriented dialogue systems, the dialogue state tracking task is simplified into studying user intents with a set of given dialogue slots and their predefined value sets. In this case,

dialogue states do not represent conversations in a full manner, but directly represent task-specific user intents; that is a big advantage for task-oriented dialogue systems.

- Dialogue state tracking methods have evolved from rule-based systems to complex deep learning techniques. The state-of-the-art dialogue state trackers are mainly based on advanced transformer architectures such as BERT and GPT2. Some systems also incorporate additional techniques for enhance the state prediction performance, but I saw that in general they do not attempt to explicitly model the relationship between dialogue slots.

- Applications of structured prediction methods in dialogue processing are limited despite the fact that there are a wide range of structured prediction approaches applied to various natural language processing tasks. However, the structured approaches applied to dialogue research show that the structural properties of dialogues can be studied and made use of. Hypothetically, the structured prediction methods in NLP, in particular in the name entity recognition task that is very close to dialogue state tracking, are potentially suitable for implementing in dialogue systems.

Although there has been a lot of work on the fields of conversational analysis and formal dialogue modelling, there exists a lack of systematic study specifically of structural properties of conversational language, hence the novelty of exploiting these properties in dialogue state management in this dissertation is ensured.

For the interested reader, more surveys on dialogue state tracking methods can be read in the work by Balaraman et al. (2021), Z. Zhang et al. (2020), H. Chen et al. (2017) and J. D. Williams et al. (2016).

Chapter 3

Inter-Slot Dependencies in Dialogue States

As structured prediction works well when dialogue state inter-slot dependencies exist, my experimental work starts with an investigation of the associations among slot types[1] in dialogue state data. In essence my goal here is to determine whether or not associations exist amongst slot values in dialogue states and to quantify to what extent these dependencies exist. This chapter details the method used to detect the dialogue state slot dependencies, and presents the results on a number of dialogue datasets.

Beside detecting slot dependencies, the dialogue datasets chosen for the experiments of this dissertation and the analysis on them for slot dependencies are also presented in the following sections.

[1]Following the explanation introduced in Section 1.2, I will use the term slot as the shortened term of slot types, and will explicitly use the terms slot types and slot values when needed for clarification.

It is worth clarifying that in this research the slot dependencies are calculated between slots based on their values. In detail, the occurrences of slot value pairs are observed in the dataset and used in the analysis formulas. It is also worth noting that the research is conducted only among slots that are significant in dialogue states, i.e. highly frequent observed slots. At a high level, the dependencies between dialogue slots represent the ontological structure of topics discussed, with dialogue slots being defined in the ontology of specific domains.

This chapter begins by presenting the dialogue corpora chosen for study in Section 3.1. Then it presents the statistical testing method to be used in Section 3.2. The studied results of inter-slot dependencies are demonstrated in Section 3.3. Finally, the chapter is summarised in Section 3.5.

The work presented in this paper was largely covered in submissions to the 23rd Workshop on the Semantics and Pragmatics of Dialogue (SemDial) (Trinh et al., 2019c) and the SIGDial 2019 Conference (Trinh et al., 2019a).

3.1 Dialogue Datasets

Due to the challenges in collecting dialogue data there is a limited number of dialogue corpora available for public use. In this research I use a number of dialogue datasets of both single domain and multiple domain settings that are known for their usefulness in the dialogue state tracking task. In these data I focus on the main task, *Joint goals*, and investigate the pairwise dependencies across the slots.

The single domain dialogue corpora used include two datasets of spoken dialogues and one dataset of chat-based dialogues with a similar ontology. The two spoken dialogue datasets are DSTC2 & 3.

- DSTC2 (Henderson, Thomson, & Williams, 2014a) is a restaurant information dataset collected from spoken conversations. It consists of 3235 dialogues in total that are split into three subsets: 1612 dialogues for training, 506 dialogues for validation, and 1117 dialogues for testing. The *Joint goals* task initially consists of four slots: *food*, *price range*, *area*, and *name*. However, following common practice, in my work the slot *name* is omitted due to the lack of its appearance.

- DSTC3 (Henderson, Thomson, & Williams, 2014b) is a spoken dialogue dataset in the tourism information domain with 2275 dialogues in a complete set. Similar to the DSTC2 dataset, I solve the *Joint goals* task of only four informable slots, *food*, *price range*, *area*, and *type*, as I omit other slots due to their extremely low appearance frequency in the data.

The appearance analysis of informable slots in DSTC2 & 3 data is presented in Table 3.1. Omitting low frequency slots is a common practice in the dialogue state tracking research, whereas it is still possible to compare different systems performing predictions on DSTC DSTC2 & 3 data. The omission of slots happens both in the training and the testing phases.

The other single domain dialogue corpus considered is Wen et al. (2017)'s WOZ dataset that was collected from 1200 chat-based dialogues in the restaurant

Table 3.1: The analysis of informable slot appearance (%) in DSTC 2 & 3, calculated over the number of dialogues and turns in the whole dataset.

Slot	DSTC2		DSTC3	
	call	turn	call	turn
food	87.9	79.3	63.5	55.4
price range	73.5	62.6	68.3	60.8
area	81.8	72.3	59.5	50.6
type	-	-	98.5	91.0
name	0.8	0.5	1.5	0.6
near	-	-	8.5	6.8
has tv	-	-	7.3	5.8
has internet	-	-	7.6	5.9
children allowed	-	-	4.9	3.6

information domain. It is split into 600 dialogues for training, 200 dialogues for validation, and 400 test dialogues. The WOZ *Joint goals* task, similar to DSTC2, also consists of three informable slots: *food*, *price range*, and *area*, and these slots have the same value set as the corresponding slots in the DSTC 2 & 3 datasets.

The multiple domain dialogue corpora used in my research are the MultiWOZ 2.0 (Budzianowski et al., 2018) and MultiWOZ 2.1 (Eric et al., 2020) datasets. These two chat-based dialogue datasets are identical to each other in term of dialogue data, but MultiWOZ 2.1 has cleaner label annotations than MultiWOZ 2.0 sets due to human efforts. Overall, these corpora include over 10,000 dialogues that consist of more than 100,000 turns across seven domains. However, following the common practice of handling the data in these datasets, two domains are omitted: *hospital* and *police*, as their appearances are extremely low; and, this reduces the number of slots available for tracking dialogue states to 30 in total.

The summary of dialogue corpora chosen for this work is presented in Table 3.2.

Table 3.2: Overview information of the chosen dialogue corpora for this research.

Corpus	#domains	#dialogues	#turns	#slots	#values
DSTC2	1	3235	25501	3	105
DSTC3	1	2275	18824	4	58
WOZ	1	1200	5012	3	105
MultiWOZ	5	10438	143048	30	4510

Here the reported multiple domain dataset is MultiWOZ 2.1 (Eric et al., 2020).

3.2 Analysis Methods

Since dialogue state tracking can essentially be thought of as categorical data classification, Pearson's chi-square method, which is popular for investigating bivariate statistics, was chosen to investigate associations across dialogue slots in training data. Specifically statistical tests are performed to detect dependencies between dialogue slots in a pairwise fashion. Following the confirmation that pairwise slot dependencies exist in dialogue state data, the strength of these associations is measured using various chi-square-based techniques. The analysis begins with a discussion of the chi-squared test.

3.2.1 Pearson's Chi-Square Test

Pearson's chi-square test is a significance test to detect bivariate association between variables. In order to apply it to the dialogue data, I first create contingency tables for all pairs of dialogue slots with their values presented in the corpora. Given this contingency table for two slots A and B, let $P(A_i)$ and $P(B_j)$ be the

probability of appearance in the population of the slot values A_i and B_j. The probability of appearance is calculated with the formula:

$$P(A_i) = \frac{O_i}{N}; P(B_j) = \frac{O_j}{N} \tag{3.1}$$

where O_i and O_j are the observed appearances of the slot values A_i and B_j, and N is the population size, i.e. the number of turns.

The dependency between these two slots is tested with the algorithm presented as follow:

Step 1. Hypotheses for the task are first defined.

H_0: The two slots are independent

$$P(A_i \cap B_j) = P(A_i)P(B_j) \tag{3.2}$$

H_1: The two slots are dependent

$$P(A_i \cap B_j) \neq P(A_i)P(B_j) \tag{3.3}$$

Step 2. The expected frequency of $\{A_i, B_j\}$ is calculated based on the input.

$$E_{ij} = P(A_i) * P(B_j) * N \tag{3.4}$$

where N is the population size.

Step 3. The chi-square error is then computed as follows:

$$\mathcal{X}^2_{\mathcal{V}} = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3.5}$$

where $\mathcal{V}$ is degree of freedom, and O_{ij} and E_{ij} are the observed and expected frequencies of slot values A_i and B_j.

Step 4. The hypothesis H_0 is rejected if the computed test statistics $\mathcal{X}^2_{\mathcal{V}}$ is high and the significance coefficient $p < 0.05$.

This statistical test is performed on all the dialogue corpora that were introduced in Section 3.1.

3.2.2 Measuring Slot Dependencies

In general, Pearson's chi-square statistical test presented in Section 3.2.1 can only detect the existence of the dependencies between dialogue slot types. Practically these dependencies might vary between different slot pairs. Therefore, following the confirmation of dialogue slots' association existence, it is useful to measure the strength of these dependencies using other techniques. There are several measurements of association strength directly related to the chi-square statistics (Field, 2017). These measures are normally scaled between 0 and 1 indicating the range from no relationship to a perfect association among slots.

Among all the measurement methods, there are three popular coefficients that use the chi-square statistics $\mathcal{X}^2$ from Equation 3.5 to measure the association

strength:

- The ϕ coefficient is calculated by adjusting the chi-square statistic by the population size:

$$\phi = \sqrt{\frac{\mathcal{X}^2}{N}} \tag{3.6}$$

  where $\mathcal{X}^2$ is the chi-square statistic value, and N is the number of samples in the dataset.

- The contingency coefficient C is slightly different from the ϕ coefficient, although it is also computed based on the adjustment of the chi-square statistics by the population size:

$$C = \sqrt{\frac{\mathcal{X}^2}{N + \mathcal{X}^2}} \tag{3.7}$$

- Cramer's V coefficient is another chi-square-based measure of association, different from the two above, it includes information on the contingency table's dimensions:

$$V = \sqrt{\frac{\mathcal{X}^2}{N \min(r-1, c-1)}} \tag{3.8}$$

  where r and c are the number of rows and columns in the contingency table.

These measures provide more details on dependencies that I wish to investigate among dialogue slots. The inter-slot associations may vary according to the measured result. Therefore it is important to correctly interpret the strength of inter-slot dependencies. For example, Cramer's V values vary between 0 and 1,

Table 3.3: Interpretation of Cramer's V coefficient (Field, 2017).

Cramer's V	Type of association
$V > 0.5$	Very strong
$V > 0.3$	Strong
$V > 0.1$	Moderate
$V \geq 0$	No or weak

but any value larger than 0.3 indicates a strong relationship between the pair of slots (Field, 2017). An interpretation chart is presented in Table 3.3, and will be used further to analyse the detected slot dependencies in the dialogue data.

3.3 Dialogue Slot Dependencies Analysis

In this section, the Pearson's chi-square statistical test is performed to investigate bivariate dependencies of dialogue slots in the dialogue corpora presented in Section 3.1. The strength of these dependencies are measured with chi-square-derived methods and also presented.

3.3.1 Single Dialogue Domain

For single dialogue domains cases, the analysis was conducted on 3 datasets: DSTC2 (Henderson, Thomson, & Williams, 2014a), DSTC3 (Henderson, Thomson, & Williams, 2014b), and WOZ (Wen et al., 2017).

Firstly, the statistics of DSTC2 data are reported in Table 3.4. In DSTC2 data as outlined above, the focus is on the *Joint goals* task that consists of three slots (*food*, *price range*, and *area*), hence the investigation of the relationships is

Table 3.4: Statistical assessment of slot dependencies in the DSTC2 data.

DSTC2		food - price	food - area	price - area
Chi-square	$\mathcal{X}^2$	9430.5	12739.0	3937.9
	$\mathcal{V}$	176	180	24
	p	$<$ 2.2e-16	$<$ 2.2e-16	$<$ 2.2e-16
Coefficients	ϕ	0.608	0.707	0.393
	C	0.520	0.577	0.366
	V	0.272	0.267	0.176

Table 3.5: Statistical assessment of slot dependencies in the WOZ and DSTC3 data in the Cramer's V coefficient.

WOZ				DSTC3				
	food	price	area		food	price	area	type
food	-			food	-			
price	0.316	-		price	0.248	-		
area	0.302	0.180	-	area	0.163	0.232	-	
				type	0.300	0.195	0.220	-

conducted among these three slot pairs.

In the results it is observed that all statistical significance values show $p < 0.05$, that confirms the existence of slot dependencies within the DSTC2 dialogue data. Furthermore the measured strength of these dependencies indicate that they are moderate associations (as interpreted using Table 3.3).

Following the work done on the DSTC2 dataset, the study of pairwise dependencies in the DSTC3 and WOZ data was conducted. In these datasets the slots are confirmed to be dependent on each other pairwise, that is shown by all statistical significance values having $p < 0.05$. Therefore the association strengths are measured, and Table 3.5 only reports the measured result to avoid repetition. The association strength result is reported with Cramer's V coefficient.

In the findings, these dependencies are consistently strong ($V > 0.3$) and moderate ($V > 0.1$) as interpreted using Table 3.3. It can be concluded that all single domain datasets in my research contain strong and moderate slot dependencies.

3.3.2 Multiple Dialogue Domain

Multiple dialogue domains are investigated using two chat-based dialogue datasets, MultiWOZ 2.0 (Budzianowski et al., 2018) and MultiWOZ 2.1 (Eric et al., 2020). Here, the MultiWOZ 2.1 dataset contains better label annotation thanks to human efforts. Therefore, it is considered to be a better source for the investigation of slot dependencies in dialogue states. Similar to the single domain data, the statistical test is performed for all the dialogue slots in a pairwise fashion across the domains. In the results, all statistical significance values are found to be significant with $p < 0.05$. Further, the dependence strengths are measured with the Cramer's V coefficient. As this dataset contains 30 slot types, the assessment of slot dependencies in Cramer's V values is presented in a heat map format (Figure 3.1).

It is observed that most dialogue slots are dependent on each other to various extents. The dialogue slot relationships range from a weak dependency such as *hotel.name – restaurant.name* to a very strong association such as *hotel.area – hotel.type*. It is also observed that there are "red lines" representing equal association of particular slots such as *hotel.internet* and *hotel.parking* to all other slots. This is explained by the nature of these slots being boolean values, therefore they are

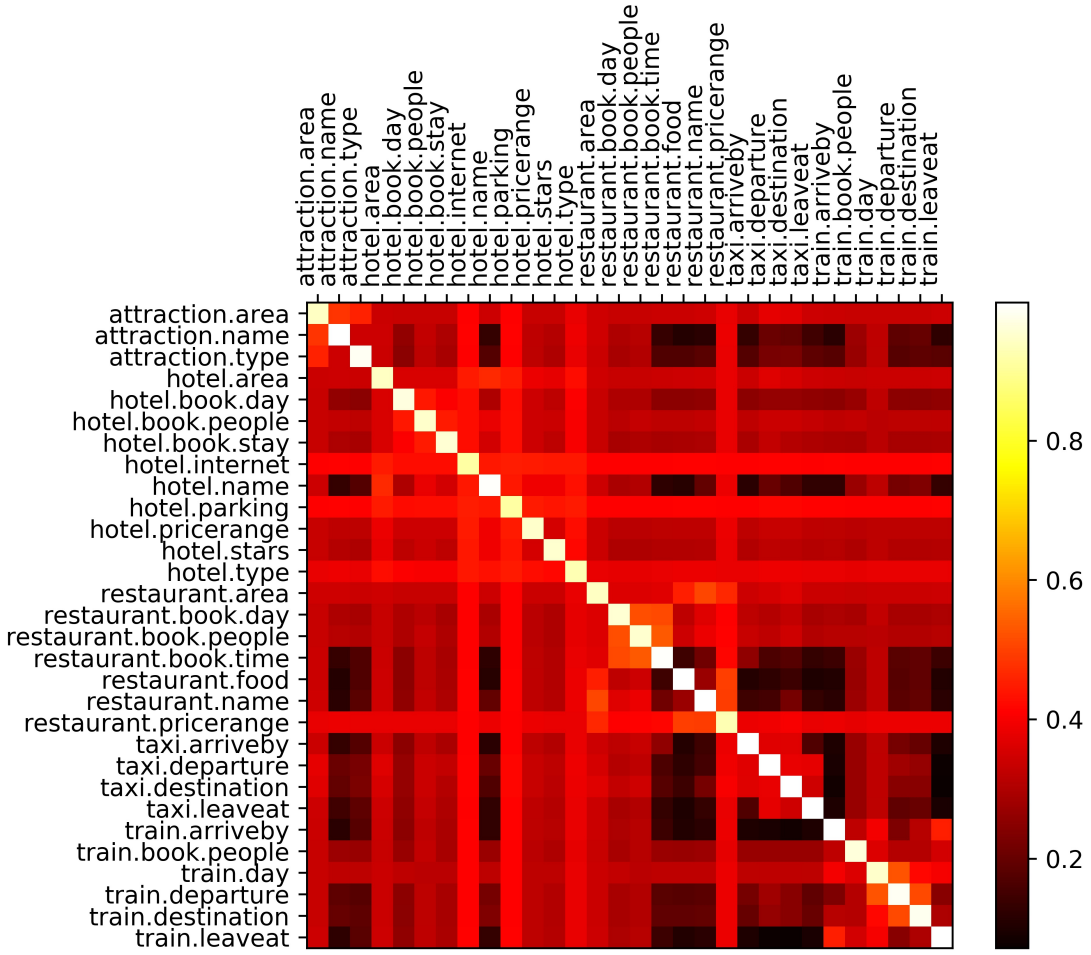

Figure 3.1: Cramer's V assessment of slot type dependencies in MultiWOZ 2.1 data

equally associated to all other slots across domains.

In this observation, the dependencies between dialogue slots are not limited within their own domains. Such cross-domain dependencies provide a strong foundation for my research in general.

3.4 Discussion

In the previous section, the existence of dialogue inter-slot dependencies has been confirmed with statistical tests in task-oriented dialogue domains, where all statistical significance values show $p < 0.05$. At a high level, the dependencies between dialogue slots represent the ontological structure of the task. It is worth clarifying that the idea of tasks, slots, and frames, and the codependencies between them in dialogue management indicate the conceptual structure rather than the linguistic structure. Dialogue management makes use of this phenomenon in the way that they are treated as extra information for more accurate dialogue understanding.

Here, the dialogue inter-slot dependencies provide a number of useful insights. It is proven that in task-oriented dialogue domains, the interrelations of dialogue slots indicate that they interact with each other to some extent. Therefore, a dialogue manager should pick dialogue slots not in isolation, but in a collective way. That leads to the question of whether the slots should be filled in a collective manner to take advantage of this information. On the other hand, the measured results of slot relationships show that strong dependencies are not universal even for a small corpus. In detail, the dependencies vary from weak to strong and that indicates how much impact the information of one slot has on another slot. Combining these phenomena, dialogue state management can be considered a complex task with inter-slot dependencies.

In multiple domain corpora, dialogue inter-slot dependencies were observed not only among slots within a domain, but among slots across domains as well. For

example, Figure 3.1 demonstrated various relationships among slots of *restaurant* and *hotel* domains. These dependencies reflect users' intents in reality, such that users often require different pieces of information across domains in their queries rather than single out a particular domain. Here, the conceptual structure of the dialogue state tracking task goes beyond single domain dependence, hence dialogue management making use of it can perform at a higher level.

Overall, studying various dependencies between dialogue slots in a number of different domains is beneficial in many ways. Since slot dependencies serve as extra features for dialogue management, it is natural to incorporate them into the dialogue state tracking process to improve the accuracy of slot prediction. Furthermore, widening the concept of dialogue variables to a broader sense such as multimodalities or users' preferences and personalities is beneficial in the long term.

3.5 Summary

This chapter presented two contributions of note:

- The chosen task-oriented dialogue corpora for this research were presented, those included three single domain datasets and two multiple domain datasets. The coverage of both settings enables the diversity in my work and ensures the generalisability of my methodology.

- Pearson's chi-square statistical tests were performed on the dialogue data to

detect slot dependencies presented in dialogue states, followed by measuring the strength of these dependencies.

Overall, it is found that in all dialogue datasets, the dialogue slots are dependent on each other pairwise. In single domain corpora these dependencies are consistently strong. While in multiple domain corpora they vary from a weak association to a very strong relationship. Furthermore, the slot dependencies were observed across dialogue domains. These findings provide a strong motivation for my further research.

Investigation of dialogue slot dependencies in various dialogue domains is an important piece of work in this dissertation, in that the existence of these dependencies back up my hypothesis of structural properties in dialogues. Such knowledge motivates further research on integrating slot dependencies into dialogue processing.

Chapter 4

Harnessing Domain Structure with Multi-Task Learning

Since dialogue states consist of a number of components, of which each can potentially be viewed as an individual tracking task, the dialogue state tracking problem has most frequently been solved by developing a separate model for a single subtask. However, this approach already presumes the independence between all the components in dialogue states. But as I argued earlier and showed in the previous chapter, there exist relationships between dialogue slots, thus there may be an advantage to analysing the whole dialogue state tracking task as a multi-task problem. I therefore propose to incorporate these relationships in the dialogue state tracking processing with a multi-task learning method (Caruana, 1997).

Here I emphasise the fact that the multi-task learning method focuses on the relationships between slots during the training process by sharing the training

signals. However it should be noted that this type of dependencies is at early stage among latent variables, not explicitly for slots and their values as detected in Chapter 3. The study is conducted on DSTC2 data (Henderson, Thomson, & Williams, 2014a), of which the dialogue states consist of several components, to investigate the efficiency of shared training among these tasks.

This chapter is structured as follow: firstly, an overview of the multi-task learning method is provided in Section 4.1. Then the design and development of the multi-task dialogue state tracker are presented in Section 4.2, followed by the experimental results and analysis in Section 4.3. The chapter is concluded with a brief summary of the studied method in Section 4.4.

This work has been published at the 21st and the 22nd Workshops on the Semantics and Pragmatics of Dialogue (SemDial) (Trinh et al., 2017, 2018).

4.1 Overview of Multi-Task Learning

In recent years multi-task learning methods have been studied intensively in natural language processing (Worsham & Kalita, 2020; S. Chen et al., 2021). The multi-task learning techniques are applied to a wide range of computational linguistics tasks such as text classification (P. Liu et al., 2017), semantic parsing (Peng et al., 2017), and sequence labelling (Rei, 2017). Subsequently, many works in the dialogue field also make use of the multi-task learning approaches. The interpretation of multi-tasks in dialogue systems ranges from the functions of different dialogue system components (T. Zhao & Eskenazi, 2016) to the subtasks

within a dialogue system component (Q. Chen et al., 2019).

The core concept of multi-task learning methods is that they aim to improve the performance of the system on several related tasks by making use of the shared information in the training phase (Caruana, 1997). The training signals are shared between multiple tasks via a simultaneous optimising process of the metrics applied on them. This process makes the multi-task learning approach different from the transfer learning method, where the knowledge of tasks is learned in a sequential manner. In transfer learning, a model is often pretrained with an auxiliary task, then applied to training the main task with the purpose to boost the performance on this task (Ruder, 2019).

To illustrate, let us consider an example of a multi-task learning problem: given two tasks A and B that are to be learned, and the loss functions for these tasks are L_A and L_B respectively. Among the trainable parameters θ, I denote that W are shared weights between the two tasks, while W_A are weights for task A, and W_B are weights for task B. The common practice to formulate the objective function in multi-task learning is summing the two losses:

$$L = \alpha L_A + (1 - \alpha) L_B \tag{4.1}$$

where α is the loss coefficient that implies the importance of the task in the learning process.

The gradients on the objective functions are calculated according to the com-

ponental losses:

$$\nabla_\theta L = \alpha \frac{dL_A}{d\theta} + (1-\alpha)\frac{dL_B}{d\theta} \tag{4.2}$$

where $\theta = \{W, W_A, W_B\}$ are the total set of trainable parameters.

In the backpropagation process, the trainable parameters are updated as such:

$$\begin{aligned} W &= W - \lambda\left(\alpha \frac{dL_A}{dW} + (1-\alpha)\frac{dL_B}{dW}\right) \\ W_A &= W_A - \lambda\alpha \frac{dL_A}{dW_A} \\ W_B &= W_B - \lambda(1-\alpha)\frac{dL_B}{dW_B} \end{aligned} \tag{4.3}$$

where λ is the learning rate.

Here it is observed that the shared weights are updated based on the training signals of both tasks, while the task-based learned weights are updated according to the single task to which they are related.

In this study, dialogue states of DSTC2 include three subtasks: *Joint goals*, *Search method*, and *Requested slots*. The multi-task dialogue state tracker is designed based on these subtasks. Among them, the *Joint goals* task is the most challenging problem, that requires the system to classify values for four informable slots at the same time. Three of these four informable slots were studied for the dialogue slot dependencies between them in the previous chapter.

4.2 Multi-Task Dialogue State Tracker

DSTC2 dialogues contain multiple turns, each of those includes machine acts and user utterance. In this work, the dialogues are treated as sequences of words and tokens, and a Recurrent Neural Network (RNN) architecture is used to handle sequential data. Different components for the multi-task dialogue state tracking system are also developed to process different dialogue input entities. Overall, the tracker includes one input layer, one output layer, and two hidden recurrent neural layers that consist of multiple recurrent neural cells. In this architecture, all recurrent neural cells are of Long Short-Term Memory (LSTM) type (Hochreiter & Schmidhuber, 1997). RNN, and in particular LSTM, was chosen due to their good performance on sequential data such as natural language and conversations.

Prior to developing the tracking system itself, it is important to preprocess dialogue input data. As machine acts are provided in semantic format, they are parsed with similar techniques proposed by Henderson, Thomson, & Young (2014b), and an autoencoder is pretrained to reduce the dimensionality of the machine act representation vector (Figure 4.1). As the dialogue acts in DSTC2 domain were provided in the format *acttype(slot=value)*, Henderson, Thomson, & Young (2014b) proposed a n-gram type feature extraction technique to establish a list of features: *acttype*, *slot*, *value*, *acttype-slot*, *slot-value* and *acttype-slot-value*. This technique resulted in high dimensionality vector representations for machine dialogue acts. Therefore, the reduction of the machine act dimensionality is important as it makes the machine act have the same dimensions as the embedding

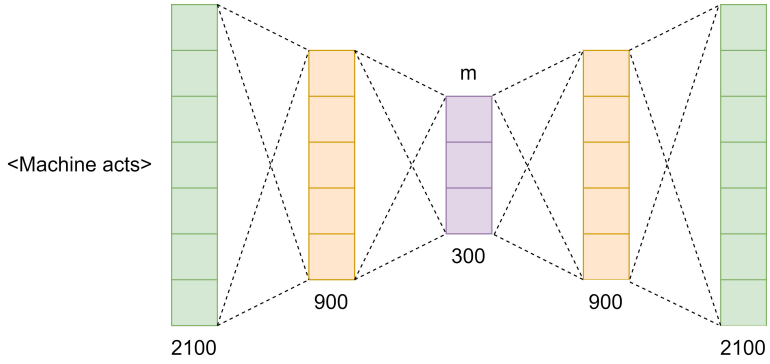

Figure 4.1: Machine act autoencoder.

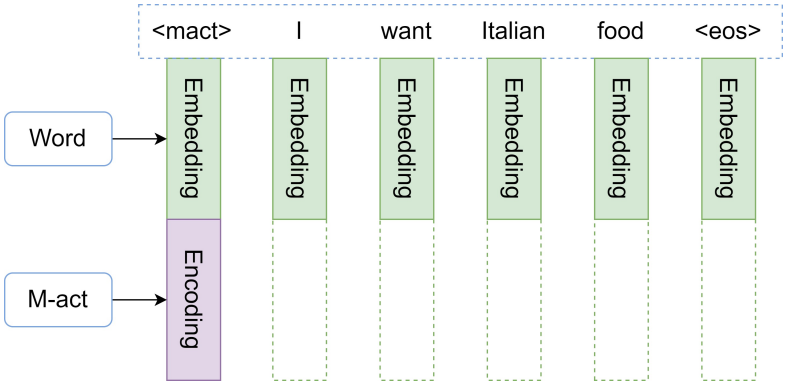

Figure 4.2: Sequential dialogue input.

for a token in a user utterance.

The machine act tokens are then added into the dialogue sequence of user utterance words (Figure 4.2). User utterances are provided in the format of speech transcriptions, therefore they do not require special preprocessing techniques. The vocabulary is defined using full words without punctuation. Here two special tokens are used: <*mact*> to mark the beginning of a dialogue turn and the position where the embedding for the turn's machine act is inserted into the sequence by concatenating it with the <*mact*> token embedding, and <*eos*> to mark the end of a dialogue turn where dialogue states are produced. All the words and tokens except <*mact*> are embedded with an online-trained embedding layer.

Furthermore to develop multi-task learning systems I use a two-layered unidirectional LSTM structure that is formulated with timesteps to roll over dialogue sequences (Figure 4.3). In this section and following chapters, all the deep learning architectures such as LSTM and energy-based models are implemented using Python 3.7 and Tensorflow 1.14 (Abadi et al., 2015).

In this research, I develop two systems to propose my multi-task learning method:

- I develop a multi-task baseline model, namely model *a*, that contains only task-specific LSTM cells and classifiers.

- On the other hand, in order to make use of shared training signals across the tasks, a concatenation layer is added on top of the first hidden LSTM layer (model *b*). The concatenation layer can leverage the signals going through it, thus making all LSTM cells of the first hidden layer equal. Meanwhile the LSTM cells of the second hidden layer still stick to the specific tasks. This model is considered a true multi-task learning approach.

The training process, therefore, is executed through two different mechanisms across the true multi-task and baseline models. At each time step, dialogue input is transformed into a vector representation and fed into the networks. Here, the learning mechanisms are differentiated as such:

- The baseline model, model *a*, with the simpler mechanism uses only task-specific LSTM cells to process input, and task-specific classifiers to perform

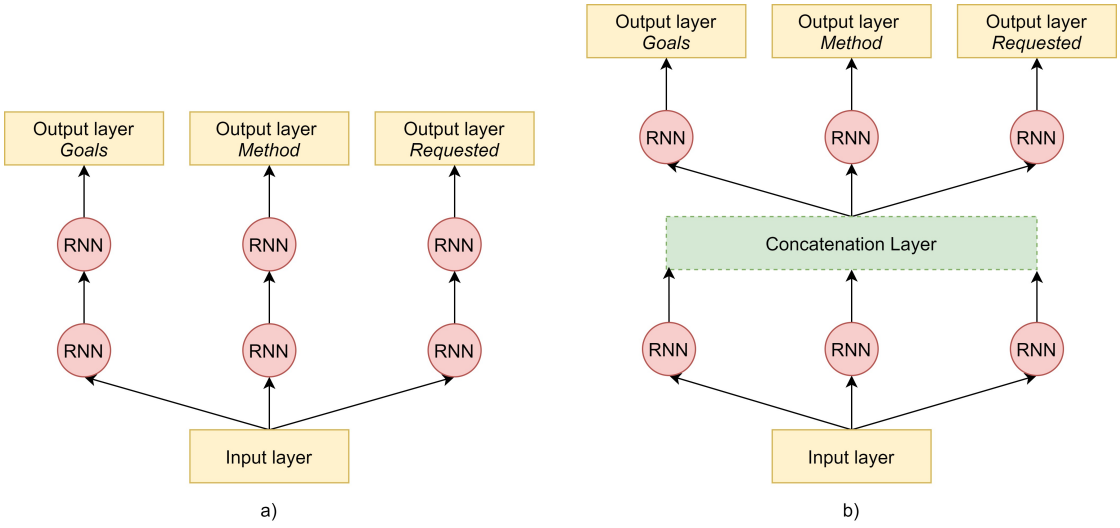

Figure 4.3: The multi-task baseline model (a) and the proposed multi-task dialogue state tracker (b).

the prediction. All the task-specific parameters can be trained either separately or in the multi-task learning fashion, i.e. using the loss calculation and weight update as set out in Equations 4.1, 4.2 and 4.3. In this work they are trained in the multi-task learning fashion, thus making model a a baseline for the multi-task learning approach.

- Meanwhile the true multi-task system, model b, processes the dialogue input with a more complex architecture. At the first layer, all LSTM cells process the input vector and produce multiple hidden states. Then these hidden states are concatenated into a joint vector representation, that is hypothesised as the representation of the whole dialogue state until the current time. Following that, the dialogue state representation is fed into the task-specific LSTM layers and classifiers to perform predictions.

The hyper-parameters were carefully selected with a grid search method during

the development phase. The details of the hyper-parameters of the multi-task models are presented as *Feature network* in Appendix A.

To train my models in the multi-task learning manner, a *joint loss function* of all the subtasks is calculated and used to backpropagate through the whole network (as presented in Equation 4.1). During the backpropagation process, model b updates the parameters of the task-specific LSTM cells based on the loss of the appropriate subtasks, and the parameters of the shared LSTM cells based on the contribution of all these errors (as shown in Equation 4.3). Meanwhile, model a updates the LSTM parameters only based on their related individual subtask.

4.3 Results & Analysis

Within the DSTC2 dataset, each turn defines full dialogue states as a combination of three subtasks, *Joint goals*, *Search methods*, and *Requested slots*. The performance of my multi-task learning models on the DSTC2 testset across all three tasks is presented and benchmarked against the state-of-the-art systems with the same setting in Table 4.1. The results are evaluated with the DSTC2 feature accuracy metric.

The results on the DSTC2 testset demonstrates that the proposed multi-task learning-based models achieved competitive performance to other related state-of-the-art sequence-to-sequence dialogue state trackers at the time of development: EncDec Framework (Platek et al., 2016), LecTrack (Zilka & Jurcicek, 2015b), and

Table 4.1: Performance of the proposed multi-task models and related state-of-the-art systems on DSTC2 testset evaluated with the accuracy metric.

Model	Joint Goals	Requested Slots	Search Method
Sequence-to-sequence model (Feng et al., 2021)	**0.850**	-	-
Word-based RNN model (Henderson, Thomson, & Young, 2014b)	0.768	0.978	0.940
EncDec Framework (Platek et al., 2016)	0.730	-	-
LecTrack (Zilka & Jurcicek, 2015b)	0.72	0.97	0.93
CNET Tracker (Jagfeld & Vu, 2017)	0.714	0.972	-
This work			
MTL Model *b*	0.728	**0.980**	**0.946**
MTL Model *a*	0.720	0.978	0.944
DSTC baseline (Henderson, Thomson, & Williams, 2014a)	0.719	0.879	0.867

CNET Tracker (Jagfeld & Vu, 2017). In detail, the true multi-task model *b* yields the best results in two subtasks, *Requested slots* and *Search method*, while performing relatively well on the *Joint goals* subtask. It is important to note that my trackers are capable of predicting full dialogue states with comparable performance with the best tracker with the same input processing technique, EncDec Framework (Platek et al., 2016), that is capable of tracking only *Joint goals*. Here the difference in the *Joint goals* result between this work and EncDec Framework is as small as 0.2%.

When comparing the two multi-task learning-based models in this work, it is observed that model *b* generally outperforms model *a* in all tasks. The key factor of this result is held within the shared LSTM layer of model *b*. As this true multi-task model is structured in such a way that the dialogue information is extracted and

shared across all tasks at an early stage, the control over correlation between slots is enhanced. Meanwhile, a certain level of independence in performing predictions of the subtasks is still ensured by using a task-specific LSTM layer and classifiers.

This hypothesis is proved by the performance of the multi-task model *b*, and the EncDec Framework (Platek et al., 2016), that also accounts for shared signals between informable slots in *Joint goals*. Together they outperform other models that are either trained in a multi-task learning fashion with little influence of the tasks on each other such as the baseline model *a*, or developed as a set of combined separate trackers such as LecTrack (Zilka & Jurcicek, 2015b) and CNET tracker (Jagfeld & Vu, 2017).

Furthermore, the hypothesis is proved by other works that process dialogues on a turn-based basis that compare models with shared and non-shared parameters between slots such as StateNet (L. Ren et al., 2018).

Although the parsing technique for dialogue acts from word-based RNN model (Henderson, Thomson, & Young, 2014b) is adopted to preprocess machine acts in my approach, the word-based model processes user input with word features extracted directly from user utterances and performs dialogue state predictions on a turn-based manner, that is completely different from my modelling setting. The word-based RNN model also does not account for the relationships among dialogue slots. Therefore, it is not relevant to us here.

Table 4.2: Detailed performance of the proposed multi-task models on DSTC2 informable slots.

	Food	Price	Area	Name
Turns	9890			
Model *b*	0.848	0.893	0.920	0.995
Model *a*	0.847	0.881	0.919	0.995
Turns with change	1596	932	1046	9
Model *b*	0.786	0.804	0.870	0.000
Model *a*	0.780	0.767	0.856	0.000

4.3.1 Analysis of Slot-Based Performance

To further investigate the effectiveness of the multi-task learning method, a detailed performance analysis of the proposed multi-task models is conducted on the *Joint goals* task, which is the most challenging subtask of DSTC2. The analysis result is presented in Table 4.2.

According to Henderson, Thomson, & Williams (2014a), user intents of slot *food* change most frequently, up to 40.9% dialogues in the testset, and it is the most difficult slot to track. The analysis was conducted on the dialogue level. However, user intents are expected to change also on a turn-by-turn basis. My analysis shows that the DSTC2 testset includes 9890 turns in total, in which there are 1596 (16.14%) turns where users change the *food*, 932 (9.42%) turns where the *price range* value is changed, 1046 (10.58%) turns with the change in *area*, and only 9 (0.09%) turns with regard of slot *name*.

The result demonstrates that the true multi-task model *b* consistently outperforms the baseline model *a* in both cases, tracking overall slot-based results and tracking value changes of each slot. This observation is well presented in the

tracking results of three out of four informable slots, *food*, *price range* and *area*, and overfit in tracking slot *name*. It can be understood in that users rarely mention the name of restaurants, thus creating the lack of training data. However, the omission of slot *name* in *Joint goals* does not affect the overall performance of dialogue state trackers as shown in the case of EncDec Framework (Platek et al., 2016). It is a common practice to omit unimportant slots, whose appearance frequency is very low in dialogue data such as the DSTC2 slot *name* in the *Joint goals* task, as mentioned in Chapter 3. Therefore, in my further work in the following chapters, the same practice is followed to reduce the resource requirements while not sacrificing the performance of my models.

4.4 Summary

This chapter demonstrated that multi-task learning is an appropriate approach for dialogue state tracking tasks where associations between dialogue state components (subtasks) are taken into account. The result suggests that the proposed multi-task model achieve state-of-the-art results. The novelty of this work lies in the proposed architecture as such: on the one hand, it accounts for the relationships between dialogue slots and dialogue state subtasks with a shared LSTM layer at an early state; on the other hand, it ensures a certain level of independence between these entities with a task-specific LSTM layer and classifiers.

In this work the dependencies between dialogue state slots are, however, not accounted for explicitly as presented in Chapter 3. This is the limitation of the

multi-task learning methodology. However, the multi-task learning approach is considered the baseline for performing dialogue state tracking with slot dependencies, and further studies are conducted with other methods for capturing explicit slot dependencies in dialogue states in the following chapters.

Chapter 5

Studying Slot Dependencies with Energy-Based Learning

In the previous chapter I demonstrated that approaches that consider the dependencies between dialogue state subtasks such as multi-task learning outperform their counterparts that ignore these features. However, it is arguable that the correlations that exist among the slots that the study conducted in Chapter 3 revealed, are not yet included in the multi-task learning approach. The exclusion of slot dependencies in dialogue states means potential improvements are overlooked. Accounting for these slot dependencies in the prediction process casts the dialogue state tracking task into a structured prediction problem. Hence it aligns with my research objectives of investigating the structural properties of conversation in dialogue states.

Structured prediction approaches have been successfully applied to various NLP

tasks, where the output labels are not assumed to be independent of each other (Tu, Pang, & Gimpel, 2020; Tu, Liu, & Gimpel, 2020; Tu & Gimpel, 2019, 2018). This is similar to the case of dialogue states where slot values influence each other. Thus I believe it is a strong motivation to interpret the dialogue state tracking task as a structured prediction problem. For the interested reader, more details of structured prediction methods in NLP can be read in the work by Dev et al. (2021).

In this research presented here, I apply the energy-based learning approach (LeCun et al., 2006) to solve the structured dialogue state tracking task, since energy-based learning is notably effective for capturing slot dependencies, and performing structured predictions (Osogami, 2017a,b). There have been published research where energy-based approaches are successfully applied to solve other NLP problems such as and part-of-speech tagging and named entity recognition (Tu, Liu, & Gimpel, 2020; Tu, Pang, & Gimpel, 2020). My energy-based dialogue state tracker is developed with deep learning modelling based on Structured Prediction Energy Networks (SPEN) (Belanger & McCallum, 2016; Belanger et al., 2017) and Deep Value Networks (DVN) (Gygli et al., 2017). In practice, applying the energy-based learning methodology to dialogue processing, in particular dialogue state tracking, is a novel modelling approach.

In the work below, the energy-based learning method is applied to the dialogue state tracking task of two single domain dialogue datasets, DSTC2 & 3 (Henderson, Thomson, & Williams, 2014a,b). In these dialogues, the focus is only for tracking

Joint goals that consist of informable dialogue slots with highly frequent appearance in training data (Table 3.1). The slots with low frequency of appearance are not significant for the study and dialogue state tracking performance overall, therefore they are omitted in my experiments following the common practice in the community.

The structure of this chapter begins with an overview of the energy-based learning methodology in Section 5.1. Then the design and development of my energy-based dialogue state tracker is presented in Section 5.2, followed by the detail of energy-based modelling strategies in Section 5.3. Section 5.4 demonstrates the experimental results and analysis. The chapter is concluded with a brief summary in Section 5.5.

The work in this chapter has been published at the 1st Workshop on NLP for Conversational AI (Trinh et al., 2019b) and the SIGDial 2019 Conference (Trinh et al., 2019a).

5.1 Overview of Energy-Based Learning

The core mechanism of energy-based methods is to measure the goodness of fit between a structured output Y and an input X using a so-called energy function $E(\cdot)$. Due to various formats of raw inputs in practice, the input is often preprocessed in a domain appropriate way to achieve more useful feature representations referred as a feature function $F(X)$. From here, the energy function $E(F(X), Y)$ measures relationships between the input features, instead of the raw input, and

the structured output. This energy function returns a scalar value, that is called *energy*, that captures the relationship measurement. Training a good energy function is crucially important for energy-based learning methods.

This work approaches the energy-based learning method from a deep learning perspective. The two components of my energy-based model are summarised as follows:

- **Feature function** $F(X)$ can be implemented with deep structures to generate feature representations. Commonly a CNN is used for image processing, and an RNN is used for language processing. I thus refer to the deep learning-based feature function as a feature network. This network can be either pretrained or online-trained with the whole model.

- **Energy function** $E(F(X), Y)$ when developed with a neural architecture should be differentiable. The benefit of this deep learning structure is that it can be trained with popular techniques such as gradient descent (Belanger & McCallum, 2016). The neural energy function is thus referred to as an energy network.

The implementation of the feature network $F(X)$ and the energy network $E(F(X), Y)$ in my dialogue state tracking system will be explained in Section 5.2.

In the general case, the working mechanism of an energy-based model is split into learning and inference processes that have different roles (LeCun et al., 2006):

- **Learning process** is the phase in which the network is trained such that it produces minimal energy values for valid input and output configurations,

and higher energy for other invalid configurations. Let us review an example to clarify this learning process. I denote the ground truth energy $E^* = E(F(X), Y^*)$ for an input X and a target Y^*, and a predicted energy $E = E(F(X), Y)$ for the same input X and any output Y. For the trained energy network, the desired result should be $E^* \leq E$, where the equality occurs only in case $Y = Y^*$. The main challenge here is to define an objective function to guide the learning process, for instance a loss function $L(E, E^*)$ between the predicted energy E and the ground truth energy E^* (LeCun & Huang, 2005).

- **Inference process** is the phase in which the network produces structured predictions based on the trained energy network $E(F(X), Y)$, of which the input X is known while predicted output Y is not yet known. Thus at runtime, the process begins with an initial hypothesis $Y^{(0)}$, that is usually a random hypothesis, and then performs an inference loop to update the output Y so as to find the best fitting $Y \rightarrow Y^*$ according to my learned differentiable energy function (i.e., the Y for which $E(F(X), Y)$ returns the lower energy).

There are two strategies to run the two processes above. On the one hand, the learning and inference processes can be separated from each other, where the latter is run only at test time to produce predictions. This strategy is implemented in Deep Value Networks (Gygli et al., 2017). On the other hand, the inference process can be jointly trained with the learning process of energy-based models in an end-

to-end fashion as is implemented in the Structured Prediction Energy Networks (Belanger et al., 2017). Both of these strategies will be detailed in Section 5.3.

5.2 Energy-Based Dialogue State Tracking Model

Since a typical energy-based model consists of two components, I develop my energy-based dialogue state tracker broadly following the same lines. The first component is the feature function $F(X)$, that is implemented in the format of a hierarchical recurrent neural network to transform raw dialogue input X into fixed-size vector representations. The second component is the energy function $E(F(X), Y)$ that is implemented with a deep neural network to measure the alignment between a structured representation – set of values Y – and a set of features $F(X)$ in dialogue data. Here the feature network is based on the multi-task learning approach proposed in Chapter 4 with significant modifications to fit the new process.

5.2.1 Hierarchical Recurrent Neural Feature Network

Since dialogues generally consist of a sequence of turns that include machine acts and user utterances, the feature network in this work is implemented with a hierarchical recurrent neural architecture (Figure 5.1). The feature model contains three core structures to handle different dialogue entities:

- **User utterances** – a bidirectional LSTM (bi-LSTM) architecture (Huang et al., 2015) is used to generate a vector representation of the user utterance

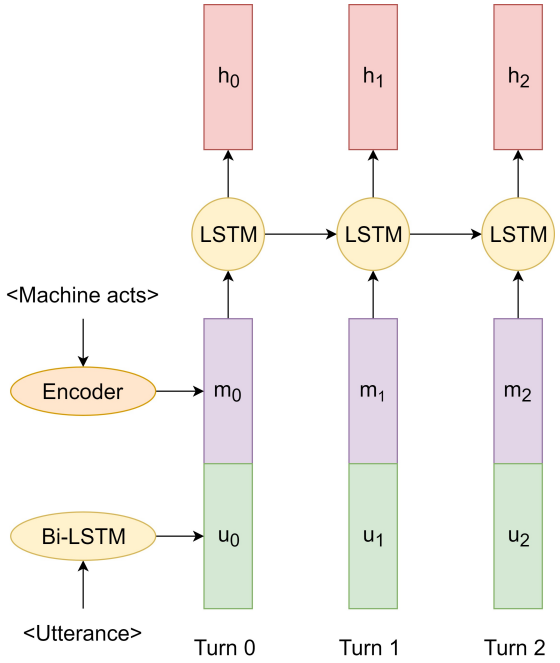

Figure 5.1: The hierarchical recurrent neural architecture to transform dialogue input into fixed-size vector representations. For the sake of simplicity of presentation, the unidirectional LSTM and concatenation layers are jointly presented as LSTM cells rolling up dialogue turns. m denotes encoded machine acts, u denotes vector representations for user utterances, and h denotes hidden states representing dialogue turn information.

at each turn.

- **Machine acts** – the machine acts at each turn, provided in a semantic format $act(slot = value)$, are parsed into a vector representation (Henderson, Thomson, & Young, 2014b), and fed through an encoder of two fully connected hidden layers to reduce vector dimensionality.

- **Dialogue turns** – the encoded vectors of machine acts and user utterances are concatenated to form dialogue turn input vectors. A unidirectional LSTM (Hochreiter & Schmidhuber, 1997) layer consisting of a number of LSTM cells is then used to roll up dialogue turns in order to build

up a representation of the dialogue, that includes dialogue history and current context. The outputs of this layer at each turn are then concatenated into a joint vector representation and treated as dialogue features for the energy-based model, that is detailed in the following section.

As explained by Belanger & McCallum (2016), the feature network should ideally be pretrained to improve the quality of features. Therefore I pretrain my feature network by plugging it into the multi-task learning architecture for dialogue state tracking in the style of Chapter 4. The focus of this research from here is only on the *joint goals* task, wherein each informable slot in joint goals is treated as a subtask and a multinomial classifier is developed for each slot. The output of each slot in the multi-task feature network is sampled with a *softmax* distribution and an *argmax* operation as per the common approach to the multinomial classification problem. All the feature network components are trained together in an end-to-end multi-task system. This pretrained feature network is then used to preprocess the inputs for the energy network, where the inputs are the results of the fully connected concatenation layer as described above.

5.2.2 Deep Neural Energy Network

In the energy-based learning method it is important to define an energy function that accounts for different types of associations in the system; in particular two types of relationships are considered: (i) the goodness of fit between inputs and candidate structured outputs; and (ii) the associations between output labels for

the structured prediction task. My differentiable energy function $E(F(X), Y)$ is formulated based on Belanger & McCallum (2016)'s Structured Prediction Energy Networks in that it is the summation of two energy terms, *local* and *global*, that serve as the two types of associations above.

The energy function $E(F(X), Y)$ is formalised as follow. To begin, the total energy is split into two component as follows:

$$E(F(X), Y) = E_{local}(F(X), Y) + E_{global}(Y) \tag{5.1}$$

where $E_{local}(F(X), Y)$ and $E_{global}(Y)$ are local and global energy terms respectively (see Figure 5.2).

Local energy $E_{local}(F(X), Y)$ is the measurement of goodness of fit between processed inputs $F(X)$ and structured outputs Y, and is computed as such:

$$E_{local}(F(X), Y) = \sum_{i=1}^{M} y_i W_i^\top F(X) \tag{5.2}$$

where weights W are trainable parameters, and M is the number of classes in the target.

Global energy $E_{global}(Y)$ captures the dependencies between outputs independently of the input features:

$$E_{global}(Y) = W_2^\top f(W_1^\top Y) \tag{5.3}$$

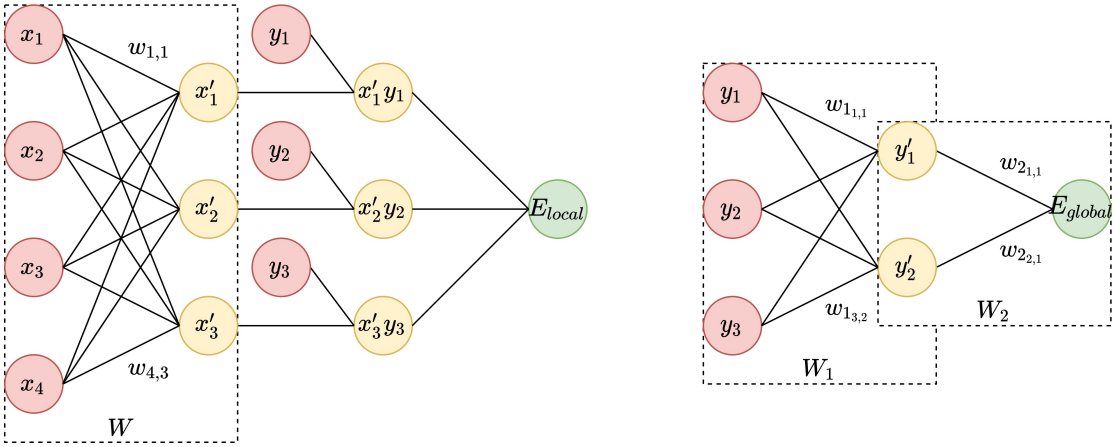

Figure 5.2: The deep neural structures of local and global energy functions.

where weights W_1 and W_2 are trainable parameters, and $f(\cdot)$ is a non-linearity function. In the experiments of this work, the non-linearity function is *softplus* that is a common differentiable activation function used in neural networks.

In this modelling, the output Y contains all the values of all informable slots in *joint goals*. The energy function $E(F(X), Y)$, presented in Equation 5.1, is not only used in the learning process to train the energy-based dialogue state tracker, but it is also used in the inference process to perform dialogue state predictions. It is noted that the training of the energy network is separated from the training of the feature network. Later sections will focus on the training of this energy network and how to use it to perform structured predictions. The predicted output is then sampled by selecting the value with the highest probability for each slot in the output Y.

5.3 Energy-Based Modelling Strategies

As mentioned earlier, the mechanism of energy-based modelling is split into two processes where a first learning process allows the model to learn to optimise the energy function, and a second inference process at runtime helps the model to produce predictions. Since these processes can be performed either separately or together in an end-to-end fashion, I experiment with both strategies in training my energy-based dialogue state tracker, and I describe each of these in detail below. In particular, in Sections 5.3.1 and 5.3.2 I describe the separate learning and inference processes, and in Section 5.3.3 I describe the end-to-end approach.

However, before proceeding further, it is worth noting that both of the approaches implemented in this work differ from the original concept of energy-based learning (LeCun et al., 2006), which was described in the example above (Section 5.1). In the standard energy-based learning framework, the energy function is trained to minimise energy values for correct input and output configurations while producing higher energy for incorrect sets of input and output. Meanwhile in my approaches, the energy network is either (i) trained to maximise the energy function that estimates an oracle F_1 score in the separate learning and inference setting, or (ii) indirectly trained through a series of predictions in the end-to-end setting. My approaches are considered variations of the energy-based learning methodology in the machine learning context (Tu, Pang, & Gimpel, 2020; Tu, Liu, & Gimpel, 2020; Osogami, 2017a).

5.3.1 Learning Process

The main challenge for the learning process is to define an appropriate objective function that can guide the training of the energy network. There are many options for designing the loss function for the learning process of an energy-based model depending on the architecture and the setting (LeCun & Huang, 2005).

In my work when separating the learning and inference processes, the energy-based dialogue state tracker is based on the Deep Value Networks algorithm (Gygli et al., 2017). In this setting the energy function $E(F(X), Y)$ is designed to estimate the compatibility of an input X and an output Y pairing $E(F(X), Y)$ with an oracle F_1 measurement, denoted as $E^*_{F_1}(Y, Y^*)$, between the said output Y and the ground truth Y^*:

$$E(F(X), Y) \sim E^*_{F_1}(Y, Y^*) \tag{5.4}$$

Here a cross entropy loss function $L(E, E^*_{F_1})$ is designed between the compatibility energy and the oracle F_1 value, since F_1 score falls into the range $[0, 1]$:

$$L(E, E^*_{F_1}) = -E^*_{F_1} \log E - (1 - E^*_{F_1}) \log(1 - E) \tag{5.5}$$

where $E = E(F(X), Y)$ is a predicted energy given an input X and a structured output Y, and $E^*_{F_1} = E^*_{F_1}(Y, Y^*)$ is the ground truth energy value measured between an output Y and the ground truth Y^*.

From Equation 5.5 I define two energy terms E and $E^*_{F_1}$:

- The first term, $E = E(F(X), Y)$, is the energy formulation based on Struc-

tured Prediction Energy Networks (Belanger & McCallum, 2016). The detail formulation is already described in Section 5.2.2.

- The second term, $E^*_{F_1} = E^*_{F_1}(Y, Y^*)$, is a variant of the dice coefficient form of the F_1 score, and serves as the ground truth energy during the training process. In my experiments this formulation is effective in the evaluation of multi-label classification outputs. The definition of this term is based on the formulation developed in the Deep Value Network model (Gygli et al., 2017), that defines the quality of any output Y with respect to the ground truth label Y^* by measuring an oracle value with an F_1 metric:

$$E^*_{F_1}(Y, Y^*) = \frac{2(Y \cap Y^*)}{(Y \cap Y^*) + (Y \cup Y^*)} \tag{5.6}$$

  where $Y \cap Y^* = \sum_i \min(y_i, y_i^*)$; and $Y \cup Y^* = \sum_i \max(y_i, y_i^*)$, that are modified from the original meaning to fit my continuous outputs.

In this setting, the energy function is trained to measure the quality of an output configuration Y given an input X with respect to the target Y^*, therefore it can be treated as a loss estimator. Here, the ground truth energy value $E^*_{F_1}(Y, Y^*)$ is defined in a non-standard supervised learning context, that is different from the target energy $E^* = E(F(X), Y^*)$ mentioned in Section 5.1.

A detailed explanation of the learning process of the energy-based dialogue state tracker is visualised in Figure 5.3. During the training process, all the parameters before and in the feature network are frozen as shown in the grey area,

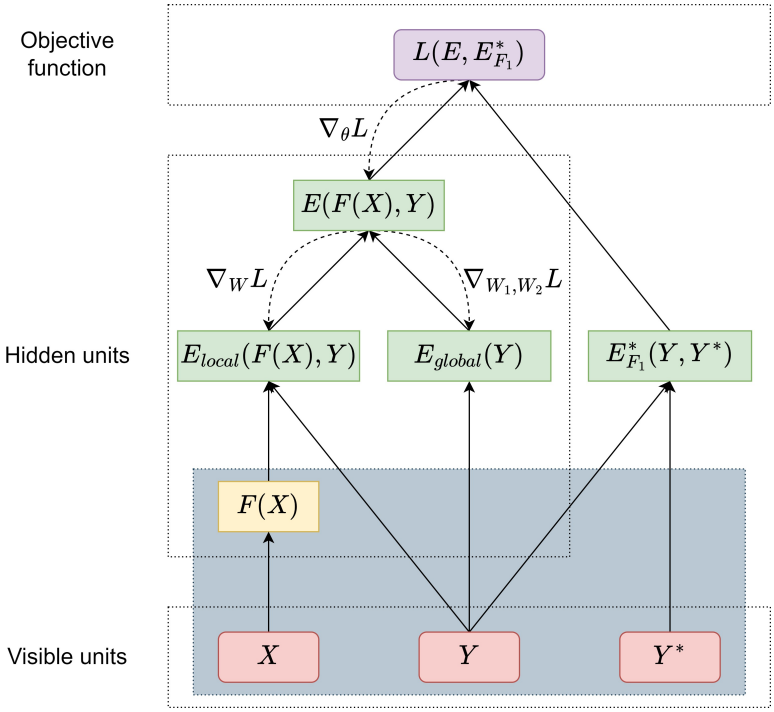

Figure 5.3: The learning process of the energy-based dialogue state tracker.

since they are pretrained in a multi-task learning model. These parameters include machine act encoder, word embedding for dialogue input, and bi-LSTM and LSTM parameters used in the feature network $F(X)$ (Section 5.2.1).

The trainable parameters θ for energy-based models are now only the energy network $E(F(X), Y)$ parameters, that are detailed in Section 5.2.2 as such:

$$\theta = \{W, W_1, W_2\} \tag{5.7}$$

where weights W are trainable parameters of the local energy, and weights W_1 and W_2 are trainable parameters of the global energy.

During the training process, the gradients of the errors are backpropagated

through the energy network, and the trainable parameters are updated accordingly:

$$\theta = \theta - \lambda \nabla_\theta L(E, E^*_{F_1}) \tag{5.8}$$

where λ is the learning rate, θ is the network's trainable parameters, and $\nabla_\theta L(E, E^*_{F_1})$ is the gradients of the errors with respect to trainable parameters.

In detail, and with the loss function as defined in Equation 5.5 the gradients are calculated as:

$$\nabla_\theta L(E, E^*_{F_1}) = \frac{dL}{dE}\frac{dE}{d\theta} = \left(-\frac{E^*_{F_1}}{E} + \frac{1 - E^*_{F_1}}{1 - E} \right) \frac{dE}{d\theta} \tag{5.9}$$

For a specific configuration of X, Y and Y^*, the term in brackets of Equation 5.9 is a fixed value with the contribution of both energy value $E(F(X), Y)$, that in turn includes global and local energy values, and oracle value $E^*_{F_1}(Y, Y^*)$. Meanwhile the differential $\frac{dE}{d\theta}$ varies based on the parameters to be updated.

Since the energy value is the summation of global and local energy terms, we can write:

$$\frac{dE}{d\theta} = \frac{dE_{local}}{d\theta} + \frac{dE_{global}}{d\theta} \tag{5.10}$$

where $E_{local} = E_{local}(F(X), Y)$ is the local energy in Equation 5.2, and $E_{global} = E_{global}(Y)$ is the global energy in Equation 5.3.

As defined above, $\theta = \{W, W_1, W_2\}$, where W contributes to only the local energy E_{local}, while W_1, W_2 appear in the global energy E_{global}. Therefore we now

can calculate the differential with respect to each parameter set:

$$\begin{aligned}
\frac{dE}{dW_2} &= \frac{dE_{global}}{dW_2} \\
\frac{dE}{dW_1} &= \frac{dE_{global}}{dW_1} = \frac{dE_{global}}{dW_2}\frac{dW_2}{dW_1} \\
\frac{dE}{dW} &= \frac{dE_{local}}{dW}
\end{aligned} \tag{5.11}$$

Subsequently by plugging Equation 5.11 into the gradients for backpropagation in Equation 5.9, we can update the trainable parameters, for example:

$$\begin{aligned}
\nabla_W L(E, E^*_{F_1}) &= \frac{dL}{dE}\frac{dE}{dW} = \left(-\frac{E^*_{F_1}}{E} + \frac{1 - E^*_{F_1}}{1 - E} \right) \frac{dE_{local}}{dW} \\
W &= W - \lambda \nabla_W L(E, E^*_{F_1})
\end{aligned} \tag{5.12}$$

Overall the proportional contribution of a specific parameter is defined by the gradients of the appropriate energy term error with respect to this parameter.

5.3.2 Inference Process

At runtime the predictions are performed through an inference process that is based around the energy function (Figure 5.4). This approach to producing predictions is different from the common feedforward deep learning prediction process.

During the inference process, the main challenge is that the energy $E(F(X), Y)$ must be constructed without output Y, while only having input X, that are preprocessed into feature representations $F(X)$. Therefore inference is performed as follows:

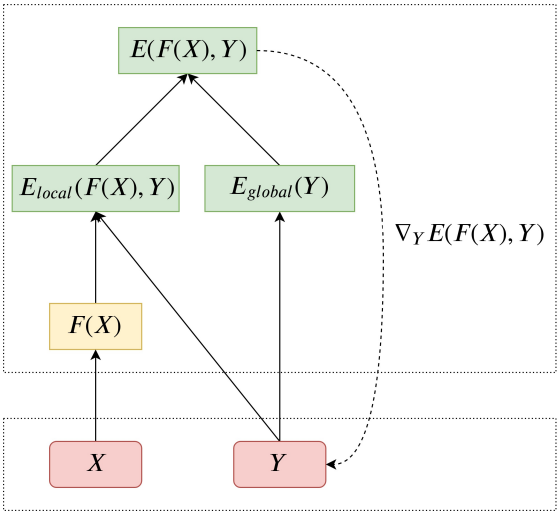

Figure 5.4: The inference process of the proposed energy-based dialogue state tracker.

- The process begins with an initial hypothesis $Y^{(0)}$, that is usually either a random hypothesis $Y^{(0)} = \{random(\cdot)\}^M$ or a zero vector $Y^{(0)} = \{0\}^M$, where M is the output dimensionality.

- An inference loop is then performed. The output hypothesis $Y^{(t)}$ is passed through the energy network to produce an energy estimate against the input X. Then the gradients of the energy with respect to the output are computed and used to update the output $Y^{(t+1)}$ of the next iteration in order to find the best fit according to the learned differentiable energy function:

$$Y^{(t+1)} = \mathcal{P}_Y\left(Y^{(t)} + \eta \nabla_Y E(F(X), Y^{(t)})\right) \tag{5.13}$$

where $\mathcal{P}_Y(\cdot)$ is the projection operation on the output, η is the inference learning rate, and $\nabla_Y E(F(X), Y^{(t)})$ is the gradients of the energy value with respect to the output.

Here, the operation $\mathcal{P}_Y(\cdot)$ is used to project the predicted output Y to the output range, i.e. $Y^{(t+1)} \in [0,1]^M$. A simple method to project the output is to clip the predicted probabilities by the value range.

- The inference outcome is treated as the prediction of the desired structured output:

$$Y^{(T)} \rightarrow Y^* \tag{5.14}$$

where T is the fixed number of iterations in the inference loop, $Y^{(T)}$ is the end prediction, and Y^* is the ground truth.

During the inference process gradient ascent techniques (Equation 5.13) are used within the loop to maximise the energy value in order to reach higher F_1 scores, since it is supposed to converge the prediction $Y^{(T)}$ towards the ground truth Y^*:

$$E^*_{F_1}(Y^{(T)}, Y^*) \rightarrow 1 \tag{5.15}$$

5.3.3 End-to-End Learning

In the previous sections I presented the energy-based learning experiment strategy that separates the learning and inference processes. An alternative approach for structured prediction is to model and train in an end-to-end fashion based on the Structured Prediction Energy Networks (Belanger et al., 2017). This modelling approach is visualised in Figure 5.5.

In this approach during training the loss of the network for a given training

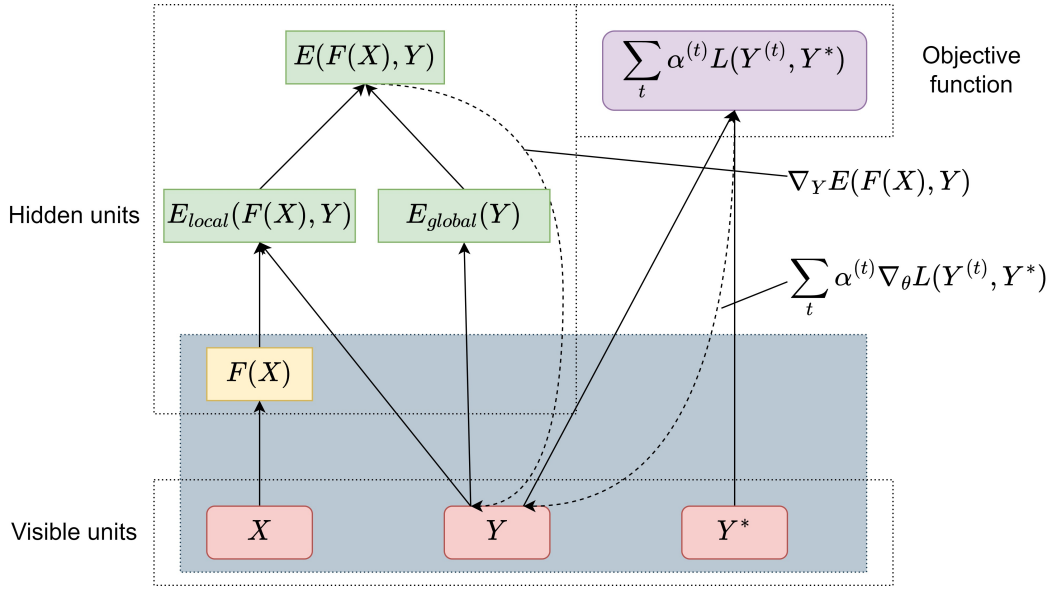

Figure 5.5: The working mechanism of the end-to-end energy-based dialogue state tracker.

instance is calculated through an iterative process wherein: (i) the input of this example is presented to the network, (ii) the system makes a series of output predictions as it converges on a desired energy state in response to this input, and (iii) the training loss is calculated by summing the losses of all the intermediate predictions with respect to the ground truth. This iterative calculation of the loss is defined as follows:

$$L(Y, Y^*) = \frac{1}{T} \sum_{t=1}^{T} \alpha^{(t)} L(Y^{(t)}, Y^*) \tag{5.16}$$

where T is the fixed number of iterations in the inference process, $L(Y^{(t)}, Y^*)$ is the loss function between the predicted output and the target, and $\alpha^{(t)} = \frac{1}{T-t+1}$ is the loss coefficient for the prediction at the iteration t.

In the end-to-end setting the objective function (Equation 5.16) implies that

the energy function $E(F(X), Y)$ is not evaluated directly as in the separate process setting (Equation 5.5), but through the predictions it helps produce. The reason the loss values of all generated output should be calculated along the inference path is that it encourages the energy function to produce good quality prediction at every iteration.

The inference algorithm of the end-to-end energy-based dialogue state tracker is adopted from the inference process used in Belanger et al. (2017)'s end-to-end Structured Prediction Energy Networks, wherein the process starts with a random output hypothesis and loops through a number of iterations to produce a structured prediction:

$$Y^{(t+1)} = \mathcal{P}_Y\big(Y^{(t)} - \eta^{(t)} \nabla_Y E(F(X), Y^{(t)})\big) \tag{5.17}$$

where $\mathcal{P}_Y(\cdot)$ is the projection operation on the output (same as in Equation 5.13 in the previous section), $\eta^{(t)}$ is the inference learning rate of the current iteration t, and $\nabla_Y E(F(X), Y^{(t)})$ is the gradients of the current energy with respect to the output.

To summarise the energy-based methodology, a brief comparison of the end-to-end and separate process algorithms is presented in Table 5.1. There are many similarities between both algorithms such as how to construct an energy network for all input features and structured outputs, and how the structured predictions are produced through an inference process. On the other hand, the differences between these two settings lie mainly in the manner of how the objective function is

Table 5.1: Comparison of the end-to-end and separate process algorithms for energy-based learning.

End-to-End	Separate process
Function *LEARNING (dataset $\mathcal{D}$, train parameters θ, learning rate λ)* **while** *not end of $\mathcal{D}$* **do** *Training sample* $(x, y^*) \in \mathcal{D}$ *Output generation* $y^{(T)} \leftarrow y^{(0)} - \sum_{t=1}^{T} \eta^{(t)} \nabla_y E(F(x), y^{(t)})$ *Objective function* $L \leftarrow \sum_{t=1}^{T} \alpha^{(t)} L(y^{(t)}, y^*)$ *Backpropagation* $\theta \leftarrow \theta - \lambda \nabla_\theta L$ **end** **end**	**Function** *LEARNING (dataset $\mathcal{D}$, train parameters θ, learning rate λ)* **while** *not end of $\mathcal{D}$* **do** *Training sample* $(x, y^*) \in \mathcal{D}$ *Output generation* $y \leftarrow Generate(x, \theta)$ *Predicted energy* $E \leftarrow E(F(x), y)$ *Ground truth energy* $E^*_{F_1} \leftarrow E^*_{F_1}(y, y^*)$ *Objective function* $L \leftarrow L(E, E^*_{F_1})$ *Backpropagation* $\theta \leftarrow \theta - \lambda \nabla_\theta L$ **end** **end**
Function *INFERENCE (input x)* *Output initialisation* $y^{(0)} \leftarrow Random(\cdot)$ *Output prediction* $y^{(t+1)} \leftarrow \mathcal{P}_Y\left(y^{(t)} - \eta^{(t)} \nabla_y E(F(x), y^{(t)})\right)$ *Return $y^{(T)}$* **end**	**Function** *INFERENCE (input x)* *Output initialisation* $y^{(0)} \leftarrow Random(\cdot)$ *Output prediction* $y^{(t+1)} \leftarrow \mathcal{P}_Y\left(y^{(t)} + \eta \nabla_y E(F(x), y^{(t)})\right)$ *Return $y^{(T)}$* **end**

defined, and subsequently the learning process. In detail, the end-to-end learning process depends on all the generated outputs on the inference path, while the learning process in the separate algorithm (Section 5.3.1) trains the energy network to be a loss estimator for the output against the given input and guides it towards the ground truth energy value. The separate learning process conceptually fits the general idea of energy-based learning better, as in this setting the energy function is specifically trained to measure the goodness of fit between input and output.

In terms of cost between these two energy-based methods, the end-to-end model requires much longer time to train, while the separate process approach can save time by making use of pretrained feature networks. However, the real-time performance on dialogue state predictions of these two models are not significantly different. When comparing the real-time performance of the energy-based models detailed in this chapter and the multi-task dialogue state tracker from Chapter 4, it is observed that there is no significant difference as the predictions are performed in the matter of milliseconds.

The hyper-parameters of my energy-based dialogue state trackers were optimised using a grid search method during the development phase and based on the literature review (Belanger et al., 2017; Gygli et al., 2017). The details of the model hyper-parameters are presented in Appendix A.

5.4 Results & Analysis

In this section I present the performance of my energy-based dialogue state tracking systems on DSTC2 & 3 data (Henderson, Thomson, & Williams, 2014a,b) while benchmarking them against state-of-the-art trackers.

5.4.1 Energy-based Modelling Performance

First the performance of the energy-based dialogue state trackers on the DSTC *joint goals* task is reported, and compared with the performance of state-of-the-art trackers (Table 5.2). The selected state-of-the-art trackers were chosen on the basis that they yield the best result, include some unique engineering techniques, or are otherwise related to this work. The feature system presented in Section 5.2.1 is pretrained in the multi-task learning manner detailed in Chapter 4, and the pretrained results are included to compare with the energy-based learning performances. All performances are reported with the *accuracy* metric as was historically common for work based on the DSTC2 dataset.

The main findings of applying the energy-based method to the dialogue state tracking tasks on DSTC2 & 3 data are two-fold:

- On the one hand, the energy-based trackers can improve the *joint goals* result on top of a feedforward deep learning architecture, namely the multi-task feature system. In detail the accuracy of tracking joint goals is improved up to 5% for DSTC2 and 9% for DSTC3. I believe that the key factor of this improvement lies within the interactions between slot outputs that my

Table 5.2: Performances of state-of-the-art and the energy-based dialogue state tracking systems on DSTC2 & 3 data.

Model	DSTC2	DSTC3
Sequence-to-sequence model (Feng et al., 2021)	**0.850**	-
Hybrid model with ASR features (Vodolan et al., 2017)	0.796	-
Web-style ranking system (J. D. Williams, 2014)	0.784	-
Multi-domain system (Mrksic et al., 2015)	0.774	**0.671**
Word-based system (Henderson, Thomson, & Young, 2014b)	0.768	-
Unsupervised RNN model (Henderson, Thomson, & Young, 2014a)	-	0.646
Global-locally self-attentive tracker (Zhong et al., 2018)	0.745	-
EncDec framework (Platek et al., 2016)	0.730	-
Conditional random fields tracker (S. Kim & Banchs, 2014)	0.601	-
This work		
Energy-based system (Separate processes)	0.760	0.622
Energy-based system (End-to-end)	0.749	0.585
Multi-task feature system	0.709	0.531
Baseline systems (Henderson, Thomson, & Williams, 2014a,b)	0.719	0.575

trackers account for in the learning process.

- On the other hand, when evaluating the two energy-based learning strategies, I find that the strategy where the learning and inference processes are separated outperforms the end-to-end approach. The improvement is more than 1% for DSTC2 and nearly 4% for DSTC3. This phenomenon, as I believe, is based on the fact that the energy function when explicitly learned performs better than indirectly trained via a series of predictions it helps produce. That said, better performance is achieved with more explicit slot dependencies learning when the energy function is trained specifically to measure the goodness of fit between input features and structured outputs, and among

different slot types.

Among dialogue state trackers on DSTC2 & 3 data, the results of the energy-based models are not yet competitive with the state-of-the-art systems such as the sequence-to-sequence model (Feng et al., 2021) and the multi-domain system (Mrksic et al., 2015). However, these trackers typically include a multitude of enhancements to achieve their high quality results. For example, Feng et al. (2021)'s sequence-to-sequence model makes use of the pretrained BERT architecture (Devlin et al., 2019), and implements an attention mechanism for the dialogue representations and the dialogue state decoding process. Meanwhile, Mrksic et al. (2015)'s multi-domain tracker is trained in several different dialogue corpora that also cover the DSTC3 tourism domain, and in the result the multi-domain tracker overcame the issue of low volume of training data and unseen states at test time.

It is worth noting that Vodolan et al. (2017)'s hybrid tracker has a similar approach to the energy-based model, but implements manual differential rules on top of a pretrained feature network instead of an energy function. The hand-crafted rules help their hybrid tracker outperform the energy-based method, but at the same time limit their system to the specific domain due to the lack of flexibility in data adaptation. On this point, the energy-based approach is more data-driven and flexible.

On the other hand, it should be highlighted that the web-style ranking system proposed by J. D. Williams (2014) achieved the highest result in the *joint goals* task among the DSTC2 entries during the competition time. However, this particular

system is very unique and hard to compare directly to my energy-based approaches.

The parsing technique proposed in the word-based tracker by Henderson, Thomson, & Young (2014b) has been found to be important and hence adopted into a number of state-of-the-art works (Vodolan et al., 2017; Henderson, Thomson, & Young, 2014a). In this work, I also implement this technique to handle machine acts, but did not achieve this system's performance. I believe that the reason behind this lies in the network architecture, as the word-based tracker was developed with a recurrent neural network whose cells are specialised for each slot and value in the domain.

There are several systems that also attempt to account for the relationships of slots during the learning process. However, their performance is limited for different reasons. The attention-based model (Zhong et al., 2018) considers the slot dependencies through a global-locally self-attentive encoder and a scoring module before producing the probability prediction. The limitation of this attention-based tracker is believed to lie in its end-to-end training fashion that does not focus on representation learning in a specific domain. In fact, while it performs averagely on DSTC2 data, it achieved state-of-the-art results at the published time on other datasets such as WOZ (Wen et al., 2017) and MultiWOZ 2.0 (Budzianowski et al., 2018). Meanwhile the EncDec framework (Platek et al., 2016) is limited within the incremental context, that has shown the disadvantage against turn-based tracking, while the conditional random field tracker (S. Kim & Banchs, 2014) did not perform well due to the manual feature representation technique it employed.

Table 5.3: Performances of the energy-based dialogue state tracking systems per slot and for *Joint goals* of those present in the task.

Dataset	Model	Slot				Joint goals
		food	price	area	type	
DSTC2	Energy-based tracker	0.872	0.938	0.923	-	0.768
	Multi-task feature system	0.825	0.929	0.919	-	0.717
DSTC3	Energy-based tracker	0.802	0.860	0.817	0.940	0.666
	Multi-task feature system	0.730	0.844	0.781	0.937	0.587

5.4.2 Analysis of Slot-Based Performance

To further investigate the performance of the energy-based approach, I conducted an analysis of the trackers' performance based on the prediction of each slot value and the *joint goals* of present slots (Table 5.3). Here, I see a gap between DSTC evaluation and my results as I omit the low frequency slots from the tracking process to focus on learning the slot dependencies. It should be noted that from here all analyses are conducted only for the energy-based system with separate processes as it achieved superior results over its end-to-end counterpart. As before, in this analysis the metric reported is *accuracy*.

In the results I observe that the energy-based technique improves the tracking results for each individual slot and for the overall *joint goals* as compared with the multi-task feature system from Chapter 4. The improvement varies from very small margins such as 0.3-0.4% accuracy for slots (*DSTC2.area*, *DSTC3.type*) that have small sets of values to a big change such as up to 7% accuracy for the slot *food* – the most challenging slot in both domains. Meanwhile, the overall *Joint goals* result is improved even more: 5.1% for DSTC2, and 7.9% for DSTC3.

Table 5.4: Proportional reduction in errors (%) of the energy-based system for each slot and the *Joint goals*.

Dataset	Slot				Joint goals
	food	price	area	type	
DSTC2	0.27	0.13	0.04	-	0.18
DSTC3	0.27	0.10	0.16	0.05	0.19

In addition, another analysis on the effectiveness of the energy-based approach is conducted with the use of a statistical method called proportional reduction in prediction errors (Kviz, 1981). This method measures the improvement on the predicted results of the energy-based model over the feature network by quantifying the reduction in the rate of errors in predictions. This improvement clearly indicates the effectiveness of the energy-based method on the dialogue state tracking task over a deep learning feed-forward architecture. The analysis result for each informable slot and the *joint goals* is reported in Table 5.4.

The analysis also shows that for more challenging slots such as *food*, the energy-based model reduces the error rate significantly, more than a quarter of errors are corrected (27%). Although the improvement for less challenging slots such as *area* in DSTC2 and *type* in DSTC3 is small, in both DSTC2 & 3 domains the error rates in *Joint goals* are reduced by nearly 20%.

Generally, the proportional reduction in error aligns with the overall performance of the energy-based tracking models.

5.4.3 Analysis of Slot Dependencies in Predicted States

Although I demonstrate that the energy-based approach improved the dialogue state tracking performance when applied on top of a feedforward deep learning architecture, I argue that the *accuracy* metric does not reflect the full capacity of my tracker's performance.

As outlined above, the structured prediction approach focuses on accounting for the slot dependencies in the state tracking process. Therefore I explicitly analyse the inter-slot dependencies in the performance of my models, and present the results of that analysis in Table 5.5. The slot dependencies analysis, described in Chapter 3, was performed on the values of the DSTC2 & 3 test data, and the predicted dialogue states of my energy-based model and multi-task feature system. Here, I performed Pearson's chi-square statistical tests to detect the slot dependencies, followed by a measure of the association strength with Cramer's V coefficient. The results confirm that inter-slot dependencies exist among all the slot types in DSTC2 & 3 test data, therefore Table 5.5 presents only an assessment of these dependencies in the Cramer's V coefficient.

The interpretation of the Cramer's V analysis is that better performance is reported by smaller margins in the Cramer's V coefficients between the tracker's evaluation and the *test values*. It should be noted that stronger associations do not necessarily indicate better tracking performance, hence my goal is to capture valid associations not to arbitrarily increase the number of associations seen in test data outputs.

Table 5.5: Analysis of slot dependencies on the DSTC2 & 3 test data. The results are reported in the Cramer's V coefficient.

	DSTC2				DSTC3			
Test data		food	price	area		food	price	area	type
	food	-			food	-			
	price	0.272	-		price	0.248	-		
	area	0.267	0.176	-	area	0.163	0.232	-	
					type	0.300	0.195	0.220	-
Energy-based system		food	price	area		food	price	area	type
	food	-			food	-			
	price	0.258	-		price	0.234	-		
	area	0.268	0.176	-	area	0.173	0.233	-	
					type	0.291	0.198	0.219	-
Multi-task system		food	price	area		food	price	area	type
	food	-			food	-			
	price	0.234	-		price	0.213	-		
	area	0.279	0.184	-	area	0.184	0.210	-	
					type	0.321	0.207	0.211	-

As expected the analysis results in Table 5.5 demonstrate that the energy-based tracker captures the slot dependencies seen in the test labels more consistently than does the multi-task approach. I argue that the ability to capture these slot dependencies as additional features for the prediction of dialogue states is the reason why the energy-based method outperforms the multi-task learning approach.

5.5 Summary

Since output slots in dialogue states are not assumed to be independent of each other, there exists a strong motivation to apply structured prediction approaches to the dialogue state tracking process. In my study I chose the energy-based learn-

ing method due to its notable effectiveness on capturing inter-slot dependencies and performing structured predictions. Implementing energy-based dialogue state tracking systems is a novel modelling approach.

The contributions of my study on energy-based learning were two-fold. Firstly, the results of my work strengthen the hypothesis that accounting for the slot dependencies while tracking dialogue states has a positive impact on the outcomes. Secondly, I demonstrated how slot dependencies can be addressed in the dialogue state tracking process with a structured prediction method. These findings have been verified on the second and third DSTC datasets.

Although my results do not in themselves improve on the state of the art, the difference relative to a multi-task model (from Chapter 4) is significant enough to indicate that the energy-based learning method is promising and can lead to improvements if combined with other methods.

Overall, I suggest that since the energy-based modelling enhancement is actually quite modular with respect to the baseline, my method when incorporated into other state-of-the-art models is likely to enhance state-of-the-art performance.

Chapter 6

Slot Value Regularisation for Energy-Based State Tracking

In the previous chapter, I demonstrated how a structured prediction method helps improve dialogue state tracking performance. The method approaches the task as a multi-label classification problem. However, a multi-label classification approach does not ensure the dialogue state requirement in task-oriented dialogue domains such that at any turn of the dialogue one and only one value should be classified for a particular slot. Consequently, the energy-based learning approach suffers from this limitation. Therefore, in this chapter an enforcement approach to energy-based dialogue state tracking is proposed, so that (i) the constraints on dialogue state slots are ensured, and (ii) the structured prediction method is improved.

To demonstrate the consistency of this improvement, the dialogue data are expanded to include another corpus, Wizard-of-Oz (WOZ) 2.0 (Wen et al., 2017),

beside the two DSTC2 & 3 datasets (Henderson, Thomson, & Williams, 2014a,b). The common characteristic of these three datasets is that they all cover a single dialogue domain. However as discussed in Chapter 2, the WOZ dataset contains chat-based dialogues, which makes it different from DSTC2 & 3 datasets which are based on spoken conversations.

This chapter is structured as follows: first, an overview of dialogue state constraints is presented in Section 6.1. Then the modifications of the energy-based dialogue state tracking system are detailed in Sections 6.2 and 6.3. The results and their analysis are detailed in Section 6.4. The chapter is concluded in Section 6.5.

The work in this chapter was published at the 29th International Conference on Artificial Neural Networks (ICANN) (Trinh et al., 2020b).

6.1 Overview of Dialogue State Constraints

As outlined in previous chapters, dialogue states in task-oriented dialogue systems are typically defined as sets of slot and value pairs, therefore the dialogue state tracking task can be interpreted as a multi-task classification problem, where assigning correct values for each slot can be treated as an individual classification task. A common specific requirement of this slot and value assignment is that for each turn of the dialogue only one correct value from the ontology is assigned for the said slot. Let us look at an example of dialogue states with and without slot value constraint rules presented in Table 6.1.

In this example the predicted dialogue states {*food = Italian, price range =*

Table 6.1: An example of dialogue states with and without slot value constraint rules.

Slot	without Constraint Rules	with Constraint Rules
food	**P(Italian) = 0.995** P(Chinese) = 0.900 P(American) = 0.110	**P(Italian) = 0.995** P(Chinese) = 0.003 P(American) = 0.002
price range	**P(cheap) = 0.950** P(expensive) = 0.885	**P(cheap) = 0.950** P(expensive) 0.050
area	**P(centre) = 0.950** P(south) = 0.921	**P(centre) = 0.950** P(south) = 0.050

cheap, area = centre} are correct in both cases. However, if I set the threshold for activated values below 0.9, the classifier without slot value constraint rules activates two values for slot *food* at the same time, and this can lead to confusion in the interpretation of dialogue state predictions. Namely that the prediction from a system can be correct as far as for each slot the top-ranked value for the slot is correct. But the prediction still does not follow the slot value constraint rules in terms of the set of activated values as such: (i) if more than one value has an activation value above the threshold; or (ii) conversely the activation value of the top-ranked value is below the threshold. On the other hand, the classifier that obeys these rules can clear this confusion by predicting only one value for the said slot at the same time.

Here, the tracking task for a specific slot is in itself a multinomial classification problem. Thus the requirement of assigning only one value for each output can be easily achieved with different techniques, for example applying a *softmax* activation function to normalise the output probability distribution and using an

argmax sampling method to select the output with the highest probability. Various deep learning approaches have been proposed to tackle the dialogue state tracking problem in this manner, both as a combination of individual models (Mrksic et al., 2017; Vodolan et al., 2017) or in a multi-task learning-based fashion (Trinh et al., 2018; L. Ren et al., 2018).

However, when proposing the structured prediction method, specifically the energy-based learning modelling, for the dialogue state tracking problem, the enforcement of the constraint across multiple slots becomes less straightforward. This practice is similar to various approaches such as the Global-locally self-attentive model (GLAD) (Zhong et al., 2018) and Globally-conditional encoding system (GCE) (Nouri & Hosseini-Asl, 2018). While classic multi-label classification methods assume independence between class values, structured prediction approaches aim to explore the impact of value dependencies in the task. From a practical point of view as outlined in Chapter 5, my energy-based dialogue state tracking system demonstrates significant improvements over a classic deep learning approach. In general, in the energy-based methodology value dependencies are captured via an energy function, that can be implemented with various machine learning techniques, and in my case a deep learning architecture (LeCun et al., 2006).

To achieve the goal of enforcing the constraint, I propose a value regularisation approach to Energy-Based Learning (EBL) to enforce the rule that there is only one activated value for each tracked slot at any time during the conversation. In the following sections I detail this slot value constraint approach, and conduct

a number of analyses on the impact of this approach on structured prediction performance.

6.2 Modified EBL Architecture

As presented in the previous chapter, the architecture of my EBL dialogue state tracking model consists of two components: a feature network $F(X)$, and an energy network $E(F(X), Y)$, where X and Y are input and output respectively. As in the previous chapter, I apply the dialogue state tracker to three corpora: WOZ 2.0 (Wen et al., 2017), DSTC2 (Henderson, Thomson, & Williams, 2014a) and DSTC3 (Henderson, Thomson, & Williams, 2014b). As the *joint goals* task in these domains are similar to each other, the energy network $E(F(X), Y)$ is designed in the same manner as presented in Section 5.2.2. Here the modification is mainly focused on the feature network $F(X)$ due to the difference in dialogue input between the datasets.

6.2.1 Multi-Task Recurrent Neural Feature Network

As presented in Section 5.2.1, the feature network $F(X)$ is designed with a hierarchical recurrent neural network architecture to transform dialogue data into fixed-size vector representations (Figure 6.1). The architecture consists of three main components:

- **User input**, given in the format of sentences, is processed with an embedding layer and a bidirectional LSTM layer (Huang et al., 2015).

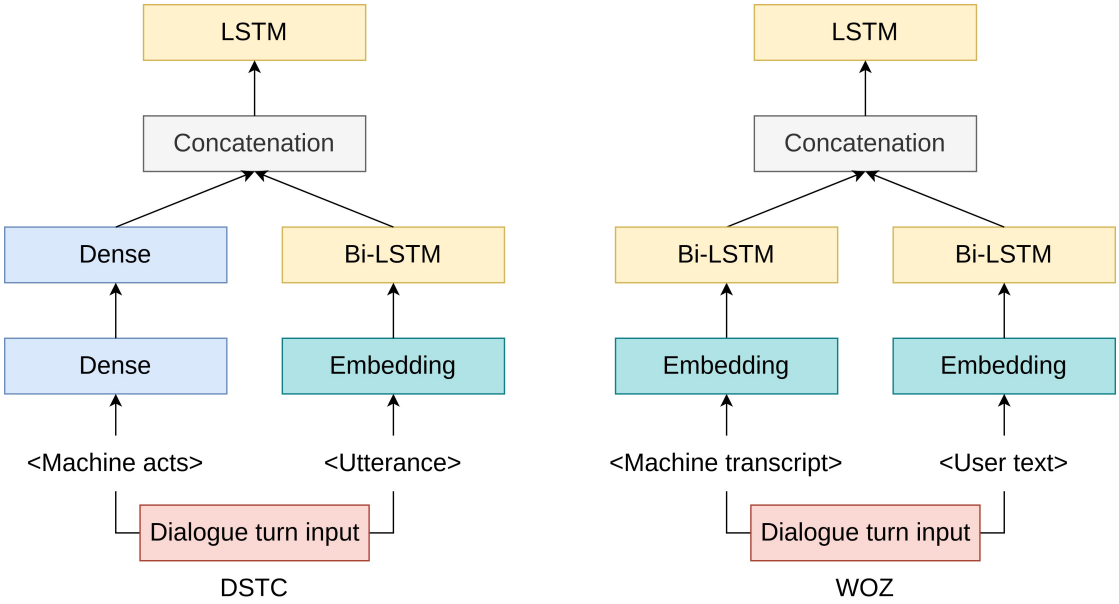

Figure 6.1: Multi-task Recurrent Neural Feature Network for DSTC and WOZ datasets.

- **Machine input** is provided in different formats in the DSTC and WOZ datasets. In the DSTC data, the machine input has a semantic format, that is parsed with the technique proposed by Henderson, Thomson, & Young (2014b), and an encoder consisting of two dense layers is developed to reduce the vector dimensionality. On the other hand, in the WOZ data, the machine input is provided in transcript format, thus it is processed similarly to user input, with an embedding layer and a bidirectional LSTM layer.

- **Dialogue turn input** is a concatenated vector of processed user and machine input, that is handled by a LSTM layer (Hochreiter & Schmidhuber, 1997). The output of this LSTM layer is treated as the dialogue representation on a turn-based basis.

Following the common practice, and as described in the last chapter, my feature

network $F(X)$ is pretrained in a multi-task learning manner to achieve higher results (Belanger & McCallum, 2016). The dialogue representations extracted with this feature network are in turn fed into the subsequent energy network to perform dialogue state predictions. As mentioned above, the architecture of my energy network $E(F(X), Y)$ remains unchanged (details in Section 5.2.2).

6.3 A Modified Learning Process

As detailed in Section 5.3, the working mechanism of an energy-based model is split into learning and inference processes. In the previous chapters, I demonstrated that the energy-based model yielded better results when these processes were run separately. For that reason, in this experiment I focus only on that separate strategy.

On the one hand, the inference process remains unchanged as presented in Section 5.3.2. On the other hand, the learning process depends on how the objective function is defined to guide the energy function to produce desired energy values for correct input and output configurations. In my energy-based system, I implement a variant of the energy-based learning methodology based on the Deep Value Networks (Gygli et al., 2017), that uses a F-measurement to evaluate the compatibility between predicted output and ground truth. That being said, the energy function here is trained to estimate the quality of an output Y given an input X with respect to the ground truth label Y^* (details in Section 5.3.1). Thus in this section I propose two modifications for the learning process: (i) redefining

the F_1 measurement to better fit multi-label classification problem, and (ii) introducing value regularisation into the objective function to enforce the slot value constraints of dialogue states. It is worth noting that the model hyper-parameters in this chapter are the same as of the energy-based models developed in Chapter 5 (see Appendix A).

6.3.1 Ground Truth Energy

As presented in the previous chapter, the ground truth energy $E^*_{F_1}$ is defined through the use of the dice coefficient F_1 metric in a differentiable format (Gygli et al., 2017). For presentation purpose only, I repeat Equation 5.6 in a detailed formulation:

$$E^*_{F_1}(Y, Y^*) = \frac{2\sum_i \min(y_i, y_i^*)}{\sum_i \min(y_i, y_i^*) + \sum_i \max(y_i, y_i^*)} \tag{6.1}$$

where $Y = \{y_i\}^M$ is the predicted output, and $Y^* = \{y_i^*\}^M$ is the ground truth.

However, in this formulation it can be seen that $\sum_i \min(y_i, y_i^*)$ is the lower boundary and $\sum_i \max(y_i, y_i^*)$ is the upper boundary of these values, that indicate the extreme values among all output configurations. Meanwhile, from another perspective in a multi-label classification task the differentiable F_1 metric has a more relaxed form (B. Wang et al., 2017), defined as follow:

$$E^*_{F_1}(Y, Y^*) = \frac{2\sum_i y_i y_i^*}{\sum_i y_i + \sum_i y_i^*} \tag{6.2}$$

where $Y = \{y_i\}^M$ is the predicted output, and $Y^* = \{y_i^*\}^M$ is the ground truth.

Here it is important to pay attention to the fact that any ground truth y_i^* can hold only the value 0 or 1, while the output y_i holds a continuous value in the range $[0, 1]$. Therefore, when I compare the terms on the right side of Equations 6.1 and 6.2, it is not difficult to mathematically prove that:

$$\begin{aligned} \sum_i \min(y_i, y_i^*) &= \sum_i y_i y_i^* \\ \sum_i \min(y_i, y_i^*) + \sum_i \max(y_i, y_i^*) &= \sum_i y_i + \sum_i y_i^* \end{aligned} \tag{6.3}$$

Despite what is shown in Equation 6.3, it is arguable that the differential process is discontinuous in Equation 6.1 based on the nature of the operations min and max. However, this is not an issue if Equation 6.2 is used.

For training purposes, the loss function between predicted and ground truth energy values remains a cross entropy loss as in Section 5.3.1. A slight difference here is that the ground truth energy formula is replaced:

$$L(E, E_{F_1}^*) = -E_{F_1}^* \log E - (1 - E_{F_1}^*) \log(1 - E) \tag{6.4}$$

where $E = E(F(X), Y)$ is the predicted energy, and $E_{F_1}^* = E_{F_1}^*(Y, Y^*)$ is the ground truth energy defined in Equation 6.2.

6.3.2 Slot Value Regularisation

The dialogue slot value constraint rules require that only one value be assigned to a slot at any time during the conversation. Hence I propose a slot value regularisation

approach that penalises predictions on the basis of the difference between the sum of the predicted activations and the sum of ground truth activations, using the sum of activations as the basis for the comparison between the predicted and ground truth activations, rather than the count of the activations, allows us to define a differentiable regularisation term. This slot value regularisation term is formulated as follow:

$$R(Y, Y^*) = \left(\frac{\sum_i y_i - \sum_i y_i^*}{\sum_i y_i^*} \right)^2 \tag{6.5}$$

where $Y = \{y_i\}^M$ is the predicted output, and $Y^* = \{y_i^*\}^M$ is the ground truth.

Here, the slot value regularisation penalises the predictions by measuring the difference in the sum of activated values between the predicted output and the ground truth. The use of regularisation term in Equation 6.5 is based on the more general meaning of regularisation, and is fundamentally different from the L_2 or L_1 regularisation techniques that instead penalise excessive parameter values. This slot value regularisation formula is differentiable in the training process.

Now that the slot value regularisation term is defined, it should be used in the objective function to guide the learning process of my energy network. The ultimate objective function including the slot value regularisation term is thus formulated as follow:

$$\begin{aligned} \mathcal{L} &= L(E, E^*_{F_1}) + \alpha R(Y, Y^*) \\ &= \Big(- E^*_{F_1} \log E - (1 - E^*_{F_1}) \log(1 - E) \Big) + \alpha \left(\frac{\sum_i y_i - \sum_i y_i^*}{\sum_i y_i^*} \right)^2 \end{aligned} \tag{6.6}$$

where α is the regularisation coefficient.

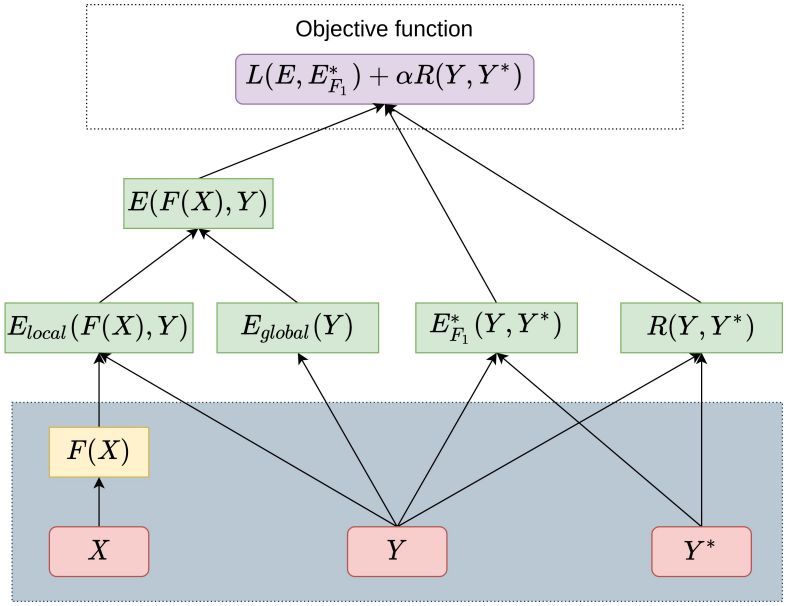

Figure 6.2: The learning process including the value regularisation of the proposed energy-based dialogue state tracker.

The new learning process with the slot value regularisation is visualised in Figure 6.2. In this process, the slot value constraints are measured between the generated output Y and the ground truth Y^* beside computing the target energy value $E^*_{F_1}(Y, Y^*)$. The predicted energy value $E(F(X), Y)$ is constructed as usual between the generated output and the given input. All these three items are then used in the objective function for the learning process.

6.4 Results & Analysis

In this section, I first report the overall performance of my modified energy-based model with respect to three dialogue corpora: DSTC2 (Henderson, Thomson, & Williams, 2014a), DSTC3 (Henderson, Thomson, & Williams, 2014b), and WOZ (Wen et al., 2017). The reported performance also repeats the result of my original

energy-based approach in Chapter 5 for ease of comparison.

The section then proceeds to conduct a number of analyses on the impact of my modifications detailed above. As the contributions of my energy-based approach in this chapter are based around the redefinition of the F_1 measurement and the regularisation of label constraint rules of dialogue states, I conduct a series of analyses to investigate these phenomena in particular.

6.4.1 Modified Energy-Based Modelling Performance

The modified model is evaluated in three single domain dialogue corpora and benchmarked against a number of state-of-the-art systems (Table 6.2). The reported task is *joint goals*, and the evaluation metric is *accuracy*.

In the results, we can observe two improvements in the performance of the energy-based learning approach. On the one hand, the energy-based learning method boosts the results over the multi-task learning methodology by up to 12% accuracy, across the datasets. I believe that the energy-based system achieved this improvement based on the significant impact of slot dependencies in dialogue domains. On the other hand, we also find that the modified energy-based system outperforms the original by 1.4% accuracy for the DSTC2 data and 2.9% accuracy for the DSTC3 data. I believe the performance improvement is the result of the modification techniques presented above. The detailed analyses of this question is presented in the following sections.

The result in Table 6.2 demonstrates that although the performance of my

Table 6.2: Performances of the state-of-the-art and the proposed dialogue state tracking systems on the DSTC 2 & 3 and WOZ data.

Model	DSTC2	DSTC3	WOZ
Amendable generation model (Tian et al., 2021)	-	-	**0.914**
Sequence-to-sequence model (Feng et al., 2021)	**0.850**	-	0.912
Globally-conditioned encoder (GCE) (Nouri & Hosseini-Asl, 2018)	-	-	0.885
Hybrid model with ASR features (Vodolan et al., 2017)	0.796	-	-
Multi-domain system (Mrksic et al., 2015)	0.774	**0.671**	-
Word-based system (Henderson, Thomson, & Young, 2014b)	0.768	-	-
Global-locally self-attentive tracker (GLAD) (Zhong et al., 2018)	0.745	-	0.881
Unsupervised RNN-based system (Henderson, Thomson, & Young, 2014a)	-	0.646	-
This work			
Modified energy-based system	0.774	0.651	0.875
Energy-based system (Chapter 5)	0.760	0.622	-
Multi-task feature system	0.709	0.531	0.841
Baseline (Henderson, Thomson, & Williams, 2014a,b; Mrksic et al., 2017)	0.719	0.575	0.844

modified energy-based model is competitive, it does not yet overcome the state-of-the-art systems such as the sequence-to-sequence model for DSTC2 (Feng et al., 2021), the multi-domain system for DSTC3 (Mrksic et al., 2015), and the amendable generation model for WOZ (Tian et al., 2021). However, none of other published systems, to the best of my knowledge, apply the slot dependencies in an explicit manner for predicting dialogue states. Thus I argue that their performance could be improved if structured prediction, in particular the energy-based learning, is applied.

Table 6.3: Performances of the energy-based dialogue state trackers with different F_1 metrics on the DSTC2 & 3 data.

Energy-based DST model	DSTC2	DSTC3
With redefined F_1 metric (Equation 6.2)	0.769	0.642
With dice coefficient F_1 (Equation 6.1)	0.760	0.622

6.4.2 Effectiveness of Redefined F_1 Metric

The first analysis focuses on the performance improvement of the modified energy-based dialogue state tracker based on the F measurement, and is presented in Table 6.3. Originally I implement the dice coefficient F_1 metric proposed in Deep Value Networks (Gygli et al., 2017). My updated version of the F_1 metric is proposed for continuous outputs in multi-label classification problems. The F_1 formulas are detailed respectively in Equations 6.1 and 6.2.

To evaluate the improvement based on the F_1 metric, I compare my energy-based system during the development phase with the one developed in Chapter 5. Here my system does not include the value regularisation, hence the performance is lower than reported in Table 6.2. The result is reported in *accuracy* for the *joint goals* of the DSTC2 and DSTC3 data.

The analysis result demonstrates the energy-based model with a redefined F_1 metric yields a slightly improved performance, that is approximately 1% higher accuracy for DSTC2 and 2% higher accuracy for DSTC3. This result strengthens my hypothesis regarding the impact of F_1 measurement for continuous outputs in the multi-label classification context.

Table 6.4: Performances of the energy-based dialogue state trackers with and without value regularisation on the DSTC2 & 3 and WOZ data.

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

6.4.3 Effectiveness of Slot Value Regularisation

I investigate the effectiveness of the slot value regularisation approach in a series of analyses in this and the following sections. Firstly, I evaluate the energy-based model's overall performance with and without the value regularisation term. Secondly, I evaluate the quality of dialogue states by studying the proportion of correct dialogue state predictions that satisfy the slot-value constraint rules with different threshold settings. And finally, I investigate the error distributions among dialogue state predictions to analyse the impact of slot value regularisation.

The result of the first slot value regularisation analysis is presented in Table 6.4 where I benchmark my energy-based model both with and without the value regularisation term presented in Equation 6.5. The model's performance is evaluated with the *accuracy* metric for the *joint goals* task in all three corpora.

The analysis result demonstrates that the performance accuracy is improved slightly across all three datasets, with nearly 1% accuracy improvement in all 3 cases. This finding shows that slot value regularisation has a clear role in the performance increase. However, it is arguable that the accuracy metric does not contain much information for the evaluation and analysis purpose. Thus, I further rely on other analyses.

Table 6.5: Analysis of the value regularisation on the energy-based dialogue state tracking on the DSTC 2 & 3 and WOZ data.

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

6.4.4 Analysis of Slot Value Constraint Rules

The second analysis on the impact of the slot value regularisation approach determines whether this regularisation term enforces the slot-value constraints or not. As mentioned above, the slot-value constraint rules for dialogue states require that at every conversation turn there is only one value assigned for each slot. In this analysis different thresholds are set, thus in this setting a value is considered activated if its predicted probability exceeds the set threshold. I then calculate the proportion of predictions where dialogue states are correct and follow the constraint rules with different thresholds. The analysis result is presented in Table 6.5. Here, three thresholds 0.5, 0.7 and 0.9 are chosen, and the presence and absence of the value regularisation term are denoted with *+/–Reg*.

The analysis result demonstrates that the value regularised energy-based system consistently outperforms the system without this value regularisation term with different thresholds across all three datasets. The interpretation of this outperformance is that more dialogue states produced with the value regularisation term satisfy the slot-value constraint rules of the chosen dialogue domains. This result indicates the impact of the proposed value regularisation term, that along-

side improving the system's performance can guide the system's prediction process towards the specific domain requirements.

6.4.5 Analysis of Error Distributions

I conduct a final analysis on the error distributions among dialogue states to emphasise the effectiveness of the slot value regularisation approach to slot value constraints. Among analyses on different modelling methods to dialogue state tracking, the comparative error analysis proposed by Smith (2014) offers a view of three error types that account for possible deviations from the joint goals in the dialogue in DSTC2. These errors indicate the advantages and disadvantages of each tracking algorithm to the given *joint goals* task with respect to produced errors. I find this analysis useful in my research, but it requires some modifications to fit my purposes with an example demonstrated in Table 6.6:

- **Missing attributes (MA)** is the error where a value for a slot is mentioned in data but not predicted by the model. In my interpretation, this error occurs when my slot value regularisation aware system assigns the number of activated values less than the number of slots.

- **Extraneous attributes (EA)** is the error where the tracker overpredicts unnecessary values for a slot even if they do not appear in data. Here in my work, this error means the number of activated values is bigger than the number of slots.

Table 6.6: An example of the three error types of dialogue state tracking. *MA* denotes missing attributes, *EA* denotes extraneous attributes, and *FA* denotes false attributes.

Utterance	*I want Chinese food not too expensive*
Correct	*food = Chinese* *pricerange = moderate*
MA type	*food = Chinese* *pricerange* – ***missing***
EA type	*food = Chinese* *pricerange = moderate* *area = centre* – ***extra***
FA type	*food = Asian fusion* – ***false*** *pricerange = moderate*

- **False attributes (FA)** is the error of a false value being assigned to a slot. In this situation the number of activated values equals the number of slots, that satisfies the slot value constraint rules, but the predicted dialogue state is still wrong due to an incorrect value.

In this analysis I compare the behaviours of my energy-based dialogue state tracking model when the slot value regularisation is included and excluded (*+/– Reg*), with the activation threshold set to 0.5. The analysis result (Table 6.7) reports the absolute error count and the proportion (%) of these errors with the respective type among the total number of wrongly predicted turns. The absolute error count result shows that when the slot value regularisation is included, the number of errors decreases overall. I note that the number of wrongly predicted turns varies when the model includes and excludes the value regularisation, therefore the main indication of this analysis is based on the proportion result.

The analysis results demonstrates a shifting trend of errors under the influence

Table 6.7: Error distributions of the energy-based dialogue state tracker on the DSTC2 & 3 and WOZ data.

Dataset	Label	#Turns	Error distributions (%)		
			MA	FA	EA
DSTC2	+Reg	6094	19.5	57.4	23.1
	–Reg	6222	28.9	40.2	30.9
DSTC3	+Reg	6588	26.4	50.8	22.8
	–Reg	6795	32.0	37.8	30.2
WOZ	+Reg	641	15.3	63.6	21.1
	–Reg	696	33.6	35.1	31.3

of slot value regularisation term. When the value regularisation is excluded, the errors are distributed more evenly among all three types. The special case is observed in WOZ data. On the other hand, my energy-based tracking model with a value regularisation shifts the errors towards the *FA* type, that means the majority of errors despite being wrong predictions for dialogue states still follow the slot value constraint rules. This finding is vital for my proposed approach such that the slot value regularisation term is effective in guiding the training process of my energy-based tracker.

As I base this analysis on the comparative analysis of Smith (2014), it is important to note that my findings are different. Here, I outline the error distributions within the slot-value constraint rules that are required by dialogue domains. Meanwhile the analysis of Smith (2014) shows the error distributions with respect to the difficulty of dialogue slots. For instant, in the DSTC2 data the errors are distributed in order $\{food >> area >> price\ range\}$, that follows the setting of the ontology where the slot *food* has a much bigger set of values than the other two. This difference can be explained by the approach to the dialogue state track-

ing task, whereas the task is treated as a multi-label classification problem in my energy-based structured prediction approach, while most of the DSTC2 trackers solve it in a multinomial classification manner.

6.5 Summary

In this chapter, I demonstrate a two-pronged improvement approach to the energy-based structured prediction method in the context of dialogue system. The improvement consists of an optimisation of the quality measurement, and a value regularisation for constraint integration. I demonstrate that the overall performance of my energy-based model is increased in a number of dialogue datasets. Further analyses show that my energy-based model's behaviours achieve a high level of performance with respect to constraints. My systemic analyses, in particular the analysis on error distributions, are essential to understand the mechanism of the dialogue state tracking process, subsequently it would help to improve future models.

In detail, changing the quality measurement from a dice coefficient F_1 (Equation 6.1) to a differentiable F_1 metric (Equation 6.2) boosts the overall performance of my energy-based model. That demonstrates the effectiveness of differentiable metrics in deep learning when working with continuous values.

Task-oriented dialogue domains have specific requirements towards dialogue states such as slot value constraint rules. A simple multi-label classification does not satisfy this strictness. However, including a slot value regularisation term into

the learning process of my energy-based model has proven to be effective. On the one hand, it helps improve the overall performance of my model. On the other hand. the analyses above indicate that this value regularisation also guides the prediction process to satisfy slot value constraints in dialogue domains.

Overall, to the best of my knowledge the redefinition of the objective function, that includes redesigning the ground truth energy and introducing the slot value regularisation, is a novel contribution to structured prediction in dialogue state tracking.

Chapter 7

Generalisability to Multiple Dialogue Domains

In the previous chapters I have shown that energy-based learning, a specific form of structured prediction methods, has proven to be effective for the dialogue state tracking task in the single domain setting. Before concluding this dissertation, in this chapter I demonstrate the generalisation ability of this method to multiple dialogue domains. The generalisability is an important characteristic for the structured prediction methodology on the application level, as modern dialogue systems are often developed in a multi-domain environment (Ram et al., 2017; Khatri et al., 2018; Gabriel et al., 2020).

In order to investigate the generalisability of my energy-based learning approach, I choose two common multiple domain dialogue corpora for the task, namely MultiWOZ 2.0 (Budzianowski et al., 2018) and MultiWOZ 2.1 (Eric et

al., 2020). The large number of classes across different domains covered by these datasets brings a challenge for applying my approach.

This chapter begins with an overview of multi-domain dialogue in Section 7.1. In Section 7.2 the architecture of a large-scale energy-based dialogue state tracker is presented, followed by the processes of energy-based learning (EBL) in Section 7.3. The EBL model is modified to accommodate the structure of multiple domain dialogue data. Then, the experiment results and analysis are detailed in Section 7.4. The chapter is concluded with a brief summary in Section 7.5.

The work in this chapter was published at the 4th Workshop on Structured Prediction for Natural Language Processing (Trinh et al., 2020a).

7.1 Overview of Multi-Domain Dialogue

With task-oriented dialogue systems being widely used in various fields, expanding from single domain to multiple domains is a rising trend. Scaling up dialogue systems can improve the generalisability of models and support transferring knowledge across domains. The benefit of multiple domains processing has been demonstrated by a number of research efforts, for example Mrksic et al. (2015) shows that a tracking model trained in multiple domains yields better results across domains than single domain trackers developed with the same approach. However, it also leads to the challenges such as handling an increased number of dialogue slots and values and enlarging the work load of dialogue system components that include the dialogue state tracker. In my work, constructing an energy-based model for

multiple dialogue domains faces a similar challenge.

As presented in previous chapters, my energy-based structured prediction method approaches the dialogue state tracking task as a multi-label classification problem. This approach is in contrast to the traditional multi-task classification approach where a number of tracking models can be developed to solve the dialogue state tracking task for each domain separately (Heck et al., 2020; C.-S. Wu et al., 2019; L. Zhou & Small, 2019). That being said, my approach aligns with the recent advanced techniques to track dialogue states in a multi-domain environment, that achieve state-of-the-art results (S. Kim et al., 2020; A. Kumar et al., 2020). However, the novelty of my structured prediction method lies in the explicit accommodation of slot dependencies in the prediction process that, to my knowledge, other approaches neglect.

In this chapter I propose to construct a large-scale energy-based model and demonstrate the manner in which the dialogue state tracking process can benefit from the slot dependencies in multiple dialogue domains, in particular the associations among slots across domains. My choice of data is the two popular multiple domain datasets, MultiWOZ 2.0 (Budzianowski et al., 2018) and MultiWOZ 2.1 (Eric et al., 2020), that contain around 8000 dialogues across 7 domains. The existence of slot dependencies in these data has been investigated in Chapter 3, Section 3.3.2.

Constructing a large-scale energy-based dialogue state tracking model requires a significant modification of my single domain energy-based tracker presented in

previous chapters. Therefore I present the modified model in Section 7.2, followed by the presentation of the tracker's performance and a systematic analysis of slot dependencies across dialogue domains in Section 7.4. Thus, in this chapter my contribution focuses on the generalisability of the structured prediction method for dialogue state tracking.

7.2 Large-Scale EBL Dialogue State Tracking

As presented above, my energy-based dialogue state tracking architecture consists of a feature network and an energy network implemented with deep learning techniques. In order to accommodate the large number of slots and values in multiple domain data both these network structures need modifying. In the following I present both these modifications.

7.2.1 Large-Scale Recurrent Neural Feature Network

In my energy-based learning method, a feature function $F(X)$ is used to transform raw dialogue input into a distributed representation format. Similar to the concept of a feature network presented in previous chapters, I structure the large-scale feature network for multiple domain dialogue data with a combination of embeddings and recurrent neural networks, specifically an LSTM structure (Hochreiter & Schmidhuber, 1997) and a bidirectional LSTM structure (Huang et al., 2015). The architecture of the feature network here, however, must accommodate the large number of dialogue domains, slots and values in the multiple domain data.

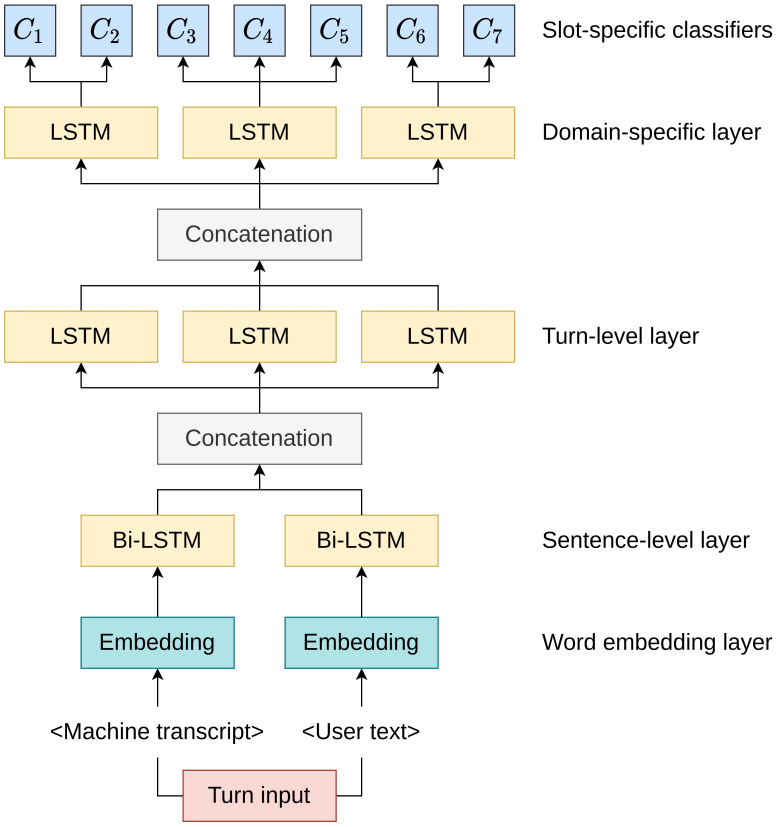

Figure 7.1: Large-Scale Recurrent Neural Feature Network for MultiWOZ data.

In Figure 7.1 I illustrate the structure of the large-scale feature network, followed by the detail of the network elements.

The MultiWOZ dialogue consists of multiple turns, each of which has a machine transcript and user text as input. In order to process this data format, my LSTM-based feature network is constructed with 5 main layers and 2 additional concatenation layers as explained below:

- **Word embedding layer** – A trained from scratch embedding layer transforms raw text input into word vector representations.
- **Sentence embedding layer** – As is common practice I transform sentences into vector representations with a bidirectional LSTM (Bi-LSTM) (Huang

et al., 2015). Here, I use separate Bi-LSTM cells for machine and user transcripts, then concatenate their output to form the dialogue turn input.

- **Dialogue turn layer** – Following my multi-task modelling approach developed for single domain dialogue data featured in previous chapters, I make use of a number of LSTM cells to roll out the dialogue turn-by-turn. The output produced by these LSTM cells are then concatenated to form the information representations of current dialogue turns.

- **Domain-specific layer** – Each domain of MultiWOZ data is assigned an LSTM cell, that specialises the information downstream from the overall dialogue to the domain level.

- **Slot-specific classifier layer** – This is the output layer of the network, that consists of a number of slot-specific classifiers. These classifiers produce the output prediction for the corresponding slots.

Due to the big increase in the number of slots and their values in the multiple domain data, the large-scale feature network is significantly modified in comparison with the single domain feature network in previous chapters. The modification mainly lies in the presence of multiple turn-level LSTM cells and the domain-specific layer. Overall, I can treat the extracted information of either the turn-level layer or the domain-specific layer as the dialogue representations for further use. However, based on my experiments it is observed that the domain-specific layer produces output that is more beneficial for the energy structure. The model

hyper-parameters were carefully optimised with a grid search method during the development phase (see Appendix A).

7.2.2 Large-Scale Deep Neural Energy Network

In the energy-based learning approach, the energy function $E(F(X), Y)$ implicitly captures the dependencies among feature inputs and slot outputs in the system. As presented in Section 5.2.2, this function is based on Belanger & McCallum (2016)'s Structured Prediction Energy Networks architecture. The formulation of the energy function was detailed in Equations 5.1, 5.2 and 5.3.

Although the architecture of the deep learning energy network remains as presented in previous chapters, the scale of this network has been changed radically to accommodate the large number of slots in MultiWOZ datasets. In the multiple dialogue domains, the slot dependencies are accounted for both within a particular domain and across different domains. For example, as reported in the investigation of slot dependencies in the MultiWOZ 2.1 data in Section 3.3.2, the associations are present between two slots *hotel.area* and *hotel.type* of the same domain *hotel*, as well as between two slots *hotel.name* and *restaurant.name* of two different domains *hotel* and *restaurant*.

7.3 Energy-Based Learning Processes

As detailed in previous chapters, the working mechanism of the energy-based learning method consists of two processes, namely the *learning process* and the *inference*

process, with different functionalities. My previous experimental results (see Chapter 5) suggest that these two processes should be performed separately to obtain better performance.

During the *learning process*, the energy function is typically trained to recognise the correct input and output configurations by assigning desired energy values to them. In my approach, the energy function is trained to be a loss estimator for the generated outputs, that is based on the algorithm of Gygli et al. (2017)'s Deep Value Networks. The crucial point of the learning process is to define a suitable objective function to serve the purpose. In my approach, the objective function is constructed with two components: (i) a cross entropy loss function between the predicted and ground truth energies, and (ii) a slot value regularisation term between the generated output and ground truth. The details of this objective function are presented in Section 6.3.

The *inference process*, as detailed in Section 5.3.2, is used to produced predicted dialogue states. The prediction procedure is different from the standard feedforward deep learning modelling, such that the prediction starts with a random output hypothesis and goes through an inference loop to reach the desired output result. This process makes use of the well trained energy function from the learning process, and performs the output updates based on the gradients of the energy value with respect to the outputs.

Table 7.1: Performances of state-of-the-art and the energy-based dialogue state tracking systems on MultiWOZ 2.0 & 2.1 data.

Model	MultiWOZ 2.0	MultiWOZ 2.1
TripPy + SaCLog model (Dai et al., 2021)	-	**0.6061**
TripPy + CoCoAug model (S. Li et al., 2021)	-	0.6053
TripPy + SCoRe model (T. Yu et al., 2021)	-	0.6048
Knowledge-aware graph-enhanced GPT2 model (W. Lin et al., 2021)	**0.5486**	-
Transformer model (Zeng & Nie, 2021)	0.5464	0.5535
DST-Picklist model (J.-G. Zhang et al., 2020)	0.5439	0.5330
Multi-task PPTOD (Y. Su et al., 2022)	0.5389	0.5745
TripPy model (Heck et al., 2020)	-	0.5529
Schema-guided with graph attention model (L. Chen et al., 2020)	0.5117	0.5523
Question-answering model (L. Zhou & Small, 2019)	0.5144	0.5117
Transferable state generator (TRADE) (C.-S. Wu et al., 2019)	0.4862	0.4560
Scalable globally-conditioned encoder (GCE) (Nouri & Hosseini-Asl, 2018)	0.3627	-
Global-locally self-attentive tracker (GLAD) (Zhong et al., 2018)	0.3557	-
This work		
Energy-based system	0.488	0.547
Multi-task feature system	0.349	0.366

7.4 Results & Analysis

The performance of both the multi-task feature system and the energy-based tracker in this work are evaluated with the *accuracy* metric as is common in the DSTC data. The results are reported in Table 7.1 alongside results for a number of state-of-the-art systems for comparison.

The results demonstrate that the energy-based dialogue state tracker yields

competitive results at the time of publication. When accounting for the inter-slot dependencies, the energy-based system improves the multi-task feature system dialogue state tracking results by large margins, in detail 13.9% for the MultiWOZ 2.0 and 18.1% for the MultiWOZ 2.1 data. There are at least two reasons for this large improvement:

- High quality features are extracted from dialogue data with a hierarchical recurrent neural feature network. As the input features are extracted from domain-specific neural cells, the features contain both dialogue information as well as domain information up to the current turn.

- The relationships among dialogue slots are taken into account for the prediction; hence more information is available for the classification of each slot than would be in a straightforward deep learning classification method.

The results also demonstrate that state-of-the-art systems currently employ a very wide variety of modelling techniques, wherein only a number of works focuses on the addition of a mechanism to guide final labelling. The state-of-the-art methods can be split into three groups based on their algorithms and modelling structures:

- In the first group, the TripPy model (Heck et al., 2020) and its derivations (Dai et al., 2021; S. Li et al., 2021; T. Yu et al., 2021), that achieve the highest accuracy in MultiWOZ 2.1 data, are based on a span-prediction and a number of additional mechanisms. The systems in this group are similar to

the energy-based model in the point of the algorithm concept, namely that pretraining a base dialogue state tracking system and applying additional techniques can boost the performance.

- The second group includes a number of transformer-based models such as the knowledge-aware graph-enhanced GPT2 model (W. Lin et al., 2021), the transformer model (Zeng & Nie, 2021), the scalable globally-conditioned encoder (Nouri & Hosseini-Asl, 2018), and the global-locally self-attentive tracker (Zhong et al., 2018). Thanks to the attention mechanism, these models can account for the slot dependencies among latent variables and achieve better prediction results.

- In the third group are the systems with additional knowledge graphs such as the schema-guided with graph attention model (L. Chen et al., 2020) and the question-answering model L. Zhou & Small (2019).

However, all of these, with the exception being the DST-picklist model (J.-G. Zhang et al., 2020), do not explicitly look at the slot dependencies as potentially useful factors of dialogue states. J.-G. Zhang et al. (2020)'s DST-picklist approach considers the slot relationships in a manual manner, that is different from my energy-based method. That being said, the practical use of the energy-based learning method lies in its ability to consider dialogue slot dependencies as extra factors. Since the energy-based network is developed separately from the feature network, it is possible to apply the energy-based method to state-of-the-art models to investigate the effectiveness of slot dependencies in different situations.

It is also observed that there exist differences in performance across MultiWOZ 2.0 and 2.1 datasets in all systems. Generally, dialogue state tracking systems tend to perform better on the MultiWOZ 2.1 data as it is better annotated. However, not all systems yield better results in MultiWOZ 2.1 than in MultiWOZ 2.0, for example models such as the DST-picklist model (J.-G. Zhang et al., 2020), the transferable model (TRADE) (C.-S. Wu et al., 2019) and the question-answering model (L. Zhou & Small, 2019) perform better with the original noisy data. In contrast, other state-of-the-art systems and my energy-based tracker perform better with cleaner data (MultiWOZ 2.1) following the common phenomenon in supervised learning.

7.4.1 Analysis of Slot Dependencies

Although the above results demonstrated that the energy-based model outperforms the multi-task feature network by a large accuracy margin, it is arguable that the accuracy metric itself does not verify the ability to capture slot dependencies. It is necessary to study how the energy-based learning method performs in this matter. As presented in previous chapters, an analysis of slot dependencies in the test data and the system's predicted output should be conducted. Following the statistical method to study pairwise dependencies between dialogue slots presented in Section 3.5, the Pearson's chi square tests were performed and the Cramer's V coefficient was used to measure the dependency strength. The results of the slot dependencies analysis between a number of slots is presented in Table 7.2 with

Table 7.2: Analysis of slot dependencies in the MultiWOZ 2.1 testset, and the predicted dialogue states of the energy-based model and the multi-task feature network.

Model	Domain	Slot	attraction area	hotel area	restaurant area
Test label	hotel	price range	0.200	0.225	0.214
	restaurant	price range	0.236	0.315	0.411
Energy-based model	hotel	price range	0.182	0.236	0.256
	restaurant	price range	0.173	0.336	0.419
Multi-task feature system	hotel	price range	0.291	0.147	0.287
	restaurant	price range	0.194	0.232	0.213

respect to test slot labels, dialogue states produced by the energy-based tracker and the multi-task feature network.

It is observed in the analysis results that the energy-based model performs more consistently in capturing the slot dependencies in the MultiWOZ data than the multi-task feature network, a feedforward deep learning structure. The captured dependencies are demonstrated by the fact that the margins of Cramer's V coefficient between the energy-based tracker and the test labels are smaller than the margins between the multi-task feature network and the test data. Here, stronger associations do not necessarily mean better performance, as the goal of my energy-based method is to capture valid associations. In Table 7.2, there is, however, one exception to this trend, namely, for the *attraction-area* and *restaurant-price range* slots where the multi-task feature system produced associations closer to the test label case as compared with the energy-based model.

7.4.2 Analysis of Slot Value Constraint Rules

The MultiWOZ datasets follow the common rules of task-oriented dialogue systems that the dialogue state at any turn assigns only one value for each slot in the domain. For example if the value *chinese* is assigned for the slot *food* in the domain *restaurant*, none of the other values can be assigned for the same slot at the time. In general, feedforward deep learning models can avoid breaking this rule by applying the *softmax* activation function at the output layer of all the slot-specific classifiers. However, the energy-based tracking method approaches this task as a multi-label classification task, that does not guarantee the strict following of this rule.

In the previous chapter, a value regularisation approach was proposed to overcome the challenge for the energy-based model (Section 6.3). This approach is also applied in the multiple dialogue domain setting. Hence, an analysis is conducted to determine to what extent the multi-domain energy-based tracker follows this slot value constraint rules. In this analysis, the energy-based model is trained with and without the value regularisation term, then evaluated on the MultiWOZ data. When trained without the value regularisation, the objective function is the standard case of energy-based learning approach (Equation 5.5). Meanwhile, when the value regularisation is included, the objective function has the formula in Equation 6.6. The evaluation is then conducted with different value thresholds, such that a value for a slot is activated for the belief state if its predicted probability exceeds the given threshold. The slot value constraint analysis is presented in

Table 7.3: Analysis of the impact of value regularisation on the energy-based dialogue state tracking on the MultiWOZ 2.0 & 2.1 data.

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 7.3. 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* denote the presence/absence of the value regularisation in the learning process.

The analysis result demonstrates that the energy-based method performs better when the value regularisation is included in the learning process than when the value regularisation is excluded. This observation is consistent with different belief score thresholds. It can be concluded that the value regularisation term truly helps guide the system's prediction behaviour towards the requirement of the task-oriented domains. And the impact of value regularisation on dialogue state tracking is systematic for both single and multiple dialogue domains.

7.4.3 Analysis of Error Distributions

In the previous sections the overall performance of the proposed energy-based dialogue state tracking systems on the MultiWOZ 2.0 & 2.1 data is reported at around 50% accuracy; thus the proportion of errors is still at a high level. Therefore, it is useful to conduct an error analysis to shed light on the limitations seen in cur-

Table 7.4: Error distributions of the energy-based dialogue state tracker on the MultiWOZ 2.0 & 2.1 data.

Dataset	Label	#Turns	Error distributions (%)		
			MA	EA	FA
MultiWOZ 2.0	+Reg	61660	17.4	29.5	53.1
	–Reg	71767	39.2	21.8	39.0
MultiWOZ 2.1	+Reg	54052	20.7	26.9	52.4
	–Reg	58708	38.1	23.0	38.9

rent approaches. To conduct this analysis I broadly follow the dialogue error type analysis presented by Smith (2014); to the best of my knowledge this is the only example of a comparative error analysis of dialogue state trackers that focuses on the distributions of different error types in dialogue states. Namely, the dialogue state errors are classified into three types: missing attributes (MA), extraneous attributes (EA), and false attributes (FA) (for detail see Section 6.4.5).

The analysis results are reported in Table 7.4. The threshold for the value assignment is 0.5. *+Reg/–Reg* means the value regularisation is included/excluded in the learning process.

The analysis results demonstrate a big change in error types in the predictions produced by the energy-based model when trained with and without the value regularisation term in both MultiWOZ 2.0 & 2.1 datasets. When the value regularisation is excluded, the errors are distributed into three types more evenly. On the other hand, when the value regularisation is present, the dominant type of errors is *False attributes*, more than 50%. It explains that the value regularisation indeed works well in assigning values to dialogue slots and ensuring the slot value constraint rules are followed.

As mentioned in the error distribution analysis in the previous chapter when conducting this analysis on single domain dialogue datasets, Smith (2014) concluded the alignment between the error distributions and the difficulty of the dialogue slots in the system. However, in my error analysis both in single and multiple dialogue domains, this alignment is not an issue as I approach the analysis from another perspective. Namely, the analysis is used to determine how effective the value regularisation approach is in the multi-domain dialogue data. The finding supports the effectiveness of this approach.

7.5 Summary

In this chapter, the energy-based learning method was applied to solving the dialogue state tracking task in a large-scale dialogue domain. The overall results demonstrate that the energy-based method is an effective approach for the task in the multiple domain setting. The energy-based system is capable of capturing the slot dependencies not only of the same domain, but across domains as well. My series of analyses on the energy-based system's performance shows that the structured prediction method can also follow the strict slot value constraint rules in the multiple domain setting, despite it being a multi-label classification method.

There exists one limitation in the generalisability of the energy-based learning method that lies in the increase of the computational cost when scaling the EBL model from single dialogue domain to multiple dialogue domains. In details, the development of the EBL dialogue state tracker requires a larger deep learning

architecture, mainly in the feature network's structure, to handle the number of dialogue slots across domains. Here in this chapter, the feature network was developed with an additional domain-specific recurrent neural layer. Meanwhile the number of slot-specific recurrent neural units and classifiers for multiple domains is much higher than in the case of single domain. Approximately, the multiple domain feature network has 10 times higher number of parameters. Furthermore, in practice the multiple domain feature network requires longer training time than the single domain feature network. While they both need the similar number of epochs to pretrain, the training time of one epoch of the multiple domain feature network is approximately 5 times longer than the single domain feature network.

On the other hand, the training of the energy network in multiple domains is not much different from in single domain, despite it was also developed with more parameters. The explanation for this lies in the overall simplicity of the deep energy network, wherein mainly feedforward neural layers were used.

Overall, this chapter finds that the energy-based learning method has a good generalisation property, that lets it be applied in a multiple dialogue domain setting. That is a promising finding for further development and application of the energy-based method in an open domain dialogue system, as well as investigating different type of variable dependencies such as emotions or personality in the conversation.

Chapter 8

Conclusion

This dissertation has presented a detailed study of the slot dependencies in a number of dialogue domains as well as the effectiveness of applying energy-based learning, a structured prediction method, to the dialogue state tracking task.

The findings with regard to dialogue slot dependencies is important because the existence of these dependencies supports the hypothesis of structural properties in dialogue. Such knowledge motivates further research on integrating dialogue slot dependencies into dialogue processing.

Since slots in dialogue states are not assumed to be independent of each other, there exists a strong motivation to apply structured prediction approaches to the dialogue state tracking process. In this work the energy-based learning method was chosen due to its notable effectiveness at capturing slot dependencies and performing structured predictions. Implementing energy-based dialogue state tracking systems is a novel modelling approach.

Furthermore, task-oriented dialogue domains have specific requirements with respect to dialogue states such as slot value constraint rules. A simple multi-label classification does not satisfy this strictness. However, incorporating a value regularisation term into the learning process of our energy-based model was proven to be effective. On the one hand, it helps improve the overall performance of our model. On the other hand, my analysis indicates that this value regularisation also guides the prediction process to satisfy slot value constraints in dialogue domains.

In this chapter, I summarise the contributions of my work in Section 8.1 and present a number of future work directions in Section 8.2.

8.1 Summary of Contributions

In this section the contributions made within this dissertation are summarised by key chapters:

- **Investigating inter-slot dependencies in dialogue context** – A number of task-oriented dialogue corpora which account for both single and multiple domains were identified for the research. Through a statistical analysis, the dependencies between dialogue slots were detected and measured. In the single dialogue domain datasets, the dialogue slot dependencies were found to be consistently strong. Meanwhile, in the multiple dialogue domain datasets, the slot dependencies were detected not only within a particular domain, but also across domains. They vary from a weak association to a very strong relationship. These findings provided a strong motivation for

further work to be done.

- **Harnessing task domain structure with multi-task learning** – Multi-task learning was demonstrated to be an appropriate approach for dialogue state tracking tasks where associations between dialogue state components (subtasks) are taken into account. When applied correctly to the task, the multi-task learning method produced competitive results. The novelty of the multi-task system lies in its architecture where a shared recurrent neural layer was introduced at an early stage to handle the relationships detected in training signals, while the system still maintained a level of independence between the dialogue state components with the use of task-specific recurrent neural cells and output classifiers. The limitation of the multi-task learning approach, however, is that it could not capture the slot dependencies in an explicit manner, despite the evidence that they were present in dialogue data.

- **Capturing slot dependencies with energy-based learning** – This chapter introduced a structured prediction method, namely energy-based learning, to capture the dialogue slot dependencies. In the results, the energy-based dialogue state tracking system outperformed the multi-task model. This result indicated that the slot dependencies when captured had a positive impact on the dialogue state prediction process. Furthermore, the improvement was significant enough to suggest that the energy-based network can be further applied to the state-of-the-art systems to boost state-of-the-art performance.

- **Enforcing dialogue state constraints in energy-based prediction** – In this chapter a number of improvements was proposed for the energy-based method presented in the previous chapter. The reason for these modifications was that the energy-base learning method itself treats the dialogue state tracking task as a multi-label classification problem and does not strictly follow the slot value constraint rules seen in task-oriented dialogues. The improvements included an optimisation of the quality measurement for the output, and a value regularisation for constraint integration. The results demonstrated that not only was the overall dialogue state tracking performance improved, but the system's behaviour followed the constraint rules to produce satisfying dialogue states as well. The analyses conducted in this chapter sets a precedent for how to analyse the performance of dialogue state tracking systems.

- **Generalisability of energy-based learning to multiple dialogue domains** – This chapter demonstrated that the studied energy-based method is generalisable. Specifically, this method was applied to solve the dialogue state tracking tasks in a number of large-scale multiple domain dialogue datasets. The experimental results indicated that the dialogue state tracking can benefit from modelling the slot dependencies detected between slots. The analyses on the energy-based system's performance revealed that the slot value constraint rules for the task-oriented dialogue system were followed. The generalisability of the energy-based method opens promising

future research directions for investigating and making use of the structural property of dialogue in broader dialogue systems.

8.2 Directions for Future Work

There are many possible directions for further study of the structural property of natural language in conversational activities and to investigate the technology to make use of these properties. In this section, I outline a number of potential research directions for the work done in this dissertation.

First, I would like to propose directions for future work that directly address the limitation of this research presented in the previous chapters:

- **Beyond DSTC's scope** – The DSTC's scope contains a lot of restrictions on dialogue state tracking such as data imbalance between dialogue slots and slot value constraint rules. The imbalance of dialogue slots lies in the difference between high- and low-frequency slot types, where low-frequency slots are often excluded from experiments following the common practice in the community. The potential solutions for this issue are perhaps zero-shot learning and/or data enrichment, that is worth investigating. On the other hand, dialogue state tracking systems should have the flexibility to work with or without the slot value constraint rules.

- **Improving energy-based dialogue state tracking** – As discussed in Chapters 6 and 7, the energy-based models are capable of following the slot

value constraint rules of task-oriented dialogue states by producing more *False Attribute* errors over the other two types. However, this error type contributes to incorrect predictions that limit the model performance in different dialogue domains. Although it is good to narrow down the error types in dialogue state predictions, more research to tackle the *False Attribute* error type are needed for further improvement.

- **Challenge of the computational cost** – It was confirmed that the energy-based learning approach possesses a good generalisability in Chapter 7. However, there exists a limitation in the generalisability that the energy-based model requires more resources and time to train when scaling up from a single domain to multiple domains. It presents a big challenge in terms of computational cost, especially in the case of even more domains being included in the tasks. This problem should be tackled by research in optimisation of energy-based learning for multiple-domain and open-domain task-oriented dialogue systems, and furthermore general purpose dialogue systems.

Second, there are a number of issues that can be raised from the technical perspective:

- **Improving the state of the art** – The energy-based network is a modular component of dialogue systems, therefore it can be plugged into different systems. Throughout the research presented in Chapters 5, 6 and 7, I demonstrated that the results of the structured prediction methods were competitive at the time of publication, they were not yet the state of the

art. One direction would be to apply the energy-based method in this work to the state-of-the-art systems to investigate the usefulness of dialogue slot dependencies in those systems. For example, the TripPy model (Heck et al., 2020) and its derivations (Dai et al., 2021; S. Li et al., 2021; T. Yu et al., 2021) explored the different computational techniques for dialogue state tracking, but did not explicitly study slot dependencies within dialogue states.

- **Improving dialogue representation learning** – My dialogue representations are learned through a deep learning architecture, namely the multi-task feature network in this dissertation. I presented a novel multi-task approach that accounted for the dependencies between dialogue state components (subtasks) while maintaining the independence between those components to a certain extent. My approach was based on a feedforward hierarchical LSTM architecture. However, since the development of transformer systems (Devlin et al., 2019; Radford et al., 2019; Lewis et al., 2020; Raffel et al., 2020), most of the state of the art are based on these transformers. Thus, I believe that changing from a feedforward LSTM architecture to a transformer-based network would benefit the dialogue representation learning process, that in turn would be useful for the structured dialogue state tracking.

- **Generalisability beyond dialogue state** – Many task-oriented dialogue systems are very specific in recognising user intents. For example, the Frames system stores all the user intents throughout the conversation history in a

collections of frames in order to retrieve past intents faster (Asri et al., 2017; Schulz et al., 2017). In this setting, detecting and studying the slot dependencies inside the system becomes challenging, in a way that the dependencies may exist above the state level. Further research is required if one wants to study the structural properties in Frames-like domains.

- **Expansion to multimodal dialogue systems** – Variable dependencies are not limited only within a single modality of dialogues, for example gestural alignment in dialogues (Khosrobeigi et al., 2022; Karpiński et al., 2018) or alignment in multimodal interaction (Rasenberg et al., 2020). Although in this work the investigation was conducted either on spoken or chat dialogues, I believe that multimodal dialogue systems can benefit from this study as well. That being said, the dependencies between variables of different modalities might exist and have a huge impact on the system's performance if properly used.

Meanwhile, from a linguistic and cognitive perspective, this work also leads to some other interesting pieces of future work:

- **Generalisability to general purpose conversation** – The study of this work was conducted mainly for task-oriented dialogue systems with predefined domains, while the conversation activities between humans are not bound to specific tasks. Today many modern dialogue systems have widened to the open domain direction with more ability of being general purpose conversational agents (Hardy et al., 2021; K.-H. Liang et al., 2021). In these

conversational activities, the structural properties of language, and in particular dialogue, still remain. That being said, studying these properties in an open setting is challenging, but has the potential to have a positive impact. There are a number of directions to study further, for example topic coreference in dialogue (Dobnik et al., 2022; Z. Liu et al., 2021; T. Zhao & Kawahara, 2021) and user relations in multi-party dialogue systems (Inoue et al., 2021; Si et al., 2021).

- **Enhancing the cognitive aspects in dialogue** – Recently, Zachrau (2022) has called for more studies to focus on relationality in dialogue such that conversational entities should not be studied separately and individually, but in the nexus from which they were taken. This call has gone beyond my study of dialogue slot dependencies in this work. There is a wide range of interpretations of conversational entities such as user emotions (Marques, 2022; Ishii et al., 2021), communication styles (Ward, 2021; Hewitt & Beaver, 2020), personalities (Miyazaki et al., 2021), behaviours (I. Gupta et al., 2021), and KoS model with recent incorporation of laughter, emotions, and other interactive and social information (Ginzburg, 2012; Maraev et al., 2018). Therefore, another long-term goal after this work is to apply the structured learning approach in tracking different cognitive aspects of the conversations as well as user intents.

Finally, my study of dialogue slot dependencies in task-oriented dialogue systems has demonstrated the impact of incorporating structural properties of natural

language into dialogue processing. Not only does it boost the overall performances in terms of accuracy of deep learning systems, but it also captures the slot dependencies with evidence via numerous analyses. Furthermore, although the notion of dialogue state investigated in this research is still quite reduced compared to one attempting to model more complex internal and external states of dialogue participants or more complex tasks, I believe that my work has a wide range of possible applications for further study in broader research. In the future the hope is to further study the linguistic and cognitive aspects of the conversational artificial intelligence field using my research as the starting point.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *Tensorflow: Large-scale machine learning on heterogeneous distributed systems.* Retrieved from `https://www.tensorflow.org/`

Al-Ajmi, A.-H., & Al-Twairesh, N. (2021). Building an arabic flight booking dialogue system using a hybrid rule-based and data driven approach. *IEEE Access*, *9*, 7043-7053. doi: 10.1109/ACCESS.2021.3049732

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Maier, E., ... Siegel, M. (1998). *Dialogue acts in verbmobil-2. second edition.*

Allen, J., & Core, M. (1997). *Damsl: Dialogue act markup in several layers.*

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Dohery, G., Garrod, S., ... Weinert, R. (1991). The hcrc map task corpus. *Language and Speech*, *34*, 351-366.

Asri, L. E., Schulz, H., Sharma, S., Zumer, J., Harris, J., Fine, E., ... Suleman, K. (2017). Frames: A corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the sigdial 2017 conference* (p. 207-219).

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd international conference on learning representations.*

Balaraman, V., & Magnini, B. (2019). Scalable neural dialogue state tracking. In *Proceedings of 2019 ieee automatic speech recognition and understanding workshop (asru).*

Balaraman, V., & Magnini, B. (2020). Domain-aware dialogue state tracker for multi-domain dialogue systems. In *Proceedings of the 8th dialog system technology challenge (dstc8), aaai-20 workshop* (Vol. 1). Retrieved from `http://arxiv.org/abs/2001.07526`

Balaraman, V., Sheikhalishahi, S., & Magnini, B. (2021). Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey. In *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue* (p. 239-251). doi: 10.18653/v1/2020.coling-main.42

Belanger, D., & McCallum, A. (2016). Structured prediction energy networks. In *Proceedings of the 33rd international conference on machine learning* (Vol. 48).

Belanger, D., Yang, B., & McCallum, A. (2017). End-to-end learning for structured prediction energy networks. In *Proceedings of the 34th international conference on machine learning.*

Bohus, D., & Rudnicky, A. (2006). A "k hypotheses + other" belief updating model. In *Proceedings of the aaai workshop on statistical and empirical approaches for spoken dialogue systems.*

Bowden, K. K., Wu, J., Cui, W., Juraska, J., Harrison, V., Schwarzmann, B., ... Walker, M. (2018). Slugbot: Developing a computational model andframework of a novel dialogue genre. In *Proceedings of the 2018 amazon alexa prize.* Retrieved from `http://arxiv.org/abs/1907.10658%0Ahttp://dx.doi.org/10.13140/RG.2.2.33543.96166` doi: 10.13140/RG.2.2.33543.96166

Budzianowski, P., Ultes, S., Su, P.-H., Mrksic, N., Wen, T.-H., Casanueva, I., ... Gasic, M. (2017). Sub-domain modelling for dialogue management with hierarchical reinforcement learning. In *Proceedings of the sigdial 2017 conference* (p. 86-92).

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., & Gašić, M. (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.*

Bunt, H. (1994). Context and dialogue control. *Think Quarterly*, *3*.

Bunt, H. (2009). The dit++ taxonomy for functional dialogue markup. In *Proceedings of edaml/aamas workshop "towards a standard markup language for embodied dialogue acts"* (p. 13-24).

Bunt, H., Petukhova, V., & Fang, A. (2017). Revisiting the iso standard for dialogue act annotation. In *Proceedings of the 13th joint iso-acl workshop on interoperable semantic annotation (isa-13).*

Bunt, H., Petukhova, V., Gilmartin, E., Pelachaud, C., Fang, A., Keizer, S., & Prévot, L. (2020). The iso standard for dialogue act annotation, second edition. In *Proceedings of the 12th conference on language resources and evaluation (lrec 2020)* (p. 549-558).

Byrne, B., Krishnamoorthi, K., Ganesh, S., & Kale, M. (2021). Tickettalk: Toward human-level performance with end-to-end, transaction-based dialog systems. In *Proceedings of the 59th annual meeting ofthe association for computational linguistics and the 11th international joint conference on natural language processing* (p. 671-680). doi: 10.18653/v1/2021.acl-long.55

Caruana, R. (1997). Multi-task learning. *Machine Learning*, *28*, 41-75.

Chao, G.-L., & Lane, I. (2019). Bert-dst: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. In *Proceedings of the annual conference of the international speech communication association, interspeech 2019* (p. 1468-1472). doi: 10.21437/Interspeech.2019-1355

Chen, F. (2020). Tartan : A two-tiered dialog framework for multi-domain social chitchat. In *3rd proceedings of alexa prize.*

Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, *19*, 25-35. doi: 10.1145/3166054.3166058

Chen, L., Lv, B., Wang, C., Zhu, S., Tan, B., & Yu, K. (2020). Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *Association for the advancement of artificial intelligence.*

Chen, Q., Zhuo, Z., & Wang, W. (2019). Bert for joint intent classification and slot filling. *ArXiv*. Retrieved from `http://arxiv.org/abs/1902.10909`

Chen, S., Zhang, Y., & Yang, Q. (2021). Multi-task learning in natural language processing: An overview. *ArXiv*. Retrieved from `http://arxiv.org/abs/2109.09138`

Clark, S., Rimell, L., Polajnar, T., & Maillard, J. (2016). The categorial framework for compositional distributional semantics. *Cambridge Computer Lab Tech Report*.

Coecke, B., Sadrzadeh, M., & Clark, S. (2010, 3). Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, *36*, 345-384. Retrieved from `http://arxiv.org/abs/1003.4394`

Cristofori, L., Hromei, C. D., Luzio, F. S. D., Tamantini, C., Cordella, F., Croce, D., ... Basili, R. (2021). Heal9000: An intelligent rehabilitation robot. In *Proceedings of aixia 2021 smartercare workshop* (Vol. 3060, p. 29-41).

Dai, Y., Li, H., Li, Y., Sun, J., Huang, F., Si, L., ... Group, A. (2021). Preview, attend and review: Schema-aware curriculum learning for multi-domain dialog state tracking. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (p. 879-885).

Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, *54*, 755-810. Retrieved from `https://doi.org/10.1007/s10462-020-09866-x` doi: 10.1007/s10462-020-09866-x

Dernoncourt, F., Lee, J. Y., Bui, T. H., & Bui, H. H. (2016). Robust dialog state tracking for large ontologies. In *Proceedings of the international workshop on spoken dialogue systems, iwsds.*

Dev, C., Biyani, N., Suthar, N. P., Kumar, P., & Agarwal, P. (2021). Structured prediction in nlp - a survey. *ArXiv*. Retrieved from `http://arxiv.org/abs/2110.02057`

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., ... Morency, L.-P. (2014). Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 13th international conference on autonomous agents and multiagent systems* (p. 1061-1068).

DeVault, D., & Stone, M. (2007). Managing ambiguities across utterances in dialogue. In *Proceedings of the workshop on the semantics and pragmatics of*

dialogue (decalog).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacl-hlt 2019* (p. 4171-4186). doi: arXiv:1810.04805v2

Dobnik, S., Ilinykh, N., & Karimi, A. (2022). What to refer to and when? reference and re-reference in two language-and-vision tasks. In *Proceedings of the 26th workshop on the semantics and pragmatics of dialogue* (p. 146-159).

Dufour, R., Morchid, M., & Parcollet, T. (2016). Tracking dialog states using an author-topic based representation. In *Proceedings of 2016 ieee workshop on spoken language technology* (p. 544-551). doi: 10.1109/SLT.2016.7846316

Eric, M., Goel, R., Paul, S., Kumar, A., Sethi, A., Goyal, A. K., ... Hakkani-Tür, D. (2020). Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th conference on language resources and evaluation (lrec 2020)* (p. 422-428).

Fagarasan, L., Vecchi, E. M., & Clark, S. (2015). From distributional semantics to feature norms: grounding semantic models in human perceptual data. In *Proceedings of the 11th international conference on computational semantics (iwcs 2015)* (p. 52-57). Association for Computational Linguistics. Retrieved from `https://code.google.com/p/word2vec/`

Feng, Y., Wang, Y., & Li, H. (2021). A sequence-to-sequence approach to dialogue state tracking. In *Proceedings of the 59th annual meeting of the association for*

computational linguistics and the 11th international joint conference on natural language processing (p. 1714-1725).

Field, A. (2017). *Discovering statistics using ibm spss statistics* (5th ed.). SAGE.

Fix, J., & Frezza-Buet, H. (2015). Yarbus: Yet another rule based belief update system. *ArXiv*.

Gabriel, R., Liu, Y., Gottardi, A., Eric, M., Khatri, A., Chadha, A., ... Hakkani-Tür, D. (2020). Further advances in open domain dialog systems in the third alexa prize socialbot grand challenge. In *3rd proceedings of alexa prize.*

Gao, S., Sethi, A., Agarwal, S., Chung, T., & Hakkani-tur, D. (2019). Dialog state tracking: A neural reading comprehension approach. In *Proceedings of the sigdial 2019 conference* (p. 264-273).

Garvin, P. L. (1976). The structural properties of language. *Revista de Letras*, *18*, 81-100.

Gasic, M., & Young, S. (2011). Effective handling of dialogue state in the hidden information state pomdp-based dialogue manager. *ACM Transactions on Speech and Language Processing*, *7*. doi: 10.1016/B978-0-12-396963-7.00021-0

Ginzburg, J. (2012). *The interactive stance: Meaning for conversation.* Oxford University Press.

Goel, R., Paul, S., & Hakkani-Tur, D. (2019). Hyst: A hybrid approach for flexible and accurate dialogue state tracking. In *Proceedings of the interspeech 2019 conference.*

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12*, 175-204.

Gulyaev, P., Elistratova, E., Konovalov, V., Kuratov, Y., Pugachev, L., & Burtsev, M. (2020). Goal-oriented multi-task bert-based dialogue state tracker. In *Proceedings of the 8th dialog system technology challenge (dstc8), aaai-20 workshop.* Retrieved from `http://arxiv.org/abs/2002.02450`

Gupta, A., & Durrett, G. (2019). Tracking discrete and continuous entity state for process understanding. In *Proceedings of the 3rd workshop on structured prediction for nlp* (p. 7-12).

Gupta, I., Eugenio, B. D., Ziebart, B. D., Liu, B., Gerber, B. S., & Sharp, L. K. (2021). Summarizing behavioral change goals from sms exchanges to support health coaches. In *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue* (p. 276-289).

Gygli, M., Norouzi, M., & Angelova, A. (2017). Deep value networks learn to evaluate and iteratively refine structured outputs. In *Proceedings of the 34th international conference on machine learning.*

Hardy, A., Paranjape, A., & Manning, C. D. (2021). Effective social chatbot strategies for increasing user initiative. In *Proceedings of the 22nd annual meet-*

ing of the special interest group on discourse and dialogue (p. 99-110). Retrieved from https://aclanthology.org/2021.sigdial-1.11

Harrison, V., Juraska, J., Cui, W., Reed, L., Bowden, K. K., Wu, J., ... Walker, M. (2020). Athena: Constructing dialogues dynamically with discourse constraints. In *3rd proceedings of alexa prize.* Retrieved from http://arxiv.org/abs/2011.10683

Heck, M., van Niekerk, C., Lubis, N., Geishauser, C., Lin, H.-C., Moresi, M., & Gašić, M. (2020). Trippy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the sigdial 2020 conference* (p. 35-44).

Henderson, M. (2015). Machine learning for dialog state tracking: A review. In *Proceedings of the first international workshop on machine learning in spoken language processing.*

Henderson, M., Thomson, B., & Williams, J. (2013). *Dialog state tracking challenge 2 3.* Retrieved from http://camdial.org/~mh521/dstc/

Henderson, M., Thomson, B., & Williams, J. D. (2014a). The second dialog state tracking challenge. In *Proceedings of the sigdial 2014 conference* (p. 263-272).

Henderson, M., Thomson, B., & Williams, J. D. (2014b). The third dialog state tracking challenge. In *Proceedings of 2014 ieee workshop on spoken language technology* (p. 324-329).

Henderson, M., Thomson, B., & Young, S. (2013). Deep neural network approach for the dialog state tracking challenge. In *Proceedings of the sigdial 2013 conference* (p. 467-471).

Henderson, M., Thomson, B., & Young, S. (2014a). Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. In *Proceedings of 2014 ieee workshop on spoken language technology* (p. 360-365).

Henderson, M., Thomson, B., & Young, S. (2014b). Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the sigdial 2014 conference* (p. 292-299).

Hewitt, T., & Beaver, I. (2020). A case study of user communication styles with customer service agents versus intelligent virtual agents. In *Proceedings of the sigdial 2020 conference* (p. 79-85).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735-1780. doi: 10.1162/neco.1997.9.8.1735

Hori, T., Wang, H., Hori, C., Watanabe, S., Harsham, B., Roux, J. L., ... Aikawa, T. (2016). Dialog state tracking with attention-based sequence-to-sequence learning. In *Proceedings of 2016 ieee workshop on spoken language technology* (p. 552-558).

Hosseini-Asl, E., McCann, B., Wu, C.-S., Yavuz, S., & Socher, R. (2020). A simple language model for task-oriented dialogue. In *34th conference on neural*

information processing systems (neurips 2020). Retrieved from https://arxiv.org/abs/2005.00796

Huang, Z., Xu, W., & Yu, K. (2015). *Bidirectional lstm-crf models for sequence tagging.* Retrieved from http://arxiv.org/abs/1508.01991

Ihler, A. T., III, J. W. F., & Willsky, A. S. (2005). Loopy belief propagation: Convergence and effects of message errors. *Machine Learning Research*, *6*, 905-936.

Inoue, K., Sakamoto, H., Yamamoto, K., Lala, D., & Kawahara, T. (2021). A multi-party attentive listening robot which stimulates involvement from side participants. In *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue* (p. 261-264). Retrieved from https://aclanthology.org/2021.sigdial-1.28

Ishii, E., Winata, G. I., Cahyawijaya, S., Lala, D., Kawahara, T., & Fung, P. (2021). Erica: An empathetic android companion for covid-19 quarantine. In *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue* (p. 257-260). Retrieved from https://arxiv.org/abs/2106.02325

ISO24617-2. (2020). *Language resource management — semantic annotation framework (semaf) — part 2: Dialogue acts.*

Jagfeld, G., & Vu, N. T. (2017). Encoding word confusion networks with recurrent neural networks for dialog state tracking. In *Proceedings of the 1st workshop on speech-centric natural language processing* (p. 10-17).

Jang, Y., Ham, J., Lee, B.-J., Chang, Y., & eung Kim, K. (2016). Neural dialog state tracker for large ontologies by attention mechanism. In *Proceedings of 2016 ieee workshop on spoken language technology* (p. 531-537). doi: 10.1109/SLT.2016.7846314

Jurafsky, D., & Martin, J. H. (2020). *Chatbots dialogue systems.*

Jurafsky, D., Shriberg, E., & Biasca, D. (1997). *Switchboard swbd-damsl shallow-discourse-function annotation: Coders manual.*

Kadlec, R., Vodolan, M., Libovicky, J., Macek, J., & Kleindienst, J. (2014). Knowledge-based dialog state tracking. In *Proceedings of 2014 ieee workshop on spoken language technology* (p. 348-353). doi: 10.1109/SLT.2014.7078599

Kamp, H., & Reyle, U. (1993). *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory.*

Kamp, H., van Genabith, J., & Reyle, U. (2010). *Discourse representation theory.* doi: 10.1007/978-94-007-0485-5

Karpiński, M., Czoska, A., Jarmołowicz-Nowikow, E., Juszczyk, K., & Klessa, K. (2018). Aspects of gestural alignment in task-oriented dialogues. *Cognitive Studies*. doi: 10.11649/cs.1640

Kelleher, J. D. (2003). *A perceptually based computational framework for the interpretation of spatial language* .

Kelleher, J. D. (2019). *Deep learning*. The MIT Press.

Kelleher, J. D., Costello, F., & van Genabith, J. (2005). Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, *167*, 62-102. doi: 10.1016/j.artint.2005.04.008

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. The MIT Press.

Khatri, C., Hedayatnia, B., Venkatesh, A., Nunn, J., Pan, Y., Liu, Q., ... Prasad, R. (2018). Advancing the state of the art in open domain dialog systems through the alexa prize. In *2nd proceedings of alexa prize.*

Khosrobeigi, Z., Koutsombogera, M., & Vogel, C. (2022). Gesture and part-of-speech alignment in dialogues. In *Proceedings of the 26th workshop on the semantics and pragmatics of dialogue* (p. 172-182).

Kim, D., Choi, J., eung Kim, K., Lee, J., & Sohn, J. (2013). Engineering statistical dialog state trackers: A case study on dstc. In *Proceedings of the sigdial 2013 conference* (p. 462-466).

Kim, K., Lee, C., Jung, S., & Lee, G. G. (2008). A frame-based probabilistic framework for spoken dialog management using dialog examples. In *Proceedings of the 9th sigdial workshop on discourse and dialogue* (p. 120-127). doi: 10.3115/1622064.1622088

Kim, S., & Banchs, R. E. (2014). Sequential labeling for tracking dynamic dialog states. In *Proceedings of the sigdial 2014 conference* (p. 332-336).

Kim, S., D'Haro, L. F., Banchs, R. E., Williams, J. D., & Henderson, M. (2016). The fourth dialog state tracking challenge. In *Proceedings of the international workshop on spoken dialogue systems, iwsds 2016.*

Kim, S., D'Haro, L. F., Banchs, R. E., Williams, J. D., Henderson, M., & Yoshino, K. (2016). The fifth dialog state tracking challenge. In *Proceedings of 2016 ieee workshop on spoken language technology* (p. 511-517).

Kim, S., Yang, S., Kim, G., & Lee, S.-W. (2020). Efficient dialogue state tracking by selectively overwriting memory. In *Proceedings of the 58th annual meeting of the association for computational linguistics (acl).*

Korbmacher, J., & Schiemer, G. (2018). What are structural properties? *Philosophia Mathematica*, *26*, 295-323. doi: 10.1093/philmat/nkx011

Kumar, A., Ku, P., Goyal, A., Metallinou, A., & Hakkani-Tur, D. (2020). Ma-dst: Multi-attention-based scalable dialog state tracking. In *Proceedings of the 34th aaai conference on artificial intelligence (aaai 2020).*

Kumar, J. A. (2021). Educational chatbots for project-based learning: investigating learning outcomes for a team-based design course. *International Journal of Educational Technology in Higher Education*, *18*, 65. doi: 10.1186/s41239-021-00302-w

Kurfali, M., & Ostling, R. (2019). Zero-shot transfer for implicit discourse relation classification. In *Proceedings of the sigdial 2019 conference* (p. 226-231). doi: 10.18653/v1/w19-5927

Kviz, F. J. (1981). Interpreting proportional reduction in error measures as percentage of variation explained. *The Sociological Quarterly*, *22*, 413-420.

Lai, T. M., Tran, Q. H., Bui, T., & Kihara, D. (2020). A simple but effective bert model for dialog state tracking on resource-limited systems. In *2020 ieee international conference on acoustics, speech and signal processing (icassp).* doi: 10.1109/ICASSP40776.2020.9053975

Landragin, F. (2013). *Man-machine dialogue: Design and challenges.* ISTE Ltd and John Wiley Sons, Inc. doi: 10.1002/9781118578681

Larsson, S., & Traum, D. R. (2000). Information state and dialogue management in the trindi dialogue move engine toolkit. *Natural Language Engineering*, *6*, 323-340. doi: 10.1017/S1351324900002539

Le, H., Socher, R., & Hoi, S. C. (2020). Non-autoregressive dialog state tracking. In *Eighth international conference on learning representations (iclr 2020).*

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M. A., & Huang, F. J. (2006). A tutorial on energy-based learning. *Predicting Structured Data*.

LeCun, Y., & Huang, F. J. (2005). Loss functions for discriminative training of energy-based models. In *Proceedings of the 10th international workshop on artificial intelligence and statistics (aistats'05)* (p. 206 - 213).

Lee, B.-J., Lim, W., Kim, D., & Kim, K.-E. (2014). Optimizing generative dialog state tracker via cascading gradient descent. In *Proceedings of the sigdial 2014 conference* (p. 273-281).

Lee, S. (2013). Structured discriminative model for dialog state tracking. In *Proceedings of the sigdial 2013 conference* (p. 442-451).

Lee, S., & Eskenazi, M. (2013). Recipe for building robust spoken dialog state trackers: Dialog state tracking challenge system description. In *Proceedings of the sigdial 2013 conference* (p. 414-422).

Lee, S., & Stent, A. (2016). Task lineages: Dialog state tracking for flexible interaction. In *Proceedings of the sigdial 2016 conference* (p. 11-21).

Lee, S., Zhu, Q., Takanobu, R., Li, X., Zhang, Y., Zhang, Z., ... Gao, J. (2019). Convlab: Multi-domain end-to-end dialog system platform. *ArXiv*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (p. 7871-7880). doi: 10.18653/v1/2020.acl-main.703

Li, B., Zhan, Y., Wei, Z., Huang, S., & Sun, L. (2021). Improved non-autoregressive dialog state tracking model. In *Ccris'21: 2021 2nd international conference on control, robotics and intelligent system* (p. 199-203).

Li, M., & Wu, J. (2016). The msiip system for dialog state tracking challenge 4. In *Proceedings of the international workshop on spoken dialogue systems, iwsds 2016.* doi: 10.1109/SLT.2016.7846313

Li, S., Yavuz, S., Hashimoto, K., Li, J., Niu, T., Rajani, N., ... Xiong, C. (2021). Coco: Controllable counterfactuals for evaluating dialogue state trackers. In *Proceedings of the 9th international conference on learning representations.*

Li, X., Chen, Y.-N., Li, L., Gao, J., & Celikyilmaz, A. (2017). End-to-end task-completion neural dialogue systems. In *Proceedings of the 8th international joint conference on natural language processing* (p. 733-743). Asian Federation of Natural Language Processing.

Liang, K., Chau, A., Li, Y., Lu, X., Yu, D., Zhou, M., ... Yu, Z. (2020). Gunrock 2.0: A user adaptive social conversational system. In *3rd proceedings of alexa prize.* Retrieved from `http://arxiv.org/abs/2011.08906`

Liang, K.-H., Lange, P., Oh, Y. J., Zhang, J., Fukuoka, Y., & Yu, Z. (2021). Evaluation of in-person counseling strategies to develop physical activity chatbot for women. In *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue* (p. 32-44). Retrieved from `http://arxiv.org/abs/2107.10410`

Lin, W., Tseng, B.-H., & Byrne, B. (2021). Knowledge-aware graph-enhanced gpt-2 for dialogue state tracking. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (p. 7871-7881). doi: 10.18653/v1/2021.emnlp-main.620

Lin, Z., Madotto, A., Winata, G. I., & Fung, P. (2020). Mintl: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 conference on empirical methods in natural language processing* (p. 3391-3405). doi: 10.18653/v1/2020.emnlp-main.273

Liu, B., & Lane, I. (2018). End-to-end learning of task-oriented dialogs. In *Proceedings of naacl-hlt 2018: Student research workshop* (p. 67-73). doi: 10.18653/v1/n18-4010

Liu, B., Tur, G., Hakkani-Tur, D., Shah, P., & Heck, L. (2018). Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of naacl-hlt 2018* (p. 2060-2069).

Liu, P., Qiu, X., & Huang, X. (2017). Adversarial multi-task learning for text classification. In *Proceedings of the 55th annual meeting of the association for*

computational linguistics (p. 1-10). doi: 10.18653/v1/P17-1001

Liu, Z., Shi, K., & Chen, N. F. (2021). Coreference-aware dialogue summarization. In *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue* (p. 509-519). Retrieved from `http://arxiv.org/abs/2106.08556`

Ma, Y., & Fosler-Lussier, E. (2014). A discriminative sequence model for dialog state tracking using user goal change detection. In *Proceedings of 2014 ieee workshop on spoken language technology* (p. 318-323).

Ma, Y., Hiraoka, T., & Okazaki, N. (2022). Joint entity and relation extraction based on table labeling using convolutional neural networks. In *Proceedings of the sixth workshop on structured prediction for nlp* (p. 11-21). Retrieved from `https://aclanthology.org/2022.spnlp-1.2`

Ma, Y., Raux, A., Ramachandran, D., & Gupta, R. (2012). Landmark-based location belief tracking in a spoken dialog system. In *Proceedings of the 13th annual meeting of the special interest group on discourse and dialogue* (p. 169-178).

Ma, Y., Zeng, Z., Zhu, D., Li, X., Yang, Y., Yao, X., ... Shen, J. (2020). An end-to-end dialogue state tracking system with machine reading comprehension and wide deep classification. In *Proceedings of the 8th dialog system technology challenge (dstc8), aaai-20 workshop.* Retrieved from `http://arxiv.org/abs/1912.09297`

Maraev, V., Ginzburg, J., Larsson, S., Tian, Y., & Bernardy, J.-P. (2018). Towards kos/ttr-based proof-theoretic dialogue management. In *Proceedings of the 22nd workshop on the semantics and pragmatics of dialogue.*

Marques, R. (2022). Conversation and mood in european portuguese. In *Proceedings of the 26th workshop on the semantics and pragmatics of dialogue* (p. 202-213).

Martins, A. F., Peters, B., Zerva, C., Lyu, C., Correia, G., Treviso, M., ... Mihaylova, T. (2022). Deepspin: Deep structured prediction for natural language processing. In *Proceedings of the 23rd annual conference of the european association for machine translation* (p. 325-326).

Mehri, S., Srinivasan, T., & Eskenazi, M. (2019). Structured fusion networks for dialog. In *Proceedings of the sigdial 2019 conference* (p. 165-177). doi: 10.18653/v1/w19-5921

Mehta, N., Gupta, R., Raux, A., Ramachandran, D., & Krawczyk, S. (2010). Probabilistic ontology trees for belief tracking in dialog systems. In *Proceedings of sigdial 2010: the 11th annual meeting of the special interest group on discourse and dialogue* (p. 37-46).

Meng, H. M., Wai, C., & Pieraccini, R. (2003). The use of belief networks for mixed-initiative dialog modeling. *IEEE Transactions on Speech and Audio Processing*, *11*, 757-773. doi: 10.1109/TSA.2003.814380

Metallinou, A., Bohus, D., & Williams, J. D. (2013). Discriminative state tracking for spoken dialog systems. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (p. 466-475).

Miyazaki, C., Kanno, S., Yoda, M., Ono, J., & Wakaki, H. (2021). Fundamental exploration of evaluation metrics for persona characteristics of text utterances. In *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue* (p. 178-189). Retrieved from `https://aclanthology.org/2021.sigdial-1.19`

Mou, X., Sigouin, B., Steenstra, I., & Su, H. (2020). Multimodal dialogue state tracking by qa approach with data augmentation. In *Proceedings of the 8th dialog system technology challenge (dstc8), aaai-20 workshop.*

Mrksic, N., O'Seaghdha, D., Thomson, B., Gasic, M., Su, P.-H., Vandyke, D., ... Young, S. (2015). Multi-domain dialog state tracking using recurrent neural networks. In *Proceedings of the 53rd annual meeting of the association for computational linguistics* (p. 794-799).

Mrksic, N., O'Seaghdha, D., Wen, T.-H., Thomson, B., & Young, S. (2017). Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th annual meeting of the association for computational linguistics.* doi: 10.18653/v1/P17-1163

Mrkšić, N., & Vulić, I. (2018). Fully statistical neural belief tracking. In *Proceedings of the 56th annual meeting ofthe association for computational linguistics (short*

papers) (p. 108-113). doi: 10.18653/v1/p18-2018

Nakano, M., & Komatani, K. (2020). A framework for building closed-domain chat dialogue systems. *Knowledge-Based Systems*, *204*. Retrieved from `https://doi.org/10.1016/j.knosys.2020.106212` doi: 10.1016/j.knosys.2020.106212

Nouri, E., & Hosseini-Asl, E. (2018). Toward scalable neural dialogue state tracking model. In *Proceedings of the 2nd conversational ai workshop, neurips 2018.*

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, *2*, 100033. Retrieved from `https://doi.org/10.1016/j.caeai.2021.100033` doi: 10.1016/j.caeai.2021.100033

Osogami, T. (2017a). Boltzmann machines and energy-based models. In *Ijcai-17 tutorial on energy-based machine learning.*

Osogami, T. (2017b). Boltzmann machines for time-series. In *Ijcai-17 tutorial on energy-based machine learning.*

Peng, H., Thomson, S., & Smith, N. A. (2017). Deep multitask learning for semantic dependency parsing. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (p. 2037-2048). doi: 10.18653/v1/P17-1186

Perez, J., & Radford, W. (2016). Probabilistic matching for dialog state tracking with limited training data. In *Proceedings of the international workshop on spoken dialogue systems, iwsds 2016.*

Platek, O., Belohlavek, P., Hudecek, V., & Jurcicek, F. (2016). Recurrent neural networks for dialogue state tracking. In *Proceedings of ceur workshop, itat 2016 conference* (Vol. 1649, p. 63-67).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI Blog*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*, 1-67.

Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., ... Pettigrue, A. (2017). Conversational ai: The science behind the alexa prize. In *1st proceedings of alexa prize.*

Rasenberg, M., Özyürek, A., & Dingemanse, M. (2020). Alignment in multimodal interaction: An integrative framework. *Cognitive Science*, *44*. doi: 10.1111/cogs.12911

Rastogi, A., Zang, X., Sunkara, S., Gupta, R., & Khaitan, P. (2020a). Schema-guided dialogue state tracking at dstc8. In *Proceedings of the 8th dialog system technology challenge (dstc8), aaai-20 workshop.*

Rastogi, A., Zang, X., Sunkara, S., Gupta, R., & Khaitan, P. (2020b). Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the thirty-fourth aaai conference on artificial intelligence (aaai-20)* (p. 8689-8696). doi: 10.1609/aaai.v34i05.6394

Raux, A., & Ma, Y. (2011). Efficient probabilistic tracking of user goal and dialog history for spoken dialog systems. In *Proceedings of the 2011 annual conference of the international speech communication association* (p. 801-804).

Rei, M. (2017). Semi-supervised multitask learning for sequence labeling. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (p. 2121-2130). doi: 10.18653/v1/P17-1194

Ren, H., Xu, W., & Yan, Y. (2014a). Markovian discriminative modeling for cross-domain dialog state tracking. In *Proceedings of 2014 ieee workshop on spoken language technology* (p. 342-347).

Ren, H., Xu, W., & Yan, Y. (2014b). Markovian discriminative modeling for dialog state tracking. In *Proceedings of the sigdial 2014 conference* (p. 327-331).

Ren, H., Xu, W., Zhang, Y., & Yan, Y. (2013). Dialog state tracking using conditional random fields. In *Proceedings of the sigdial 2013 conference* (p. 457-461).

Ren, L., Xie, K., Chen, L., & Yu, K. (2018). Towards universal dialogue state tracking. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (p. 2780-2786).

Ross, R. J. (2009). *Situated dialogue systems: Agency spatial meaning in task-oriented dialogue* .

Ross, R. J., & Bateman, J. (2009). Daisie: Information state dialogues for situated systems. In *Proceedings of international conference on text, speech and dialogue, tsd 2009* (p. 379-386). doi: 10.1007/978-3-642-04208-9_52

Roy, N., Pineau, J., & Thrun, S. (2000). Spoken dialog management for robots. In *Proceedings of the association for computational linguistics.*

Ruan, Y.-P., Ling, Z.-H., Gu, J.-C., & Liu, Q. (2020). Fine-tuning bert for schema-guided zero-shot dialogue state tracking. In *Proceedings of the 8th dialog system technology challenge (dstc8), aaai-20 workshop.* Retrieved from `http://arxiv.org/abs/2002.00181`

Ruder, S. (2019). *Neural transfer learning for natural language processing* . Retrieved from `http://ruder.io/thesis/neural_transfer_learning_for_nlp.pdf`

Schatzmann, J. (2008). *Statistical user modeling for dialogue systems* .

Schulz, H., Zumer, J., Asri, L. E., & Sharma, S. (2017). A frame tracking model for memory-enhanced dialogue systems. In *Proceedings of the 2nd workshop on representation learning for nlp* (p. 219-227).

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., & Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the 30th aaai conference on artificial intelligence (aaai-16)* (p. 3776-3783).

Shalyminov, I. (2020). *Data-efficient methods for dialogue systems* .

Shen, T., & Wang, X. (2020). Multi-domain dialogue state tracking with hierarchical task graph. In *Proceedings of the 2020 international joint conference on neural networks (ijcnn).* doi: 10.1109/IJCNN48605.2020.9206790

Shi, H., Ushio, T., Endo, M., Yamagami, K., & Horii, N. (2016a). Convolutional neural networks for multi-topic dialog state tracking. In *Proceedings of the international workshop on spoken dialogue systems, iwsds 2016.*

Shi, H., Ushio, T., Endo, M., Yamagami, K., & Horii, N. (2016b). A multichannel convolutional neural network for cross-language dialog state tracking. In *Proceedings of 2016 ieee workshop on spoken language technology* (p. 559-564). doi: 10.1109/SLT.2016.7846318

Shi, X., Fang, S., & Knight, K. (2020). A bert-based unified span detection framework for schema-guided dialogue state tracking. In *Proceedings of the 8th dialog system technology challenge (dstc8), aaai-20 workshop.*

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., & Carvey, H. (2004). The icsi meeting recorder dialog act (mrda) corpus. In *Proceedings of the 5th sigdial workshop on discourse and dialogue at hlt-naacl 2004* (p. 97-100).

Shu, L., Molino, P., Namazifar, M., Xu, H., Liu, B., Zheng, H., & Tur, G. (2019). Flexibly-structured model for task-oriented dialogues. In *Proceedings of the sigdial 2019 conference* (p. 178-187).

Si, W. M., Ammanabrolu, P., & Riedl, M. O. (2021). Telling stories through multi-user dialogue by modeling character relations. In *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue* (p. 269-275). Retrieved from `http://arxiv.org/abs/2105.15054`

Sieińska, W., Gunson, N., Walsh, C., Dondrup, C., & Lemon, O. (2020). Conversational agents for intelligent buildings. In *Proceedings of the sigdial 2020 conference* (p. 45-48).

Smith, R. W. (2014). Comparative error analysis of dialog state tracking. In *Proceedings of the sigdial 2014 conference* (p. 300-309).

Stratos, K. (2017). Entity identification as multitasking. In *Proceedings of the 2nd workshop on structured prediction for nlp* (p. 7-11). Retrieved from `https://github.com/karlstratos/`

Su, P.-H., Budzianowski, P., Ultes, S., Gasic, M., & Young, S. (2017). Sample-efficient actor-critic reinforcement learning with supervised data for dialogue management. In *Proceedings of the sigdial 2017 conference* (p. 147-157).

Su, Y., Li, M., & Wu, J. (2016). The msiip system for dialog state tracking challenge 5. In *Proceedings of 2016 ieee workshop on spoken language technology* (p. 525-530).

Su, Y., Shu, L., Mansimov, E., Gupta, A., Cai, D., Lai, Y.-A., & Zhang, Y. (2022). Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th annual meeting of the association for computational linguistics (acl).* Retrieved from `http://arxiv.org/abs/2109.14739`

Sun, K., Chen, L., Zhu, S., & Yu, K. (2014a). A generalized rule based tracker for dialogue state tracking. In *Proceedings of 2014 ieee workshop on spoken language technology* (p. 330-335).

Sun, K., Chen, L., Zhu, S., & Yu, K. (2014b). The sjtu system for dialog state tracking challenge 2. In *Proceedings of the sigdial 2014 conference* (p. 318-326).

Sun, K., Moon, S., Crook, P., Roller, S., Silvert, B., Liu, B., ... Cardie, C. (2021). Adding chit-chats to enhance task-oriented dialogues. In *Proceedings of the 2021 conference of the north american chapter ofthe association for computational linguistics: Human language technologies* (p. 1570-1583).

Tanaka, S., Yoshino, K., Sudoh, K., & Nakamura, S. (2021). Arta: Collection and classification of ambiguous requests and thoughtful actions. In *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue* (p. 77-88).

ter Horst, H., & Cimiano, P. (2020). Structured prediction for joint class cardinality and entity property inference in model-complete text comprehension. In *Proceedings of the 4th workshop on structured prediction for nlp* (p. 22-32). Retrieved from `http://psink.techfak.uni-bielefeld.`

Thomson, B. (2009). *Statistical methods for spoken dialogue management* .

Thomson, B., & Young, S. (2010). Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech and Language*, *24*, 562-588. doi: 10.1016/j.csl.2009.07.003

Tian, X., Huang, L., Lin, Y., Bao, S., He, H., Yang, Y., ... Sun, S. (2021). Amendable generation for dialogue state tracking. In *Proceedings of the 3rd workshop on natural language processing for conversational ai* (p. 80-92).

Traum, D. R. (2000). 20 questions on dialogue act taxonomies. *Journal of Semantics*, *17*, 7-30. doi: 10.1093/jos/17.1.7

Traum, D. R., & Larsson, S. (2001). The information state update approach to dialogue modelling. *The Trindi Book. Notes from the ESSLI*, 18-21.

Traum, D. R., & Larsson, S. (2003). The information state approach to dialogue management. *Current and New Directions in Discourse and Dialogue*, *22*, 325-353.

Trinh, A. D. (2017). Dialogue management modelling. In *Proceedings of the 13th workshop on spoken dialogue systems for phds, postdocs new researchers (yrrsds)* (p. 23-24).

Trinh, A. D. (2019). Dialogue state tracking. In *Proceedings of the 15th workshop on spoken dialogue systems for phds, postdocs new researchers (yrrsds)* (p. 18-19).

Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2017). Incremental joint modelling for dialogue state tracking. In *Proceedings of the 21st workshop on the semantics and pragmatics of dialogue (semdial)* (p. 176-177).

Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2018). A multi-task approach to incremental dialogue state tracking. In *Proceedings of the 22nd workshop on the semantics and pragmatics of dialogue (semdial)* (p. 132-145).

Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2019a). Capturing dialogue state variable dependencies with an energy-based neural dialogue state tracker. In *Proceedings of the sigdial 2019 conference* (p. 75-84). Association for Computational Linguistics. Retrieved from `https://www.aclweb.org/anthology/W19-5910` doi: 10.18653/v1/W19-5910

Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2019b). Energy-based modelling for dialogue state tracking. In *Proceedings of the 1st workshop on nlp for conversational ai* (p. 77-86). Association for Computational Linguis-

tics. Retrieved from https://www.aclweb.org/anthology/W19-4109 doi: 10.18653/v1/W19-4109

Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2019c). Investigating variable dependencies in dialogue states. In *Proceedings of the 23rd workshop on the semantics and pragmatics of dialogue (semdial)* (p. 195-197).

Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2020a). Energy-based neural modelling for large-scale multiple domain dialogue state tracking. In *Proceedings of the 4th workshop on structured prediction for nlp* (p. 33-42).

Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2020b). F-measure optimisation and label regularisation for energy-based neural dialogue state tracking models. In *Proceedings of the 29th international conference on artificial neural networks (icann)* (p. 798-810).

Truong, H. P., Parthasarathi, P., & Pineau, J. (2017). Maca: A modular architecture for conversational agents. In *Proceedings of the sigdial 2017 conference* (p. 93-102).

Tseng, B.-H., Budzianowski, P., Wu, Y.-C., & Gašić, M. (2019). Tree-structured semantic encoder with knowledge sharing for domain adaptation in natural language generation. In *Proceedings of the sigdial 2019 conference* (p. 155-164). doi: 10.18653/v1/w19-5920

Tu, L., & Gimpel, K. (2018). Learning approximate inference networks for structured prediction. In *Proceedings of the 6th international conference on learning representations (iclr).*

Tu, L., & Gimpel, K. (2019). Benchmarking approximate inference methods for neural structured prediction. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies* (Vol. 1, p. 3313-3324). doi: 10.18653/v1/n19-1335

Tu, L., Liu, T., & Gimpel, K. (2020). An exploration of arbitrary-order sequence labeling via energy-based inference networks. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (p. 5569-5582). doi: 10.18653/v1/2020.emnlp-main.449

Tu, L., Pang, R. Y., & Gimpel, K. (2020). Improving joint training of inference networks and structured prediction energy networks. In *Proceedings of the 4th workshop on structured prediction for nlp* (p. 62-73).

Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer networks. In *Advances in neural information processing systems 28 (nips 2015).*

Vodolan, M., Kadlec, R., & Kleindienst, J. (2015). Hybrid dialog state tracker. In *Proceedings of the machine learning for slu interaction nips 2015 workshop.*

Vodolan, M., Kadlec, R., & Kleindienst, J. (2017). 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, p. 205-210).

Wang, B., Li, C., Pavlu, V., & Aslam, J. (2017). Regularizing model complexity and label structure for multi-label text classification. In *Proceedings of kdd'17.* doi: 10.1145/nnnnnnn.nnnnnnn

Wang, Z., & Lemon, O. (2013). A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In *Proceedings of the sigdial 2013 conference* (p. 423–432). doi: 10.13140/2.1 .1213.1522

Ward, N. G. (2021). Individual interaction styles: Evidence from a spoken chat corpus. In *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue* (p. 27-31). Retrieved from `https://aclanthology.org/2021.sigdial-1.4`

Ward, N. G., & DeVault, D. (2015). Ten challenges in highly-interactive dialog systems. In *Aaai spring symposium on turn-taking and coordination in human-machine interaction* (p. 104-107).

Ward, N. G., & Devault, D. (2016). Challenges in building highly interactive dialogue systems. *AI Magazine*, *37*, 7-18. doi: 10.1609/aimag.v37i4.2687

Wen, T.-H., Vandyke, D., Mrksic, N., Gasic, M., Rojas-barahona, L. M., hao Su, P., ... Young, S. (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, eacl* (p. 438-449).

Williams, J., Raux, A., Ramachandran, D., & Black, A. (2013). The dialog state tracking challenge. In *Proceedings of the sigdial 2013 conference* (p. 404-413).

Williams, J. D. (2010). Incremental partition recombination for efficient tracking of multiple dialog states. In *Proceedings of the ieee international conference on acoustics, speech and signal processing* (p. 5382–5385). doi: 10.1109/ICASSP.2010.5494939

Williams, J. D. (2012). Challenges and opportunities for state tracking in statistical spoken dialog systems: Results from two public deployments. *IEEE Journal on Selected Topics in Signal Processing*, *6*, 959-970. doi: 10.1109/JSTSP.2012.2229691

Williams, J. D. (2014). Web-style ranking and slu combination for dialog state tracking. In *Proceedings of the sigdial 2014 conference* (p. 282-291).

Williams, J. D., Asadi, K., & Zweig, G. (2017). Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. In *Proceedings of the 55th annual meeting of the association for computational linguistics.* doi: 10.18653/v1/P17-1062

Williams, J. D., & Liden, L. (2017). Demonstration of interactive teaching for end-to-end dialog control with hybrid code networks. In *Proceedings of the sigdial 2017 conference* (p. 82-85).

Williams, J. D., Poupart, P., & Young, S. (2005). Factored partially observable markov decision processes for dialogue management. In *Proceedings of the work-*

shop on knowledge and reasoning in practical dialogue systems, international joint conference on artificial intelligence (ijcai).

Williams, J. D., Raux, A., & Henderson, M. (2016). The dialog state tracking challenge series: A review. *Dialogue Discourse*, *7*, 4-33. doi: 10.5087/dad.2016.301

Williams, J. D., & Young, S. (2007). Partially observable markov decision processes for spoken dialog systems. *Computer Speech and Language*, *21*, 393-422.

Worsham, J., & Kalita, J. (2020). Multi-task learning for natural language processing in the 2020s: Where are we going? *Pattern Recognition Letters*, *136*. doi: 10.1016/j.patrec.2020.05.031

Wu, C.-S., Hoi, S., & Xiong, C. (2020). Improving limited labeled dialogue state tracking with self-supervision. In *Findings of the association for computational linguistics: Emnlp 2020* (p. 4462-4472).

Wu, C.-S., Madotto, A., Hosseini-Asl, E., Xiong, C., Socher, R., & Fung, P. (2019). Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th annual meeting of the association for computational linguistics.*

Wu, P., Zou, B., Jiang, R., & Aw, A. T. (2020). Gcdst: A graph-based and copy-augmented multi-domain dialogue state tracking. In *Findings of the association for computational linguistics findings of acl: Emnlp 2020* (p. 1063-1073). doi: 10.18653/v1/2020.findings-emnlp.95

Xu, P., & Hu, Q. (2018). An end-to-end approach for handling unknown slot values in dialogue state tracking. In *Proceedings of the 56th annual meeting ofthe association for computational linguistics* (p. 1448-1457).

Yoshino, K., Hiraoka, T., Neubig, G., & Nakamura, S. (2016). Dialogue state tracking using long short term memory neural networks. In *Proceedings of the international workshop on spoken dialogue systems, iwsds 2016.*

Young, S., Gasic, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., & Yu, K. (2010). The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech and Language*, *24*, 150-174. doi: 10.1016/j.csl.2009.04.001

Young, S., Schatzmann, J., Weilhammer, K., & Ye, H. (2007). The hidden information state approach to dialog management. In *Proceedings of the ieee international conference on acoustics, speech and signal processing.* doi: 10.1109/ICASSP.2007.367185

Yu, K., Sun, K., Chen, L., & Zhu, S. (2015). Constrained markov bayesian polynomial for efficient dialogue state tracking. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *23*, 2177-2188. doi: 10.1109/TASLP .2015.2470597

Yu, T., Zhang, R., Polozov, O., Meek, C., & Awadallah, A. H. (2021). Score: Pre-training for context representation in conversational semantic parsing. In *Proceedings of the 9th international conference on learning representations.*

Zachrau, M. (2022). Relationality is not enough: The organization of dynamic structures. In *Proceedings of the 26th workshop on the semantics and pragmatics of dialogue* (p. 116-124).

Zeng, Y., & Nie, J.-Y. (2020). Multi-domain dialogue state tracking based on state graph. *ArXiv*. Retrieved from `http://arxiv.org/abs/2010.11137`

Zeng, Y., & Nie, J.-Y. (2021). Jointly optimizing state operation prediction and value generation for dialogue state tracking. *ArXiv*. Retrieved from `http://arxiv.org/abs/2010.14061`

Zhang, B., Cai, Q., Mao, J., Chang, E., & Guo, B. (2001). Spoken dialogue management as planning and acting under uncertainty. In *Proceedings of the eurospeech conference* (p. 2169-2172).

Zhang, J.-G., Hashimoto, K., Wu, C.-S., Wan, Y., Yu, P. S., Socher, R., & Xiong, C. (2020). Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. In *Proceedings of the 9th joint conference on lexical and computational semantics (*sem)* (p. 154-167).

Zhang, Z., Takanobu, R., Zhu, Q., Huang, M. L., & Zhu, X. Y. (2020). Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, *63*, 2011-2027. doi: 10.1007/s11431-020-1692-3

Zhao, G., Zhao, J., Li, Y., Alt, C., Schwarzenberg, R., Hennig, L., ... Xu, F. (2019). Moli: Smart conversation agent for mobile customer service. *Information (Switzerland)*, *10*. doi: 10.3390/info10020063

Zhao, J., Mahdieh, M., Zhang, Y., Cao, Y., & Wu, Y. (2021). Effective sequence-to-sequence dialogue state tracking. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (p. 7486-7493).

Zhao, T., & Eskenazi, M. (2016). Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *Proceedings of the sigdial 2016 conference* (p. 1-10).

Zhao, T., & Kawahara, T. (2021). Multi-referenced training for dialogue response generation. In *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue* (p. 190-201). Retrieved from `http://arxiv.org/abs/2009.07117`

Zheng, J., Salvi, O., & Chan, J. (2020). Candidate attended dialogue state tracking using bert. In *Proceedings of the 8th dialog system technology challenge (dstc8), aaai-20 workshop.*

Zhong, V., Xiong, C., & Socher, R. (2018). Global-locally self-attentive dialogue state tracker. In *Proceedings of the 56th annual meeting ofthe association for computational linguistics* (p. 1458-1467).

Zhou, J., Wu, H., Lin, Z., Li, G., & Zhang, Y. (2021). Dialogue state tracking with multi-level fusion of predicted dialogue states and conversations. In *Proceedings of the 22nd annual meeting of the special interest group on discourse and dialogue* (p. 228-238). Retrieved from `http://arxiv.org/abs/2107.05168`

Zhou, L., & Small, K. (2019). Multi-domain dialogue state tracking as dynamic knowledge graph enhanced question answering. *ArXiv*. Retrieved from `https://arxiv.org/abs/1911.06192`

Zhu, S., Chen, L., Sun, K., Zheng, D., & Yu, K. (2014). Semantic parser enhancement for dialogue domain extension with little data. In *Proceedings of 2014 ieee workshop on spoken language technology* (p. 336-341).

Zilka, L., & Jurcicek, F. (2015a). Incremental lstm-based dialog state tracker. In *Proceedings of 2015 ieee workshop on automatic speech recognition and understanding (asru)* (p. 757-762). doi: 10.1109/ASRU.2015.7404864

Zilka, L., & Jurcicek, F. (2015b). Lectrack: Incremental dialog state tracking with long short-term memory networks. In *Proceedings of the 18th international conference on text, speech, and dialogue* (Vol. 9302, p. 174-182).

Zilka, L., Marek, D., Korvas, M., & Jurcicek, F. (2013). Comparison of bayesian discriminative and generative models for dialogue state tracking. In *Proceedings of the sigdial 2013 conference* (p. 452-456).

Zue, V., Seneff, S., Glass, J. R., Polifroni, J., Pao, C., Hazen, T. J., & Hetherington, L. (2000). Jupiter: A telephone-based conversational interface for weather information. *IEEE Transactions on Speech and Audio Processing*, *8*, 85-96. doi: 10.1109/89.817460

Appendices

Appendix A

Experiment Training Details

This appendix provides the training details of the energy-based learning method. Table A.1 presents the hyperparameters used in the experiments detailed in Chapters 5 and 6.

In the experiments in Chapter 7 the number of turn-level LSTM cells was increased to 5 to handle the larger number of domains in dialogue data. The other training parameters remain the same as in the case of single domain dialogue data (see Table A.1).

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

	Hyper parameter	Value
Feature network	Machine acts encoded size	300
	Encoder output activation	*tanh*
	Word embedding size	300
	LSTM number of turn-level cells	3
	LSTM number of units	128
	LSTM drop out	0.2
	LSTM output activation	*tanh*
	Pretraining convergence epochs	10
Energy network	Energy non-linearity function	*softplus*
	Energy loss function	Cross entropy
	Regularisation coefficient	0.01
	Training optimiser	Adam
	Training learning rate	0.001
	Training convergence epochs	10
	Inference number of iterations	50
	Inference learning rate	0.001

Appendix B

List of Publications

Long Papers

- (Trinh et al., 2018) Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2018). 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).

- (Trinh et al., 2019a) Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2019a). Capturing Dialogue State Variable Dependencies with an Energy-based Neural Dialogue State Tracker. In Proceedings of the sigdial 2019 conference (pp. 75–84).

- (Trinh et al., 2019b) Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2019b). Energy-Based Modelling for Dialogue State Tracking. In Proceedings of the 1st workshop on nlp for conversational ai (pp. 77–86).

- (Trinh et al., 2020a) Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2020a). Energy-based Neural Modelling for Large-Scale Multiple Domain Dialogue State Tracking. In Proceedings of the 4th workshop on structured prediction for nlp (spnlp) (pp. 33–42).

- (Trinh et al., 2020b) Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2020b). F-Measure Optimisation and Label Regularisation for Energy-Based Neural Dialogue State Tracking Models. In Proceedings of the 29th international conference on artificial neural networks (icann) (pp. 798–810).

Extended Abstracts

- (Trinh et al., 2017) Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2017). Incremental Joint Modelling for Dialogue State Tracking. In Proceedings of the 21st workshop on the semantics and pragmatics of dialogue (semdial) (pp. 176–177).

- (Trinh et al., 2019c) Trinh, A. D., Ross, R. J., & Kelleher, J. D. (2019c). Investigating Variable Dependencies in Dialogue States. In Proceedings of the 23rd workshop on the semantics and pragmatics of dialogue (semdial) (pp. 195–197).

Position Papers

- (Trinh, 2017) Trinh, A. D. (2017). Dialogue Management Modelling. In Pro-

ceedings of the 13th workshop on spoken dialogue systems for phds, postdocs & new researchers (yrrsds) (pp. 23–24).

- (Trinh, 2019) Trinh, A. D. (2019). Dialogue State Tracking. In Proceedings of the 15th workshop on spoken dialogue systems for phds, postdocs & new researchers (yrrsds) (pp. 18–19).

Appendix C

List of Employability and Discipline Specific Skills Training

Employability Skills

- SPEC 9997 – Scientific Research & Literature (5 ECTS)
- SPEC 9160 – Problem Solving, Innovation & Communications (5 ECTS)
- MATH 9102 – Probability & Statistical Inference (5 ECTS)
- MATH 9953 – Algorithms & Approximation Theory (5 ECTS)

Discipline Specific Training Skills

- ENEH 1027 – Advanced Topics in Research: Computational Intelligence (5 ECTS)

- COMP 9001 – Deep Learning (5 ECTS)
- SPEC 9270 – Machine Learning (10 ECTS)

Additional Training

- DeepLearn 2017 – 1st International Summer School on Deep Learning (Bilbao, Spain)
- DeepLearn 2018 – 2nd International Summer School on Deep Learning (Genova, Italy)
- LxMLS 2018 – 8th Lisbon Machine Learning School (Lisbon, Portugal)
- TLMSS 2018 – Transylvanian Machine Learning Summer School (Cluj-Napoca, Romania)